Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Comparative Analysis of Machine Learning Classifiers with Feature Selection for Early Diabetes Prediction (Tabular ML Classification)

Maimona

MS Virtual University of Pakistan. Email: maimonamasood4@gmail.com

ABSTRACT

The early prediction of diabetes mellitus represents a critical objective in preventive healthcare, offering the potential to delay or prevent disease onset through targeted interventions. As electronic health records and structured clinical datasets proliferate, machine learning (ML) models have increasingly been applied to tabular healthcare data to estimate diabetes risk. However, the translation of these models into clinical practice is hindered by methodological flaws, including data leakage, inappropriate evaluation metrics, and the misapplication of feature selection techniques. This article provides a comprehensive, critical review of tabular ML classification for early diabetes prediction. We distinguish the clinical target of predicting future incidence from classifying undiagnosed baseline disease. Furthermore, we examine the unique challenges of tabular clinical data, including mixed data types, missingness, and class imbalance. A comparative analysis of foundational classifiers (e.g., Logistic Regression, Random Forests, Gradient Boosting) and feature selection paradigms (filter, wrapper, and embedded methods) is conducted. To address widespread methodological deficiencies in the empirical literature, we propose a rigorous, leakage-resistant experimental protocol utilizing nested cross-validation and pipeline integration. Finally, we advocate for multi-metric evaluation encompassing discrimination, calibration, and clinical utility, while emphasizing the imperative of model interpretability and fairness for responsible clinical deployment.

Keywords: Machine Learning, Tabular Data, Feature Selection, Diabetes Prediction, Clinical Predictive Modeling, Data Leakage, Healthcare Artificial Intelligence

Introduction

Diabetes mellitus is a chronic metabolic disorder with an accelerating global prevalence, contributing to significant morbidity, mortality, and healthcare expenditure (World Health Organization, 2021). The insidious onset of Type 2 Diabetes (T2D), which comprises the vast majority of cases, means that microvascular and macrovascular complications often develop prior to formal diagnosis (American Diabetes Association, 2023). Consequently, early prediction and risk stratification are paramount. Machine learning (ML), particularly the

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

application of supervised classification algorithms to structured, tabular clinical data, has emerged as a powerful paradigm for modeling complex, non-linear interactions among physiological, demographic, and lifestyle risk factors (Zou et al., 2018).

Despite the proliferation of predictive models, the literature is saturated with studies exhibiting critical methodological flaws. High predictive performance is frequently an artifact of data leakage, target confounding, or the absence of robust external validation (Cawley & Talbot, 2010; Vabalas et al., 2019). Furthermore, while feature selection is routinely employed to reduce dimensionality and mitigate the "curse of dimensionality," it is often executed prior to data partitioning, resulting in overfitting and inflated performance estimates (Hastie, Tibshirani, & Friedman, 2009).

This article aims to rigorously evaluate the landscape of machine learning applied to tabular diabetes data. Given the absence of a universally optimal algorithm, we propose a standardized experimental framework to benchmark ML classifiers against various feature selection strategies, emphasizing stringent data-leakage prevention and clinically relevant evaluation metrics.

Clinical Context and Definition of Early Diabetes Prediction

In the context of predictive modeling, "early prediction" is frequently conflated with disparate clinical tasks. It is imperative to precisely define the clinical target and the prediction horizon:

1. **Incident (Future) Diabetes Prediction:** The true forecasting of future disease onset (e.g., within 3, 5, or 10 years) in a currently normoglycemic or prediabetic cohort. This requires longitudinal data.
2. **Detection of Undiagnosed Diabetes:** The identification of current, existing disease in individuals who have not yet received a formal diagnosis. This relies on cross-sectional screening data.
3. **Progression and Complication Risk:** Predicting the trajectory of individuals who already possess a diabetes diagnosis (e.g., risk of diabetic retinopathy, nephropathy, or cardiovascular events).

This article focuses primarily on the prediction of future/incident diabetes and the detection of undiagnosed diabetes, as these align with early intervention. ML predictions are not substitutes for confirmatory diagnostic tests (such as HbA1c or oral glucose tolerance tests) but serve as risk stratification tools to allocate screening resources and intensive lifestyle interventions efficiently.

Characteristics and Challenges of Tabular Healthcare Data

Unlike image or signal data, tabular clinical data present distinct structural and statistical challenges that dictate preprocessing and modeling choices (Kuhn & Johnson, 2013).

- **Mixed Data Types:** Tabular records inherently contain continuous (e.g., BMI, fasting plasma glucose), ordinal (e.g., self-reported health status), categorical (e.g., ethnicity), and binary (e.g., history of hypertension) variables. Classifiers vary

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

in their ability to handle non-numerical inputs directly.

- **Missingness:** Missing data in electronic health records (EHR) is rarely Missing Completely at Random (MCAR). It is often Missing at Random (MAR) or Missing Not at Random (MNAR), where the absence of a lab test is clinically informative. Naive imputation can introduce bias.
- **Class Imbalance:** In general population screening, incident diabetes cases are the minority class. Severe imbalance can bias classifiers toward the majority class, necessitating threshold tuning, class weighting, or controlled resampling (e.g., SMOTE) (Chawla et al., 2002).
- **Correlated Predictors and Multicollinearity:** Physiological variables (e.g., systolic and diastolic blood pressure, or BMI and waist circumference) often exhibit high collinearity, which can destabilize coefficients in linear models and complicate feature importance metrics in tree-based models.
- **Dataset Shift and Representativeness:** Models trained on specific demographic cohorts (e.g., the heavily researched Pima Indians Diabetes Database) often fail to generalize to broader, more diverse populations due to covariate shift and differences in baseline prevalence.

ML Classifiers for Tabular Diabetes Prediction

Tabular data classification relies on algorithms capable of partitioning feature spaces defined by discrete and continuous variables. The selection of an algorithm entails trade-offs regarding assumptions, computational cost, and interpretability.

Table 1: ***Comparative Analysis of Machine-Learning Classifiers for Tabular Diabetes Prediction***

Classifier	Core Principle	Strengths for Tabular Data	Limitations & Overfitting Risks
Logistic Regression (LR)	Models log-odds as a linear combination of features.	Highly interpretable (coefficients = odds ratios); well-calibrated probabilities; computationally lightweight.	Assumes linear relationships; struggles with complex interactions unless explicitly engineered; vulnerable to multicollinearity.
Decision Tree (DT)	Recursive, hierarchical partitioning based on	Handles mixed data types naturally; invariant to feature scaling; interpretable	High variance; highly prone to overfitting on tabular data without

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Classifier	Core Principle	Strengths for Tabular Data	Limitations & Overfitting Risks
	feature thresholds.	structure.	strict pruning.
Random Forest (RF)	Ensemble of unpruned trees trained on bootstrapped samples with random feature subsets.	Robust to outliers and non-linearities; less prone to overfitting than DTs; provides Out-of-Bag error.	Computationally intensive; acts as a "black box"; can favor features with more categories in importance metrics.
Gradient Boosting (XGBoost, LightGBM)	Sequential ensemble of weak trees, optimizing a differentiable loss function.	State-of-the-art performance on tabular data; inherent handling of missing values (in some implementations).	Requires extensive hyperparameter tuning; highly susceptible to overfitting if learning rate/depth are not controlled.
Support Vector Machine (SVM)	Finds the optimal separating hyperplane maximizing the margin.	Effective in high-dimensional spaces; non-linear mapping via kernel trick.	Computationally prohibitive on very large tabular datasets (O(n^2) to O(n^3)); requires rigorous feature scaling.
k-Nearest Neighbours (k-NN)	Classifies based on majority vote of the *k* closest training examples.	Simple, non-parametric; adapts to local data structures.	Requires careful scaling; sensitive to irrelevant features (curse of dimensionality); slow inference.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Classifier	Core Principle	Strengths for Tabular Data	Limitations & Overfitting Risks
Multilayer Perceptron (MLP)	Feedforward artificial neural network.	Capable of approximating any continuous function; captures deep interactions.	Usually outperformed by tree ensembles on standard tabular data unless sample size is massive; highly opaque.

Note: The choice of classifier is dataset-dependent. Baseline performance must always be established using simple, interpretable models (e.g., LR) before deploying complex ensembles.

Feature-Selection Methods

Dimensionality reduction via feature selection is critical to eliminate redundant, irrelevant, or noisy variables, thereby enhancing model generalization, speeding up inference, and improving interpretability (Dash & Liu, 1997). Unlike feature extraction (e.g., PCA), feature selection retains the original variables, preserving clinical meaning.

Table 2: ***Comparative Analysis of Feature-Selection Methods for Tabular Clinical Data***

Selection Family	Mechanism	Examples	Strengths	Limitations & Leakage Risks
Filter Methods	Evaluates feature relevance based on intrinsic statistical properties, independently of the classifier.	Pearson correlation, ANOVA F-test, Mutual Information, Chi-square.	Computationally very fast; completely independent of the downstream model; robust to overfitting.	Ignores feature interactions and model biases; may select highly correlated (redundant) features.
Wrapper Methods	Uses a specific classifier to	Recursive Feature Elimination	Considers feature interactions;	Computationally expensive; high risk of

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Selection Family	Mechanism	Examples	Strengths	Limitations & Leakage Risks
	evaluate subsets of features via a search algorithm.	(RFE), Sequential Forward/Backward Selection.	specifically tailored to the chosen classifier's inductive bias.	overfitting if cross-validation is not strictly enforced within the wrapper.
Embedded Methods	Feature selection is integrated into the model training process as a penalty or structural choice.	LASSO (L1 regularization), Tree-based Gini importance.	Balances computational efficiency and interaction modeling; optimized alongside model weights.	Model-dependent; tree importance can be biased toward continuous variables or masked by collinearity.

Clinical plausibility must act as a secondary filter; features selected by algorithms must be evaluated by domain experts to ensure they are biologically sound and available at the time of prediction.

Critical Review of Empirical Studies

The academic literature on ML for diabetes prediction features high variance in methodological quality (Collins et al., 2015). A vast proportion of early studies utilized the Pima Indians Diabetes Database (Smith et al., 1988). While foundational for algorithmic benchmarking, its cross-sectional nature, high missingness (often erroneously treated as zeros), and specific demographic constrain its generalizability as a modern predictive cohort.

More recent, rigorous studies utilizing broader datasets (e.g., NHANES or institutional EHRs) consistently demonstrate that ensemble methods, particularly Gradient Boosting (Chen & Guestrin, 2016) and LightGBM (Ke et al., 2017), achieve superior discrimination. However, Dagliati et al. (2018) note that regularized Logistic Regression often achieves parity with complex ensembles when robust feature engineering is applied, highlighting the necessity of comparing against strong, simple baselines.

A systematic flaw across the literature is the application of feature selection (e.g.,

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

filtering by correlation to the target) on the *entire* dataset prior to train/test splitting. This induces severe data leakage, resulting in hyper-optimistic performance reports that fail upon external validation (Cawley & Talbot, 2010).

Proposed Conceptual Framework

To advance the field beyond flawed empirical benchmarks, we propose a conceptual framework integrating model selection, stringent feature selection, and clinical evaluation. This framework posits that no single combination is universally optimal; rather, optimization is a function of the specific dataset's dimensionality, sample size, and class prevalence.

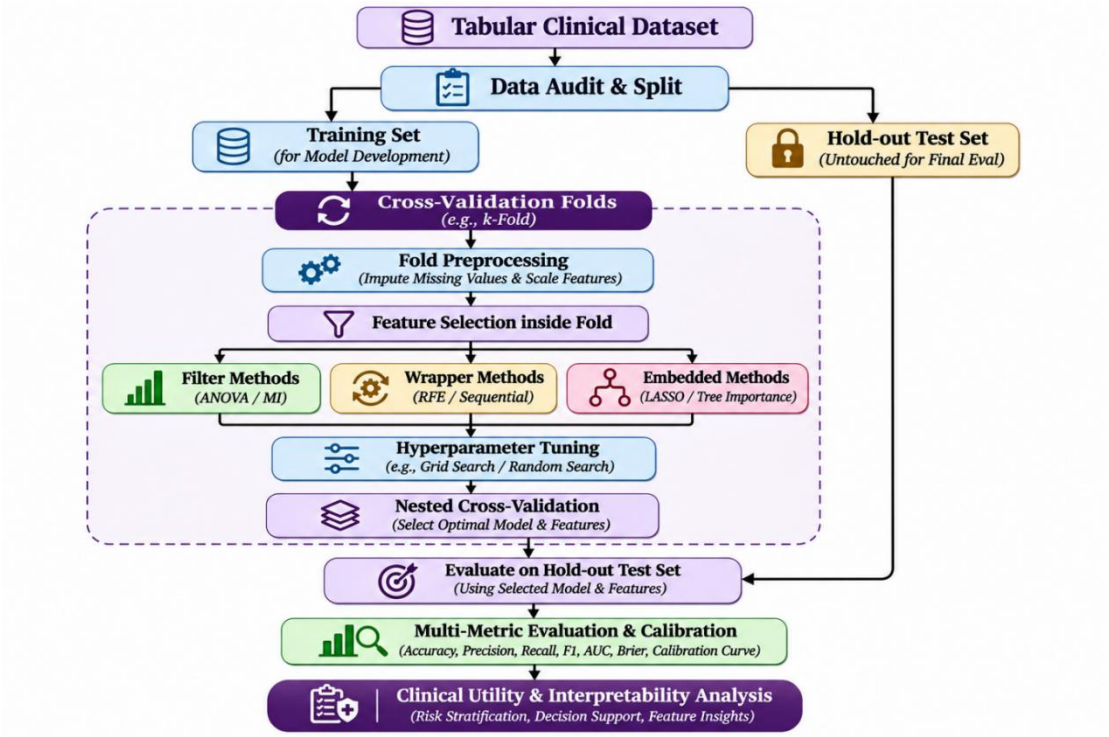

Figure 1: Architecture of Feature Selection and Tabular Classifier Evaluation.

This highlights the absolute necessity of embedding preprocessing and feature selection strictly within the cross-validation folds.

Methodology or Proposed Experimental Protocol

For researchers developing ML models for early diabetes prediction, we advocate the following rigorous experimental protocol, assuming a verified longitudinal dataset (e.g., predicting 5-year incidence):

1. **Baseline Definition:** Define LR using all eligible predictors as the baseline.
2. **Experiment Matrix:** Systematically compare:
   - Classifiers: LR, RF, XGBoost, and SVM.
   - Feature Selection: Baseline (None), Filter (Mutual Information), Wrapper (SVM-RFE), Embedded (LASSO).
3. **Data Partitioning:** Initially split the data into 70% training and 30% hold-

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

out test set, stratified by the target variable.

4. **Tuning:** Utilize randomized search with 5-fold cross-validation on the training set to optimize hyperparameters.

5. **Pipeline Construction:** Enforce leakage-resistant workflows (detailed below).

(Note: As this article is a literature-based review and protocol, specific empirical outcomes are designated as "to be evaluated" by future researchers implementing this framework).

Leakage-Resistant Development and Validation

Data leakage occurs when information from outside the training dataset is used to create the model, leading to fundamentally invalid performance metrics (Kuhn & Johnson, 2013). To guarantee research integrity, the following workflow is mandatory:

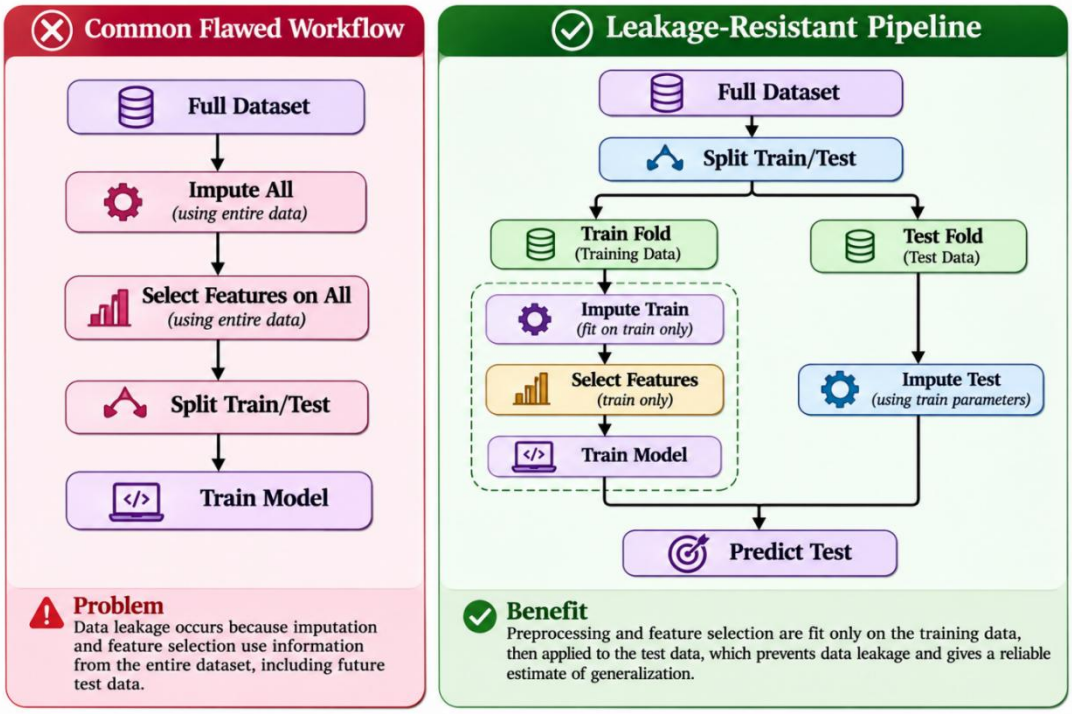

Figure 2: Contrast between flawed and leakage-resistant tabular ML workflows. Operations like mean imputation or feature selection must compute parameters solely on the training partition.

1. **Strict Pipeline Architecture:** Standard scaling, categorical encoding (e.g., target encoding), imputation, and feature selection must be instantiated as sequential steps in a mathematical pipeline (e.g., Scikit-learn's Pipeline object) (Pedregosa et al., 2011).

2. **Nested Cross-Validation:** If feature selection wrapper methods are used simultaneously with hyperparameter tuning, an outer CV loop must evaluate generalization while an inner CV loop optimizes parameters, preventing optimization bias.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

3. **Temporal Validation:** Where possible, the test set should be partitioned chronologically (e.g., train on 2015-2020, test on 2021-2023) to assess resilience to dataset shift (Altman & Royston, 2000).

Evaluation Metrics and Comparative Analysis

In healthcare applications, evaluating models using basic accuracy is medically indefensible, particularly given disease class imbalance. A model predicting that "no one has diabetes" in a cohort with 5% prevalence achieves 95% accuracy while possessing zero clinical utility.

Evaluation must cover three distinct dimensions (Steyerberg, 2019):

1. **Discrimination:**
   - **ROC-AUC:** Assesses overall ranking ability across all thresholds.
   - **PR-AUC (Average Precision):** More informative than ROC-AUC under severe class imbalance.
   - **Sensitivity/Recall:** Crucial for minimizing false negatives (missed early interventions).
   - **Specificity & PPV/Precision:** Crucial for minimizing false positives (unnecessary anxiety and confirmatory testing).
2. **Calibration:** A classifier may discriminate well but output poorly scaled probabilities. Calibration plots (reliability diagrams) and the Brier score (Brier, 1950) determine whether a predicted risk of 0.2 translates to a 20% observed incidence.
3. **Clinical Utility:** Decision Curve Analysis (DCA) calculates the "Net Benefit" of using the ML model across various clinical probability thresholds, directly comparing the model against default strategies like "treat all" or "treat none" (Vickers & Elkin, 2006).

Interpretability, Fairness, and Clinical Implications

For a diabetes prediction model to achieve clinical acceptance, it must be transparent. "Black box" predictions violate the principles of evidence-based medicine and shared decision-making.

- **Model Agnostic Interpretability:** Techniques like SHapley Additive exPlanations (SHAP) provide local (patient-level) and global (cohort-level) feature importance metrics (Lundberg & Lee, 2017). SHAP values can explain how a high BMI and a family history act in concert to push an individual's risk score beyond the decision threshold. However, researchers must categorically avoid conflating feature importance with causal inference.
- **Algorithmic Fairness:** Tabular clinical data often reflect systemic healthcare disparities. A model must be audited for disparate performance across subgroups (e.g., age, sex, ethnicity). If a model exhibits high sensitivity for one demographic but high false-negative rates for another, its deployment could exacerbate health inequities.
- **Workflow Integration:** Predictive models should trigger specific clinical pathways. For example, individuals crossing a lower threshold might receive digital

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

lifestyle interventions, while those crossing a higher threshold may be referred for urgent confirmatory HbA1c testing.

Discussion

The comparative analysis of ML classifiers and feature selection methods reveals that the "best" algorithmic choice is context-dependent. Linear models with domain-informed feature selection often yield non-inferior clinical utility to complex tree ensembles, while providing inherent transparency (Moons et al., 2015). Feature selection methodologies must prioritize the biological plausibility and clinical availability of features just as heavily as algorithmic optimization. Filter methods offer a safe, low-leakage baseline, while embedded methods in Gradient Boosting machines represent the current state-of-the-art for high-dimensional clinical tabular data, provided rigorous tuning prevents overfitting.

Limitations

This article acknowledges several limitations. As a proposed framework and critical review, it does not present new empirical clinical findings, relying instead on synthesizing established ML principles with epidemiological best practices. Furthermore, the framework assumes the availability of sufficiently powered, longitudinal data, which is often difficult to secure due to privacy constraints and fragmented EHR systems. Finally, while tabular data remain the foundation of predictive health, the future likely involves multi-modal models combining tabular data with genomics and unstructured clinical notes.

Future Research

Future research must transition away from repeatedly benchmarking algorithms on static, outdated cross-sectional datasets. Priorities should include:

1. Prospective, randomized clinical trials evaluating patient outcomes when ML risk stratification is deployed versus standard of care.
2. The development of federated learning protocols, allowing institutions to collaboratively train diabetes prediction models on tabular data without transferring sensitive patient records.
3. Enhanced methods for causal representation learning within tabular ML, helping to distinguish predictive correlations from interventional targets.

Conclusion

The application of machine learning to tabular data for the early prediction of diabetes holds immense promise for preventive medicine. However, the field is constrained by pervasive methodological errors, notably data leakage during feature selection and reliance on misleading evaluation metrics. By adopting leakage-resistant pipelines, prioritizing calibration and Decision Curve Analysis alongside discrimination, and ensuring model transparency via SHAP or interpretable baselines, researchers can develop robust algorithms. Ultimately, the success of a diabetes prediction model is not defined by its AUC, but by its capacity to be safely, fairly, and effectively integrated into clinical workflows to delay or prevent the onset

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

of disease.

References

1. Altman, D. G., & Royston, P. (2000). What do we mean by validating a prognostic model? *Statistics in Medicine*, *19*(4), 453-473. https://doi.org/10.1002/(sici)1097-0258(20000229)19:4<453::aid-sim350>3.0.co;2-5
2. American Diabetes Association. (2023). 2. Classification and diagnosis of diabetes: Standards of Care in Diabetes—2023. *Diabetes Care*, *46*(Supplement_1), S19-S40. https://doi.org/10.2337/dc23-S002
3. Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324
4. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1-3. https://doi.org/10.1175/1520-0493(1950)078<0001:vofeit>2.0.co;2
5. Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*(70), 2079-2107.
6. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321-357. https://doi.org/10.1613/jair.953
7. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. https://doi.org/10.1145/2939672.2939785
8. Collins, G. S., Reitsma, J. B., Altman, D. G., & Moons, K. G. (2015). Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD statement. *Annals of Internal Medicine*, *162*(10), 735-736. https://doi.org/10.7326/M14-0697
9. Dagliati, A., Marini, S., Sacchi, L., Giuffrida, G., Martino, M., Klersy, C., ... & Bellazzi, R. (2018). Machine learning methods to predict diabetes complications. *Journal of Diabetes Science and Technology*, *12*(2), 295-302. https://doi.org/10.1177/1932296817706375
10. Dash, M., & Liu, H. (1997). Feature selection for classification. *Intelligent Data Analysis*, *1*(3), 131-156. https://doi.org/10.1016/S1088-467X(97)00008-5
11. Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*(1), 389-422. https://doi.org/10.1023/A:1012487302797
12. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer.
13. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in*

Online ISSN 3007-3197
Print ISSN 3007-3189

Neural Information Processing Systems, *30*.

14. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, *2*, 1137-1143.
15. Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. https://doi.org/10.1007/978-1-4614-6849-3
16. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.
17. Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., ... & Collins, G. S. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis (TRIPOD): Explanation and elaboration. *Annals of Internal Medicine*, *162*(1), W1-W73. https://doi.org/10.7326/M14-0698
18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.
19. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. *Advances in Neural Information Processing Systems*, *31*.
20. Riley, R. D., Ensor, J., Snell, K. I., Harrell, F. E., Martin, G. P., Reitsma, J. B., ... & Collins, G. S. (2020). Calculating the sample size required for developing a clinical prediction model. *BMJ*, *368*, m441. https://doi.org/10.1136/bmj.m441
21. Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, 261-265.
22. Steyerberg, E. W. (2019). *Clinical prediction models: A practical approach to development, validation, and updating* (2nd ed.). Springer.
23. Vabalas, A., Gowen, E., Poliakoff, E., & Casson, A. J. (2019). Machine learning algorithm validation with a limited sample size. *PLOS ONE*, *14*(11), e0224365. https://doi.org/10.1371/journal.pone.0224365
24. Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making*, *26*(6), 565-574. https://doi.org/10.1177/0272989X06295361
25. World Health Organization. (2021). *Diabetes*. https://www.who.int/news-room/fact-sheets/detail/diabetes
26. Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in Genetics*, *9*, 515. https://doi.org/10.3389/fgene.2018.00515