

Article

Beyond Accuracy: Reliability-Aware Machine Learning for Handwriting-Based Alzheimer's Disease Detection

Uddalak Mitra ¹  and Shafiq Ul Rehman ^{2,*} 

¹ Department of Computer Science and Engineering, JIS College of Engineering, Kalyani 741235, India; uddalak.mitra@jiscollege.ac.in

² College of Information Technology, Kingdom University, P.O. Box 40434, Riffa 3903, Bahrain

* Correspondence: s.rehman@ku.edu.bh

Abstract

Reliable clinical decision support systems require not only high predictive accuracy but also trustworthy probability estimates and robust uncertainty quantification. However, most medical artificial intelligence (AI) studies primarily emphasize discrimination performance while overlooking systematic reliability evaluation. This study proposes a reliability-aware evaluation framework for Alzheimer's disease detection that integrates discrimination analysis, statistical validation, probability calibration, uncertainty quantification, robustness assessment, and clinical decision analysis within a unified pipeline. Multiple machine learning classifiers and ensemble configurations were evaluated using repeated stratified cross-validation and assessed through discrimination and calibration metrics. Support Vector Machine achieved the highest ROC-AUC (0.955 ± 0.046), while Extra Trees obtained the highest Accuracy (0.878) and F1-score (0.887). Friedman analysis confirmed statistically significant differences among classifiers ($p < 0.001$). Platt scaling consistently improved probabilistic reliability, whereas Beta calibration demonstrated stable performance under noise, feature perturbation, and reduced-data scenarios. Uncertainty-aware selective prediction increased high-confidence diagnostic accuracy by up to 8.1%, and decision curve analysis demonstrated improved clinical utility. The reliability analysis identified calibration-aware stacking as the most reliable ensemble configuration. An independent cross-dataset evaluation on a heterogeneous Alzheimer's disease clinical dataset with a substantially different feature space yielded stable discrimination (ROC-AUC = 0.858 ± 0.025) and calibration (ECE = 0.132 ± 0.019) after the STACK_CAL architecture was independently retrained from scratch. These findings provide evidence of the cross-dataset applicability of the proposed reliability-aware strategy across different clinical data modalities, while further prospective and independent validation remains necessary before real-world clinical deployment.



Academic Editors: Michel Audette, Jerzy W. Rozenblit and Christopher P. Paolini

Received: 5 August 2026

Revised: 23 August 2026

Accepted: 31 August 2026

Published: 9 September 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 \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: Alzheimer's disease; model-agnostic framework; reliability evaluation; probability calibration; trustworthy machine learning

1. Introduction

Alzheimer disease (AD) is a progressive neurodegenerative disease and one of the most common causes of dementia in the world, impacting millions of elderly individuals and representing significant clinical and socioeconomic costs [1]. Early identification and risk stratification play a significant role in intervening in, treating, and planning patient care in time [2]. Conventional clinical examinations are based on cognitive examinations and

neuroimaging; they are not only always time-consuming and costly, and they might not be readily available in low resource areas [3]. The extraction of behavioral biomarkers with the use of daily activities has become possible due to recent advancements in digital healthcare systems. The study of handwriting dynamics is one of them; an effective non-invasive digital biomarker for various neurological conditions, such as AD and Parkinson's disease, has been proposed to date [4,5]. Handwriting production involves complicated cognitive and motor functions that are vulnerable to initial cognitive decline; thus, it is an efficient measure to identify automated screening with machine learning (ML) [6–8].

Although numerous studies report high discrimination performance for handwriting-based AD detection, evaluations predominantly focus on metrics such as accuracy and area under the ROC curve (AUC) [9]. These metrics quantify separability but not the reliability of predicted probabilities. In clinical applications, model outputs are often interpreted as risk scores, and poorly calibrated models can produce over- or under-confident predictions, potentially leading to inappropriate clinical decisions [10,11].

Despite the critical role of uncertainty in clinical interpretation, most existing work does not explicitly assess model calibration or uncertainty quantification. Simple discrimination benchmarks ignore whether predicted probabilities reflect empirical risk, and few studies examine how models behave under perturbations or distribution shifts [12]. Moreover, techniques such as selective prediction and decision curve analysis—which are essential for deployment-oriented evaluation—are largely absent in handwriting-focused AD research.

1.1. Research Gap

Current handwriting-based ML frameworks for AD detection demonstrate strong classification performance but lack a systematic evaluation of probabilistic reliability, uncertainty behavior, and robustness. This gap limits understanding of whether predicted probabilities are trustworthy for clinical risk assessment, especially in high-stakes environments where calibration and confidence estimation are as important as discrimination [13,14]. In particular, the interaction between calibration placement within ensemble architectures and probabilistic reliability has not been thoroughly investigated in this domain.

1.2. Contributions

To address these limitations, we propose a comprehensive reliability-aware evaluation framework for handwriting-based AD detection. Our key contributions include:

- **Reliability-Aware Evaluation Framework:** We introduce a multi-phase methodology that jointly analyzes discrimination, calibration, uncertainty, robustness, and clinical utility, moving beyond conventional accuracy-centric assessment.
- **Comprehensive Calibration Study:** We evaluate multiple calibration strategies—temperature scaling [11], Platt scaling [15], isotonic regression [16], and beta calibration [17]—across diverse classifiers using Brier score, LogLoss, ECE, and MCE.
- **Uncertainty Quantification and Selective Prediction:** We investigate confidence–accuracy alignment and uncertainty-based rejection strategies using risk–coverage analysis for clinically safe predictions [18].
- **Robustness and Stability Evaluation:** We systematically assess stability under noise injection, reduced-data training, and feature perturbation to mimic real-world perturbations [12].
- **Clinical Threshold Optimization and Decision Curve Analysis:** We optimize thresholds for clinical criteria and evaluate net benefit via decision curve analysis [19].

Additionally, we examine how calibration placement within ensemble architectures influences reliability, providing new insight into structural model uncertainty that has not been considered in handwriting-based AD detection literature.

The following section, Section 2, presents the related work, and the proposed reliability-aware framework is introduced in Section 3, while Section 4 demonstrates the computational results. Finally, in Section 5, we conclude the study with possible future directions.

2. Related Work

2.1. Handwriting-Based Digital Biomarkers for Alzheimer's Disease

Digital handwriting analysis has emerged as a promising behavioral biomarker for early cognitive decline. Neurological impairments associated with Alzheimer's disease (AD) affect motor coordination, visuospatial organization, temporal consistency, and executive planning, all of which manifest in handwriting dynamics. Early studies demonstrated that kinematic and pressure-based features extracted from digitized tablets can capture subtle motor abnormalities [4]. Subsequent research extended these approaches to cognitive impairment screening using spatiotemporal and structural handwriting descriptors [6–8].

More recent investigations have applied supervised machine learning and ensemble classifiers to distinguish AD patients from healthy controls using static and dynamic handwriting features [9]. These studies report promising discrimination metrics, often exceeding 85–95% accuracy in controlled datasets. However, evaluation protocols are typically centered on accuracy, sensitivity, specificity, and ROC-AUC, with limited attention to probabilistic reliability or deployment-oriented validation. As a result, although handwriting-based ML systems show strong separability, their clinical trustworthiness remains insufficiently examined.

2.2. Calibration in Clinical Machine Learning

Calibration assesses the agreement between predicted probabilities and observed outcome frequencies. In clinical applications, probability estimates are frequently interpreted as individualized risk scores, making calibration as important as discrimination [20]. Modern neural networks and ensemble models are known to exhibit systematic overconfidence, even when achieving high accuracy [11,21].

Post hoc calibration techniques—including Platt scaling [15], isotonic regression [16], temperature scaling [11], and beta calibration [17]—have been widely studied in general ML and medical imaging. In radiology and diagnostic AI, improper calibration has been shown to distort uncertainty-based referral strategies [10]. Comprehensive reviews further emphasize calibration as a critical component of trustworthy AI deployment in healthcare [22,23].

Despite this growing emphasis, calibration studies remain scarce in digital biomarker-based neurological screening tasks. In particular, handwriting-driven AD detection literature rarely reports Brier score, Expected Calibration Error (ECE), reliability diagrams, or calibration gain analysis. Moreover, structural calibration strategies within ensemble architectures have not been systematically explored in this domain.

2.3. Uncertainty Quantification and Selective Prediction

Uncertainty-aware modeling aims to mitigate overconfident errors in high-stakes environments. Selective classification frameworks allow models to abstain from uncertain predictions, improving reliability among accepted cases [18]. Risk–coverage analysis quantifies the trade-off between predictive performance and abstention rate, providing a deployment-relevant reliability measure.

In medical AI, uncertainty quantification has been applied to imaging diagnostics, survival prediction, and clinical decision support systems [22]. However, several studies highlight that uncertainty estimates are meaningful only when probability outputs are well-calibrated [10]. Without calibration, confidence-based rejection strategies may inadvertently reinforce biased or unstable predictions. To date, handwriting-based AD detection studies

have not systematically incorporated uncertainty modeling or selective prediction analysis into their evaluation frameworks.

2.4. Robustness Under Distributional Shift and Clinical Utility

Real-world clinical deployment exposes models to distributional shifts, data noise, and demographic variability. Empirical evidence shows that predictive uncertainty often degrades under dataset shift, even when discrimination remains stable [24]. Adversarial and perturbation-based analyses further demonstrate the fragility of many medical AI systems [25].

Robust evaluation under reduced-data regimes, feature corruption, and noise injection has therefore become essential for responsible medical AI development. Additionally, decision curve analysis (DCA) provides a clinically interpretable framework for evaluating net benefit across varying risk thresholds [19]. While DCA is increasingly adopted in risk prediction research, it remains underutilized in digital handwriting-based AD detection systems.

2.5. Summary and Research Positioning

Collectively, existing handwriting-based AD detection studies demonstrate strong classification performance but lack systematic reliability validation (Table 1). Calibration analysis, uncertainty-aware selective prediction, robustness evaluation under perturbation, and clinical utility assessment are rarely integrated within a unified framework. Furthermore, the influence of calibration placement within ensemble architectures has not been investigated.

The present study addresses these gaps by proposing a comprehensive reliability-aware evaluation pipeline that integrates discrimination, calibration, uncertainty modeling, robustness testing, and clinical threshold optimization within a structured experimental design.

Table 1. Comparison of existing handwriting-based AD detection studies and the proposed framework. The symbols ‘✓’ and ‘-’ indicate the presence and absence, respectively, of the particular study in the corresponding article.

Study	Handwriting ML	Calibration	Uncertainty	Robustness	DCA	Ensemble Structural Analysis
Drotar et al. [4]	✓	–	–	–	–	–
Oliveira et al. [6]	✓	–	–	–	–	–
Mitra and Rehman [7]	✓	–	–	–	–	–
Rehman and Mitra [8]	✓	–	–	–	–	–
Demircioglu [9]	✓	–	–	–	–	–
Zhang et al. [10]	–	✓	✓	–	–	–
Guo et al. [11]	–	✓	–	–	–	–
Ovadia et al. [24]	–	✓	✓	✓	–	–
Proposed Study	✓	✓	✓	✓	✓	✓

3. Proposed Reliability-Aware Framework

This study proposes a structured reliability-aware evaluation framework for handwriting-based Alzheimer’s disease (AD) detection that extends beyond conventional accuracy-centric machine learning pipelines. Unlike traditional approaches that primarily assess discriminative performance, the proposed framework systematically integrates discrimination analysis, statistical validation, probability calibration, uncertainty quantification, robustness assessment, and clinical decision analysis within a unified multi-stage architecture. By jointly evaluating predictive performance, probabilistic reliability, confidence-aware decision-making, model stability under distributional perturbations, and clinical utility, the framework provides a comprehensive assessment of model trustworthiness for deployment in real-world healthcare settings.

Figure 1 illustrates the overall workflow, which consists of six major components: (1) feature processing and data preparation; (2) baseline classification with statistical performance evaluation; (3) probability calibration and reliability assessment; (4) uncertainty-aware evaluation through confidence analysis and selective prediction; (5) robustness analysis under distributional perturbations; (6) clinical decision analysis involving threshold optimization and decision curve analysis. This integrated pipeline enables systematic evaluation of both predictive performance and the reliability of machine learning models, ensuring that high discrimination is complemented by well-calibrated probabilities, reliable uncertainty estimates, robustness against data variability, and clinically meaningful decision support.

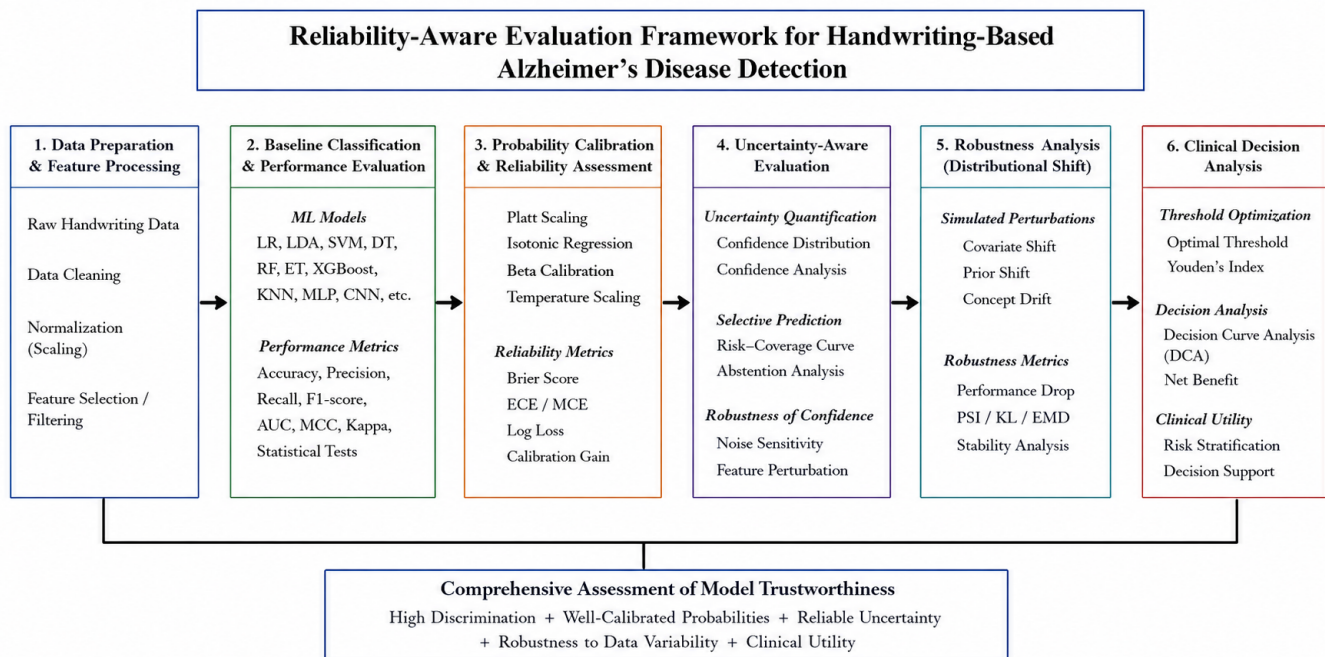


Figure 1. Proposed reliability-aware framework for handwriting-based Alzheimer's disease detection.

3.1. Datasets

To comprehensively evaluate the proposed reliability-aware framework, this study employs two publicly available Alzheimer's disease (AD) datasets representing complementary data modalities and distributional characteristics. The DARWIN handwriting dataset serves as the primary experimental platform for developing and evaluating the proposed framework, including discrimination analysis, statistical validation, probability calibration, uncertainty quantification, robustness assessment, and clinical decision analysis across multiple machine learning models. Subsequently, an independent Alzheimer's disease clinical dataset is used to examine the cross-dataset applicability of the proposed reliability evaluation strategy. Since the handwriting and clinical datasets differ substantially in their feature spaces and data modalities, the model architecture is retrained from scratch on the clinical dataset rather than transferring the handwriting-trained model. Therefore, this experiment is not intended as external validation of a handwriting-based predictive model; instead, it evaluates whether the proposed reliability assessment strategy and the identification of reliable models remain applicable under a substantially different data setting. This two-stage evaluation enables assessment of the framework's reliability across controlled distributional perturbations within the primary dataset as well as its broader applicability across heterogeneous AD datasets, while avoiding claims of direct clinical deployment or external validation of the handwriting-based model.

3.1.1. DARWIN Handwriting Dataset

The DARWIN (Diagnosis Alzheimer With handwriting) dataset (<https://archive.ics.uci.edu/dataset/732/darwin> (accessed on 1 May 2026)) contains handwriting-derived kinematic features collected from 174 participants, including 89 individuals diagnosed with Alzheimer's disease and 85 cognitively healthy controls. The dataset comprises 25 handwriting tasks categorized into Memory and Dictation (M), Graphic (G), and Copy (C) tasks. Each task is represented by 18 temporal, spatial, and motor features, yielding a total of 450 features (25×18) per subject.

The high-dimensional, sensor-derived nature of the DARWIN dataset provides a rich behavioral feature space for investigating probability stability, model calibration, and uncertainty estimation in complex machine learning models. Owing to its high feature dimensionality and controlled experimental setting, the dataset is particularly suitable for analyzing overconfidence phenomena, calibration drift, and structural ensemble behavior.

In this study, the DARWIN dataset constitutes the primary benchmark for constructing and evaluating the proposed reliability-aware framework. Multiple baseline classifiers and ensemble variants are first evaluated using discrimination metrics, statistical significance tests, probability calibration, uncertainty quantification, robustness analysis, and clinical decision measures. Furthermore, all controlled distributional shift experiments—including Gaussian noise injection, feature perturbation, and reduced-data simulations—are performed exclusively on this dataset. These controlled perturbations enable systematic investigation of calibration degradation, uncertainty behavior, and model robustness under reproducible non-stationary conditions. Based on the comprehensive reliability evaluation, the framework identifies the most reliable candidate model for subsequent cross-dataset applicability assessment.

3.1.2. Alzheimer's Disease Clinical Dataset

The Alzheimer's Disease Clinical Dataset (<https://www.kaggle.com/datasets/rabieelkharoua/alzheimers-disease-dataset> (accessed on 1 May 2026)) consists of 2149 individual records described by 35 demographic, lifestyle, medical, and cognitive attributes. The features include age, gender, education level, comorbidity indicators, lifestyle factors, and functional assessment measures, while the binary target variable indicates the presence or absence of Alzheimer's disease. The dataset contains no missing values and is provided in structured CSV format.

Unlike the DARWIN dataset, no artificial distributional perturbations are introduced in this dataset. Instead, the clinical dataset represents a substantially different data modality and feature space, incorporating demographic, lifestyle, medical, and cognitive characteristics. It therefore provides a complementary evaluation setting for examining whether the proposed reliability-aware evaluation strategy remains informative when applied to a different representation of Alzheimer's disease-related information.

Within the proposed evaluation strategy, this clinical dataset is used as a second-dataset evaluation to assess the cross-dataset applicability of the proposed reliability-aware strategy. Because the clinical and DARWIN datasets have fundamentally different feature spaces and data modalities, no feature-level transfer is performed and the model trained on the DARWIN dataset is not directly evaluated on the clinical dataset. Instead, the STACK_CAL architecture identified through the reliability-aware evaluation on the DARWIN dataset is independently retrained from scratch using the clinical dataset, followed by the same overall reliability-aware evaluation procedure. This experiment therefore examines whether key components of the proposed strategy, including probability calibration, uncertainty and confidence analysis, threshold stability, robustness assessment, and clinical decision analysis, remain informative when the framework is reapplied to a substantially

different clinical data modality. Accordingly, the results are interpreted as evidence of the cross-dataset applicability of the proposed reliability evaluation strategy, rather than as external validation or direct generalization of a handwriting-trained model.

3.2. Feature Processing

Handwriting-derived features were extracted from digitized samples, capturing spatial, temporal, and dynamic characteristics associated with cognitive impairment. Owing to the heterogeneous scales of the extracted features, Z-score normalization was applied to standardize feature distributions, ensuring numerical stability, preventing scale dominance during model training, and facilitating fair comparison across different machine learning classifiers and ensemble models.

3.2.1. Normalization

All continuous features were standardized using z-score normalization:

$$x' = \frac{x - \mu}{\sigma},$$

where μ and σ denote the mean and standard deviation computed from the training set. Standardization prevents dominance of large-scale features and improves convergence behavior in distance-based and gradient-based classifiers.

3.2.2. Feature Selection (If Applicable)

To reduce redundancy and improve generalization, feature selection was optionally applied using statistically grounded or wrapper-based approaches. Techniques such as Recursive Feature Elimination (RFE) or univariate filtering were employed when specified. Feature selection was conducted strictly within cross-validation folds to prevent information leakage. The selected subset aimed to preserve discriminative power while enhancing model stability and interpretability.

3.3. Base Classifiers and Machine Learning Algorithms

A diverse set of supervised machine learning classifiers was employed to capture complementary decision boundaries, learning biases, and probabilistic characteristics. The selected models encompass tree-based ensembles, linear discriminative methods, probabilistic classifiers, distance-based learning, margin-based classifiers, and neural networks, enabling comprehensive evaluation of discrimination performance, probability calibration, and uncertainty behavior across structurally distinct learning paradigms.

The evaluated classifiers include:

- Extreme Gradient Boosting (XGB)
- Random Forest (RF)
- Extra Trees (ET)
- Decision Tree (DT)
- Support Vector Machine (SVM)
- Gaussian Naïve Bayes (GNB)
- Linear Discriminant Analysis (LDA)
- k-Nearest Neighbors (KNN)
- Multilayer Perceptron (MLP)

Fixed classifier configurations were used consistently across all experiments. Unless otherwise specified, the classifiers were initialized using the implementation-default hyperparameter values, with the principal configurations of all evaluated classifiers explicitly reported in Table 2 (Default hyperparameter configurations of the evaluated classifiers).

No hyperparameter tuning, cross-validated grid search, or inner optimization loop was performed, as the objective was to ensure a consistent and reproducible comparison among classifiers without introducing an additional nested optimization layer into the repeated reliability evaluation. The same fixed configurations were maintained across all runs, while the repeated stratified 10-fold cross-validation with 5 repetitions (10-fold $\times$ 5) was used exclusively for performance and reliability estimation rather than hyperparameter selection. A fixed random seed of 42 was used for stochastic operations to further improve reproducibility. The relevant implementation and experimental details are provided to facilitate independent reproduction of the reported results.

Table 2. Default hyperparameter configurations of the evaluated classifiers.

Classifier	Hyperparameter	Default Setting
Logistic Regression (LR)	penalty	l2
	C	1.0
	solver	lbfgs
	max_iter	100
Linear Discriminant Analysis (LDA)	solver shrinkage	svd None
Gaussian Naive Bayes (GNB)	var_smoothing	1×10^{-9}
Decision Tree (DT)	criterion	gini
	splitter	best
	max_depth	None
	min_samples_split	2
Random Forest (RF)	n_estimators	100
	criterion	gini
	max_depth	None
	min_samples_split	2
	min_samples_leaf	1
	max_features	sqrt
Extra Trees (ET)	n_estimators	100
	criterion	gini
	max_depth	None
	min_samples_split	2
	min_samples_leaf	1
	max_features	sqrt
k-Nearest Neighbors (KNN)	n_neighbors	5
	weights	uniform
	algorithm	auto
	metric	minkowski
	p	2
Support Vector Machine (SVM)	C	1.0
	kernel	rbf
	gamma	scale
Multilayer Perceptron (MLP)	hidden_layer_sizes	(100)
	activation	relu
	solver	adam
	alpha	0.0001
	max_iter	200
Extreme Gradient Boosting (XGB)	n_estimators	100
	max_depth	6
	learning_rate	0.3
	subsample	1.0
	colsample_bytree	1.0

3.3.1. Decision Tree (DT)

Decision Tree recursively partitions the feature space by selecting splits that maximize impurity reduction using criteria such as Gini impurity or information gain.

3.3.2. Random Forest (RF)

Random Forest constructs an ensemble of decision trees using bootstrap aggregation and random feature selection. Final predictions are obtained through majority voting for classification or probability averaging, improving robustness and reducing variance.

3.3.3. Extra Trees (ET)

Extra Trees extends Random Forest by introducing additional randomness through randomly selected split thresholds, increasing ensemble diversity and influencing probability calibration characteristics.

3.3.4. Extreme Gradient Boosting (XGB)

XGBoost is a gradient boosting framework that sequentially constructs decision trees by minimizing a regularized objective function:

$$\text{Obj}(\Theta) = \sum_{i=1}^n \mathcal{L}(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (1)$$

where $\mathcal{L}$ denotes the prediction loss and $\Omega(f_k)$ is the regularization term controlling model complexity.

3.3.5. Linear Discriminant Analysis (LDA)

LDA projects the original feature space onto discriminative directions that maximize between-class separation:

$$\mathbf{y} = \mathbf{W}^T \mathbf{x} \quad (2)$$

where $\mathbf{W}$ represents the projection matrix.

3.3.6. Gaussian Naïve Bayes (GNB)

Gaussian Naïve Bayes estimates posterior probabilities using Bayes' theorem under the conditional independence assumption:

$$P(y|\mathbf{x}) \propto P(y) \prod_{i=1}^n P(x_i|y) \quad (3)$$

3.3.7. K-Nearest Neighbors (KNN)

KNN predicts the class label according to the majority class among the k nearest neighbors, using a distance metric such as Euclidean distance. Class probabilities are estimated from the class proportions within the neighborhood.

3.3.8. Support Vector Machine (SVM)

Support Vector Machine identifies the maximum-margin separating hyperplane by solving

$$\begin{aligned} \max \quad & \frac{2}{\|\mathbf{w}\|} \\ \text{subject to} \quad & y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad i = 1, \dots, n. \end{aligned} \quad (4)$$

Probability estimates are obtained using post hoc Platt (sigmoid) scaling.

3.3.9. Multilayer Perceptron (MLP)

MLP is a feedforward artificial neural network trained using backpropagation. Each neuron computes

$$z_j^{(l)} = \sum_i w_{ji}^{(l)} a_i^{(l-1)}, \quad (5)$$

followed by a nonlinear activation

$$a_j^{(l)} = \sigma(z_j^{(l)}), \quad (6)$$

where $\sigma(\cdot)$ denotes the activation function.

3.3.10. Ensemble Learning Configurations

Based on the discrimination performance of the individual classifiers, the best-performing models (XGB, RF, ET, SVM, GNB, LDA, KNN, and MLP) were selected to construct ensemble models. Both soft-voting and stacking strategies were investigated to analyze the impact of probability calibration on discrimination performance, probabilistic reliability, uncertainty estimation, and overall model trustworthiness. Five ensemble configurations were evaluated:

- BASE_UNCAL: An uncalibrated soft-voting ensemble that directly aggregates posterior probabilities from the selected base classifiers using probability averaging.
- BASE_CAL: A calibrated soft-voting ensemble in which each base classifier is first calibrated using Platt (sigmoid) scaling before probability aggregation.
- STACK_UNCAL: A conventional stacking ensemble that uses probability outputs from uncalibrated base classifiers as meta-features for a Logistic Regression meta-classifier.
- STACK_CAL: A calibration-aware stacking ensemble in which the base classifiers are individually calibrated using Platt scaling before their calibrated probabilities are used to train the Logistic Regression meta-classifier.
- META_CAL: A stacking ensemble constructed from uncalibrated base learners, followed by post hoc Platt calibration applied only to the meta-classifier output probabilities.

These ensemble configurations enable systematic investigation of how calibration placement influences ensemble discrimination, probability calibration, uncertainty estimation, and overall predictive reliability within the proposed reliability-aware evaluation framework.

3.4. Calibration Module

Accurate classification alone is insufficient for clinical deployment; predictive probabilities must also be statistically reliable. A classifier is said to be perfectly calibrated if

$$\mathbb{P}(Y = 1 \mid \hat{P} = p) = p, \quad \forall p \in [0, 1], \quad (7)$$

where $\hat{P}$ denotes the predicted posterior probability. In practice, modern machine learning models—particularly ensemble and deep architectures—tend to produce overconfident or underconfident probability estimates. To enforce probabilistic reliability, post hoc calibration techniques were applied to the predicted probabilities of each base classifier.

Let z denote the uncalibrated score (logit or decision value) produced by a classifier. A calibration function $g(\cdot)$ maps z (or $\hat{P}$) to a calibrated probability:

$$\tilde{P} = g(z). \quad (8)$$

The following calibration methods were evaluated.

3.4.1. Temperature Scaling

Temperature scaling rescales logits using a single scalar parameter $T > 0$:

$$\tilde{P} = \sigma\left(\frac{z}{T}\right), \quad (9)$$

where $\sigma(\cdot)$ is the sigmoid function (for binary classification). The parameter T is optimized by minimizing negative log-likelihood on a validation set. Temperature scaling preserves decision boundaries while adjusting confidence levels.

3.4.2. Platt Scaling

Platt scaling fits a logistic regression model on the uncalibrated scores:

$$\tilde{P} = \frac{1}{1 + \exp(Az + B)}, \quad (10)$$

where A and B are learned parameters. This parametric sigmoid transformation corrects systematic bias in decision scores and is particularly effective for margin-based classifiers.

3.4.3. Isotonic Regression

Isotonic regression is a non-parametric calibration method that learns a monotonic mapping:

$$\tilde{P} = g(z), \quad (11)$$

subject to the constraint:

$$z_i \leq z_j \Rightarrow g(z_i) \leq g(z_j). \quad (12)$$

The function $g(\cdot)$ is obtained by minimizing squared error on the validation set. Unlike parametric methods, isotonic regression does not assume a sigmoid form and can model complex calibration distortions, though it may overfit under limited data.

3.4.4. Beta Calibration

Beta calibration generalizes Platt scaling by modeling the posterior using a Beta distribution-inspired transformation:

$$\tilde{P} = \frac{1}{1 + \exp(A \log \hat{P} + B \log(1 - \hat{P}) + C)}, \quad (13)$$

where A , B , and C are learned parameters. This formulation provides greater flexibility than standard logistic calibration and better accommodates asymmetric miscalibration patterns.

3.4.5. Calibration Evaluation Metrics

Calibration performance was assessed using strictly proper scoring rules and bin-based reliability metrics.

Brier Score

The Brier Score measures mean squared error between predicted probabilities and true labels:

$$\text{BS} = \frac{1}{N} \sum_{i=1}^N (\hat{P}_i - y_i)^2. \quad (14)$$

Lower values indicate better calibration and discrimination jointly.

Logarithmic Loss

Log loss (negative log-likelihood) evaluates probabilistic sharpness:

$$\text{LogLoss} = -\frac{1}{N} \sum_{i=1}^N [y_i \log \hat{P}_i + (1 - y_i) \log(1 - \hat{P}_i)]. \quad (15)$$

Expected Calibration Error (ECE)

ECE measures average discrepancy between confidence and empirical accuracy across M bins:

$$\text{ECE} = \sum_{m=1}^M \frac{|B_m|}{N} |\text{acc}(B_m) - \text{conf}(B_m)|, \quad (16)$$

where B_m denotes the set of samples in bin m .

Maximum Calibration Error (MCE)

MCE captures worst-case bin deviation:

$$\text{MCE} = \max_{m \in \{1, \dots, M\}} |\text{acc}(B_m) - \text{conf}(B_m)|. \quad (17)$$

3.4.6. Calibration Gain

To quantify improvement over uncalibrated baselines, calibration gain was computed as the relative reduction in calibration error:

$$\text{Gain} = \frac{\text{Metric}_{\text{uncal}} - \text{Metric}_{\text{cal}}}{\text{Metric}_{\text{uncal}}}. \quad (18)$$

This measure enables structured comparison of calibration effectiveness across classifiers and drift conditions.

Collectively, this calibration module allows systematic investigation of probability reliability, structural miscalibration patterns, and robustness of ensemble predictions under both controlled and real-world distributional variability.

3.5. Uncertainty Module

Reliable clinical machine learning systems require not only high predictive accuracy but also reliable characterization of prediction confidence and the ability to identify cases for which automated predictions may be less reliable. In this study, uncertainty is assessed through probability-based confidence analysis and selective prediction rather than through formal Bayesian uncertainty estimation. The module therefore focuses on confidence distribution, confidence–error alignment, confidence-based rejection, and selective prediction performance.

Let $\hat{P}_i \in [0, 1]$ denote the predicted posterior probability of the positive class for sample i , and let the predicted class label be defined as

$$\hat{y}_i = \mathbf{1}(\hat{P}_i \geq 0.5). \quad (19)$$

3.5.1. Prediction Confidence and Confidence Margin

Prediction confidence is defined as the probability assigned to the predicted class:

$$c_i = \max(\hat{P}_i, 1 - \hat{P}_i), \quad (20)$$

where $c_i \in [0.5, 1]$. A value close to 0.5 indicates an ambiguous prediction, whereas a value close to 1 indicates high confidence in the predicted class.

For binary classification, the distance of the predicted probability from the decision boundary at 0.5 is defined separately as the confidence margin:

$$m_i = |\hat{P}_i - 0.5|, \quad (21)$$

where $m_i \in [0, 0.5]$. The prediction confidence and confidence margin are related by

$$c_i = 0.5 + m_i. \quad (22)$$

Thus, c_i and m_i induce exactly the same ranking of binary predictions, although they use different numerical scales. Throughout this study, c_i denotes prediction confidence, whereas m_i denotes the confidence margin, i.e., the distance from the decision boundary. In the implementation of the selective prediction experiments, the confidence margin $m_i = |\hat{P}_i - 0.5|$ is used to rank predictions. Because of the monotonic relationship $c_i = 0.5 + m_i$, this produces the same ordering as ranking predictions by c_i .

3.5.2. Confidence Distribution Analysis

The empirical distribution of prediction confidence values $\{c_i\}_{i=1}^N$ was analyzed to characterize the concentration and sharpness of model predictions and to identify potential overconfidence. Let $f(c)$ denote the empirical density of confidence values. A high concentration of predictions near $c = 1$ indicates that the model frequently assigns high confidence to its predictions. Such concentration is not necessarily desirable and should be interpreted jointly with prediction correctness and calibration measures. In particular, a high density of confident predictions accompanied by a substantial error rate may indicate overconfident or poorly calibrated behavior.

Class-conditional confidence distributions were additionally examined as

$$f_k(c) = p(c \mid Y = k), \quad k \in \{0, 1\}, \quad (23)$$

to characterize differences in confidence behavior between the Alzheimer's disease and control classes. These distributions provide an additional descriptive view of confidence concentration and class-dependent uncertainty patterns.

3.5.3. Accuracy–Coverage Analysis

Selective prediction performance was evaluated by progressively rejecting the most uncertain predictions and measuring the accuracy of the predictions retained for automated classification. For each prediction, the confidence margin was defined as

$$m_i = |\hat{P}_i - 0.5|. \quad (24)$$

Smaller values of m_i indicate that the predicted probability is closer to the decision boundary at 0.5 and therefore represent greater predictive uncertainty, whereas larger values indicate greater separation from the decision boundary. Since the prediction confidence is defined as $c_i = \max(\hat{P}_i, 1 - \hat{P}_i) = 0.5 + m_i$, ranking predictions according to m_i is equivalent to ranking them according to prediction confidence c_i .

For a given rejection proportion $r \in [0, 0.5]$, a rejection threshold q_r is obtained from the r -th percentile of the confidence margins:

$$q_r = \text{Percentile}_{100r}(\{m_i\}_{i=1}^N). \quad (25)$$

The retained prediction set is then defined as

$$\mathcal{S}_r = \{i \mid m_i > q_r\}, \quad (26)$$

and the corresponding coverage is calculated as

$$\text{Coverage}(r) = \frac{|\mathcal{S}_r|}{N}. \quad (27)$$

Thus, r represents the proportion of predictions with the smallest confidence margins targeted for rejection, rather than a numerical confidence threshold. For example, $r = 0.20$ corresponds to rejecting approximately the lowest 20% of predictions according to their confidence margin and retaining approximately 80% of the predictions for automated classification. Accordingly, the term “least confident predictions” throughout this analysis refers to predictions identified through the confidence-margin ranking and does not denote a confidence value of 20%.

The selective accuracy of the retained prediction set is calculated as

$$\text{Acc}_{\text{sel}}(r) = \frac{\sum_{i \in \mathcal{S}_r} \mathbf{1}(\hat{y}_i = y_i)}{|\mathcal{S}_r|}. \quad (28)$$

Accuracy–coverage curves plot $\text{Acc}_{\text{sel}}(r)$ against $\text{Coverage}(r)$ as increasingly uncertain predictions are rejected. A desirable selective predictor should improve predictive accuracy while retaining a substantial proportion of the predictions for automated classification. The implementation evaluates rejection proportions from 0% to 50% in increments of 5%.

3.5.4. Selective Classification and Abstention Mechanism

Confidence-based rejection was used to evaluate whether abstaining from low-confidence predictions improves the reliability of the retained prediction set. The acceptance region for a given rejection proportion r is defined as

$$\mathcal{A}_r = \{i \mid m_i > q_r\}, \quad (29)$$

where q_r is the corresponding percentile-based confidence-margin threshold. The rejection region is therefore

$$\mathcal{R}_r = \{1, \dots, N\} \setminus \mathcal{A}_r. \quad (30)$$

The selective accuracy is then

$$\text{Acc}_{\text{sel}}(r) = \frac{\sum_{i \in \mathcal{A}_r} \mathbf{1}(\hat{y}_i = y_i)}{|\mathcal{A}_r|}. \quad (31)$$

The improvement associated with selective prediction can be expressed by comparing the accuracy before rejection,

$$\text{Acc}_{\text{all}} = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(\hat{y}_i = y_i), \quad (32)$$

with the accuracy after rejection,

$$\text{Acc}_{\text{sel}}(r). \quad (33)$$

The corresponding selective accuracy improvement is therefore

$$\Delta \text{Acc}(r) = \text{Acc}_{\text{sel}}(r) - \text{Acc}_{\text{all}}. \quad (34)$$

A positive $\Delta\text{Acc}(r)$ indicates that rejecting the lowest-margin predictions improves the accuracy of the retained prediction set.

For fixed-percentage rejection experiments, the term “ $q\%$ least confident predictions” refers to the lowest $q\%$ of predictions ranked according to the confidence margin $m_i = |\hat{P}_i - 0.5|$. Because $c_i = 0.5 + m_i$, this ranking is equivalent to ranking predictions by prediction confidence c_i . The value $q\%$ therefore denotes the proportion of predictions rejected and does not represent a numerical confidence threshold.

Overall, the uncertainty module evaluates three complementary properties: (i) the concentration and sharpness of prediction confidence; (ii) the alignment between confidence and prediction correctness; (iii) the change in accuracy and retained coverage achieved by abstaining from low-confidence predictions. Together with the calibration analyses, these measures provide confidence-based reliability and selective-prediction diagnostics beyond conventional discrimination metrics.

3.6. Clinical Decision Module

While discrimination, calibration, and uncertainty characterize statistical reliability, clinical deployment requires explicit decision-theoretic evaluation. A predictive model must support threshold selection consistent with clinical objectives such as screening sensitivity, risk stratification, or cost-sensitive intervention policies.

Let $\hat{P}_i$ denote the calibrated posterior probability for subject i . A binary decision rule is defined as

$$\hat{y}_i(p_t) = \mathbf{1}(\hat{P}_i \geq p_t), \quad (35)$$

where $p_t \in (0, 1)$ is the clinical decision threshold.

3.6.1. Threshold Optimization

Optimal classification thresholds were determined using clinically interpretable criteria.

Youden’s Index

Youden’s Index identifies the threshold that maximizes the trade-off between sensitivity and specificity:

$$J(p_t) = \text{Sensitivity}(p_t) + \text{Specificity}(p_t) - 1. \quad (36)$$

The optimal threshold is

$$p_t^* = \arg \max_{p_t} J(p_t). \quad (37)$$

Sensitivity-Priority Threshold

For screening-oriented deployment where false negatives are clinically costly, thresholds were selected to satisfy:

$$\text{Sensitivity}(p_t) \geq \alpha, \quad (38)$$

where α is a predefined minimum acceptable sensitivity level (e.g., $\alpha = 0.90$).

Cost-Sensitive Decision Criterion

Let C_{FN} and C_{FP} denote the costs of false negatives and false positives, respectively. The expected decision risk is

$$\mathcal{R}(p_t) = C_{FN} \cdot \text{FN}(p_t) + C_{FP} \cdot \text{FP}(p_t). \quad (39)$$

The cost-optimal threshold is obtained by

$$p_t^* = \arg \min_{p_t} \mathcal{R}(p_t). \quad (40)$$

Under Bayesian decision theory, the theoretically optimal threshold satisfies

$$p_t^* = \frac{C_{FP}}{C_{FP} + C_{FN}}. \quad (41)$$

This enables adaptation of the predictive model to distinct clinical workflows and risk tolerances.

3.6.2. Net Benefit and Decision Curve Analysis

To quantify real-world clinical utility, Decision Curve Analysis (DCA) was performed across a continuum of threshold probabilities p_t . Net Benefit (NB) is defined as

$$\text{NB}(p_t) = \frac{TP(p_t)}{N} - \frac{FP(p_t)}{N} \cdot \frac{p_t}{1 - p_t}, \quad (42)$$

where $TP(p_t)$ and $FP(p_t)$ are the numbers of true positives and false positives at threshold p_t , and N is the total sample size.

The term $\frac{p_t}{1 - p_t}$ represents the implicit harm-to-benefit ratio associated with the chosen decision threshold. DCA evaluates whether the predictive model yields greater net benefit than two default strategies:

$$\text{NB}_{\text{treat-all}}(p_t) = \pi - (1 - \pi) \frac{p_t}{1 - p_t}, \quad (43)$$

$$\text{NB}_{\text{treat-none}}(p_t) = 0, \quad (44)$$

where π denotes disease prevalence.

A clinically useful model must satisfy

$$\text{NB}_{\text{model}}(p_t) > \max\{\text{NB}_{\text{treat-all}}(p_t), \text{NB}_{\text{treat-none}}(p_t)\} \quad (45)$$

over a meaningful threshold range.

For graphical interpretation, the treat-all and treat-none strategies were plotted alongside each model's decision curve to provide the standard reference points for assessing clinical net benefit.

3.6.3. Reliability-to-Decision Bridging

Importantly, threshold-based decisions depend directly on probability calibration and uncertainty ranking. Miscalibrated probabilities distort the effective risk interpretation of p_t , potentially leading to suboptimal intervention decisions. By integrating threshold optimization, cost-sensitive analysis, and decision curve evaluation, this module formally bridges probabilistic reliability with actionable clinical utility, ensuring that statistical improvements translate into meaningful patient-level benefit.

4. Results

This section presents a comprehensive experimental validation of the proposed reliability-aware evaluation framework for early Alzheimer's disease (AD) detection by using handwriting biomarkers. Unlike conventional studies that primarily emphasize discrimination performance, the proposed framework systematically evaluates multiple complementary dimensions of model reliability, including statistical validation, probability

calibration, uncertainty-aware evaluation, robustness under distributional perturbations, and clinical decision utility. The objective is not only to identify highly discriminative models but also to determine whether their probabilistic predictions are reliable, stable, and suitable for risk-informed clinical decision support.

All experiments were conducted using 10-fold stratified cross-validation with five repetitions ($10 \times 5 = 50$ folds) to obtain robust and unbiased performance estimates while preserving the class distribution across evaluation folds. Within each fold, all data-dependent preprocessing, model fitting, and probability calibration procedures were performed using the corresponding training data, while the held-out test fold was used exclusively for final evaluation. Discrimination performance was evaluated using Accuracy, Precision, Recall, F1-score, ROC-AUC, PR-AUC, Cohen's Kappa, and Matthews Correlation Coefficient (MCC). Statistical significance was assessed using the Friedman test, Nemenyi post hoc analysis, Wilcoxon signed-rank test, and Kendall's coefficient of concordance to evaluate differences and consistency in model rankings across repeated experiments. Probability calibration was evaluated using the Brier Score, LogLoss, Expected Calibration Error (ECE), Maximum Calibration Error (MCE), and reliability diagrams.

To further assess model trustworthiness, uncertainty-aware evaluation was performed through confidence analysis, confidence–accuracy consistency, risk–coverage analysis, and selective prediction. Robustness was investigated using controlled perturbation experiments, including Gaussian noise injection, feature perturbation, and reduced-data simulations to evaluate model stability under controlled distributional changes. Finally, clinical applicability was examined through threshold optimization and decision curve analysis to determine the practical utility of calibrated probability estimates for risk-informed clinical decision support.

The experimental results are organized into the following subsections: (1) discrimination and statistical performance analysis; (2) probability calibration and reliability assessment; (3) uncertainty-aware evaluation; (4) robustness assessment under distributional perturbations; (5) clinical decision analysis.

4.1. Baseline Performance Evaluation

This experiment employs a repeated stratified K-fold cross-validation framework with 10 folds and 5 repetitions (10×5), resulting in 50 held-out test-fold evaluations. The stratification preserves the class distribution across splits and provides repeated estimates of model performance. All classifiers were evaluated using fixed hyperparameter configurations, with implementation-default values used unless otherwise specified. No inner hyperparameter optimization, grid search, or model-specific parameter tuning was performed. Thus, the 10-fold, 5-repeat stratified cross-validation was used solely for performance estimation rather than for hyperparameter selection. All reported performance and reliability metrics—including Accuracy, Precision, Recall, Specificity, F1-score, ROC-AUC, PR-AUC, Cohen's Kappa, MCC, Brier Score, Expected Calibration Error (ECE), and Log Loss—are computed exclusively from predictions on the corresponding held-out test folds to avoid information leakage and obtain unbiased performance estimates.

4.1.1. Comparative Performance Analysis

Table 3 presents the performance of ten classifiers evaluated under the repeated stratified 10×5 cross-validation framework. All metrics are reported as mean $\pm$ standard deviation across the 50 held-out test folds. Complete results are presented in Supplementary Material File S1.

Table 3. Performance comparison of ten classifiers (Mean $\pm$ Std). Bold values in each column indicate the maximum values.

Model	Acc	F1	ROC-AUC	PR-AUC	Kappa	MCC	Brier	LogLoss
SVM	0.8758 $\pm$ 0.0763	0.8825 $\pm$ 0.0719	0.9550 $\pm$ 0.0462	0.9570 $\pm$ 0.0475	0.7511 $\pm$ 0.1530	0.7580 $\pm$ 0.1520	0.0845 $\pm$ 0.0407	0.2719 $\pm$ 0.1139
ET	0.8778 $\pm$ 0.0784	0.8865 $\pm$ 0.0693	0.9467 $\pm$ 0.0451	0.9490 $\pm$ 0.0465	0.7552 $\pm$ 0.1568	0.7629 $\pm$ 0.1524	0.1197 $\pm$ 0.0210	0.4003 $\pm$ 0.0474
RF	0.8625 $\pm$ 0.0711	0.8727 $\pm$ 0.0645	0.9466 $\pm$ 0.0461	0.9495 $\pm$ 0.0497	0.7245 $\pm$ 0.1423	0.7381 $\pm$ 0.1392	0.1191 $\pm$ 0.0225	0.3975 $\pm$ 0.0497
XGB	0.8547 $\pm$ 0.0694	0.8601 $\pm$ 0.0652	0.9337 $\pm$ 0.0535	0.9467 $\pm$ 0.0460	0.7093 $\pm$ 0.1390	0.7186 $\pm$ 0.1377	0.1098 $\pm$ 0.0518	0.3475 $\pm$ 0.1542
MLP	0.8317 $\pm$ 0.0456	0.8293 $\pm$ 0.0488	0.9273 $\pm$ 0.0535	0.9354 $\pm$ 0.0487	0.6639 $\pm$ 0.0911	0.6741 $\pm$ 0.0933	0.1184 $\pm$ 0.0414	0.3865 $\pm$ 0.1699
LR	0.8144 $\pm$ 0.0831	0.8085 $\pm$ 0.0920	0.9013 $\pm$ 0.0738	0.9108 $\pm$ 0.0703	0.6289 $\pm$ 0.1669	0.6380 $\pm$ 0.1671	0.1383 $\pm$ 0.0567	0.5026 $\pm$ 0.2294
GNB	0.8294 $\pm$ 0.1042	0.8427 $\pm$ 0.0926	0.8948 $\pm$ 0.0879	0.8681 $\pm$ 0.1054	0.6579 $\pm$ 0.2100	0.6666 $\pm$ 0.2070	0.1604 $\pm$ 0.0960	3.1437 $\pm$ 2.1096
KNN	0.6164 $\pm$ 0.0808	0.3728 $\pm$ 0.2115	0.8229 $\pm$ 0.0776	0.8247 $\pm$ 0.0829	0.2453 $\pm$ 0.1551	0.3453 $\pm$ 0.1720	0.2770 $\pm$ 0.0626	5.5118 $\pm$ 2.3301
DT	0.7448 $\pm$ 0.0837	0.7341 $\pm$ 0.0999	0.7449 $\pm$ 0.0839	0.7043 $\pm$ 0.0835	0.4891 $\pm$ 0.1680	0.4984 $\pm$ 0.1683	0.2552 $\pm$ 0.0837	8.8154 $\pm$ 2.8894
LDA	0.7011 $\pm$ 0.1120	0.6819 $\pm$ 0.1368	0.7499 $\pm$ 0.1347	0.8033 $\pm$ 0.0975	0.4042 $\pm$ 0.2225	0.4156 $\pm$ 0.2262	0.2810 $\pm$ 0.1103	3.6848 $\pm$ 2.3788

4.1.2. Statistical Significance Analysis for Baseline Classifiers

To determine whether performance differences among classifiers were statistically significant, a Friedman test (Table 4) was conducted using fold-wise ROC-AUC scores from the 50 held-out test folds generated by the repeated stratified 10×5 cross-validation framework.

Table 4. Friedman Test Results (ROC-AUC).

Friedman Statistic	166.7234
<i>p</i> -value	<0.001

The Friedman test indicates a statistically significant difference among the evaluated classifiers ($p < 0.001$). Therefore, a Nemenyi post hoc analysis was performed to identify pairwise differences.

Average rank analysis revealed that SVM (2.80), RF (3.20), ET (3.32), XGB (4.05), and MLP (4.50) form a statistically homogeneous top-performing group at $\alpha = 0.05$, with no significant pairwise differences observed among them. In contrast, DT (9.25), LDA (8.65), and KNN (7.62) were significantly inferior to the leading models.

4.1.3. Wilcoxon Signed-Rank Test

To further compare the two highest-ranked models, SVM and ET, a Wilcoxon signed-rank test was conducted using their fold-wise ROC-AUC scores. The test yielded a statistic of 107.00 with a *p*-value of 0.1352. Since $p \geq 0.05$, the difference between SVM and ET is not statistically significant. This confirms that although SVM achieved the best mean ROC-AUC and average rank, its performance is statistically comparable to ET under repeated stratified evaluation.

4.1.4. Critical Difference Diagram Analysis

Figure 2 presents the Critical Difference (CD) diagram based on the Nemenyi post hoc test applied to ROC-AUC scores across the repeated stratified cross-validation folds. Lower average ranks indicate superior performance.

SVM achieved the best average rank (2.80), followed closely by RF (3.20), ET (3.32), XGB (4.05), and MLP (4.50). The horizontal connecting bars in the diagram indicate that these five models are not significantly different at $\alpha = 0.05$, forming a competitive top-performing cluster.

LR (5.72) and GNB (5.90) occupy a middle-performance tier, while KNN (7.62), LDA (8.65), and DT (9.25) demonstrate substantially inferior ranks. The absence of connecting bars between the top cluster and the lowest-ranked models indicates statistically significant performance degradation for these classifiers.

Overall, the CD diagram confirms that margin-based (SVM) and ensemble-based methods (RF, ET, XGB) consistently outperform single-tree and distance-based classifiers in terms of discriminative performance. Furthermore, the small rank differences within

the top-performing cluster suggest strong stability and competitive equivalence among advanced learning paradigms under repeated stratified evaluation.

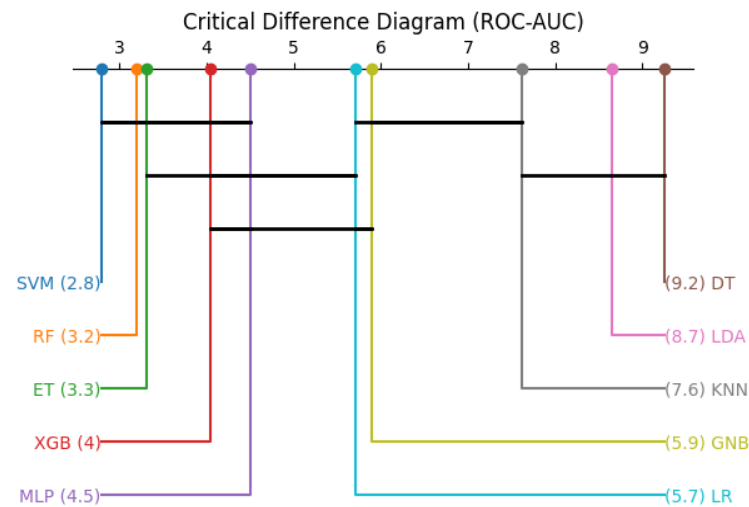


Figure 2. Critical Difference (CD) diagram based on the Nemenyi post hoc test applied to ROC-AUC scores across 10×5 Repeated Stratified K-Fold cross-validation. Lower average ranks indicate better performance. Models connected by a horizontal line are not significantly different at $\alpha = 0.05$.

Overall, the baseline discrimination analysis demonstrates that advanced ensemble and margin-based classifiers consistently achieve superior predictive performance under repeated stratified cross-validation, with statistical analyses confirming no significant differences among the top-performing models. These findings establish a statistically robust baseline and identify the most competitive candidate models for subsequent reliability-aware evaluation. The following experiments therefore extend beyond conventional discrimination assessment to investigate probability calibration, uncertainty quantification, robustness under distributional perturbations, and clinical decision utility, providing a comprehensive evaluation of model trustworthiness for Alzheimer's disease detection.

4.2. Calibration Analysis

In this study, we evaluated the calibration performance of ten widely used classifiers, including XGBoost, Random Forest, Extra Trees, Decision Tree, Linear Discriminant Analysis, Support Vector Machine, K-Nearest Neighbors, Gaussian Naive Bayes, Multi-Layer Perceptron, and Logistic Regression. Each model was trained and tested using repeated stratified 10×5 cross-validation on a normalized, filtered dataset for Alzheimer's disease prediction. We compared uncalibrated probability outputs with four post hoc calibration methods—Temperature Scaling, Platt Scaling, Isotonic Regression, and Beta Calibration—using calibration metrics such as Brier score, Expected Calibration Error (ECE), Maximum Calibration Error (MCE), and Log Loss. Statistical analyses, including Wilcoxon signed-rank tests, Friedman tests, and Nemenyi post hoc comparisons, were performed to quantify improvements and assess significant differences among calibration methods. Additionally, reliability diagrams and confidence histograms were generated to visualize calibration quality, and calibration gains (Δ Brier and Δ ECE) were computed to quantify the improvements over uncalibrated models. Complete results are presented in Supplementary Material File S2.

4.2.1. Comprehensive Evaluation of Post Hoc Calibration Methods

The probabilistic predictions of machine learning models often suffer from miscalibration, leading to overconfident or underconfident predictions. In this work, we systematically evaluated four widely used post hoc calibration methods: Temperature Scaling (TEMP), Platt Scaling (PLATT), Isotonic Regression (ISO), and Beta Calibration (BETA) on a diverse set of classifiers including tree-based, linear, kernel-based, probabilistic, and neural network models. Calibration performance was assessed using Brier Score, Expected Calibration Error (ECE), Maximum Calibration Error (MCE), and Logarithmic Loss (LogLoss). Statistical significance was analyzed using Wilcoxon signed-rank tests, Friedman tests, and Nemenyi post hoc tests.

Tree-Based Models. Table 5 presents the mean Brier scores and calibration gain (Δ Brier) for XGB, ET, RF, and DT. Platt scaling consistently reduced the Brier score for ET and DT, while TEMP scaling occasionally increased the Brier score (XGB, RF), indicating that temperature scaling may negatively impact the probabilistic accuracy for some tree ensembles. ISO and BETA showed moderate improvements in XGB and RF.

Table 5. Brier scores and calibration gain (Δ Brier) for tree-based models.

Model	UNCAL	TEMP	PLATT	ISO	BETA	Δ Brier (Best)
XGB	0.1098	0.1379	0.1239	0.1149	0.1148	−0.0281 (TEMP)
ET	0.1197	0.1197	0.0928	0.1197	0.1966	0.0269 (PLATT)
RF	0.1191	0.1297	0.0953	0.0949	0.0955	0.0241 (ISO)
DT	0.2552	0.2552	0.2490	0.2552	0.2552	0.0062 (PLATT)

Table 6 summarizes the ECE and LogLoss. Platt scaling achieved the largest reduction in ECE for ET and DT. TEMP scaling occasionally improved ECE (SVM and XGB) but increased the Brier score, highlighting a trade-off between calibration error and probabilistic accuracy.

Table 6. ECE and LogLoss for tree-based models.

Model	Method	ECE	LogLoss
XGB	UNCAL	0.1590	0.3475
XGB	TEMP	0.1474	1.0685
XGB	PLATT	0.1456	0.4367
XGB	ISO	0.1533	0.9638
XGB	BETA	0.1521	0.3850
ET	UNCAL	0.2180	0.4003
ET	TEMP	0.2180	0.4003
ET	PLATT	0.1605	0.3062
ET	ISO	0.2180	0.4003
ET	BETA	0.3126	0.5847
RF	UNCAL	0.2172	0.3975
RF	TEMP	0.1383	1.0708
RF	PLATT	0.1425	0.2985
RF	ISO	0.1408	0.5247
RF	BETA	0.1519	0.3032
DT	UNCAL	0.2552	8.8154
DT	TEMP	0.2552	8.8043
DT	PLATT	0.2436	1.1277
DT	ISO	0.2552	8.8154
DT	BETA	0.2551	2.3476

Wilcoxon signed-rank tests confirmed that the Brier score improvements by PLATT scaling were statistically significant for all tree-based models ($p < 0.001$), except TEMP in XGB. Friedman tests indicated significant differences across calibration methods ($p < 0.001$), and Nemenyi post hoc analysis highlighted PLATT as superior to UNCAL, ISO, and BETA in ensemble models.

Linear and Probabilistic Models. Linear classifiers (LDA, LR) and probabilistic classifiers (GNB) generally benefited from PLATT and BETA calibration. Table 7 shows that PLATT consistently reduced Brier scores, while TEMP and ISO showed mixed performance. ECE reduction followed a similar pattern, confirming PLATT's reliability for models with well-defined probabilistic assumptions.

Table 7. Brier score and ECE for linear and probabilistic models.

Model	Method	Brier	ECE
LDA	UNCAL	0.2810	0.2946
LDA	TEMP	0.2730	0.2874
LDA	PLATT	0.2601	0.2890
LDA	ISO	0.2817	0.2945
LDA	BETA	0.2765	0.2924
LR	UNCAL	0.1383	0.1688
LR	TEMP	0.1655	0.1808
LR	PLATT	0.1350	0.1694
LR	ISO	0.1378	0.1654
LR	BETA	0.1431	0.1675
GNB	UNCAL	0.1604	0.1672
GNB	TEMP	0.1251	0.1601
GNB	PLATT	0.1407	0.1178
GNB	ISO	0.1220	0.1389
GNB	BETA	0.1210	0.1485

Kernel-Based Models. SVM and KNN displayed heterogeneous behavior. For SVM, TEMP reduced ECE substantially but increased Brier score, indicating a trade-off between calibration accuracy and probabilistic error. KNN, being highly sensitive to local neighborhoods, benefited most from PLATT scaling, reducing both Brier score and ECE. Table 8 summarizes these results.

Table 8. Brier score and ECE for kernel-based models.

Model	Method	Brier	ECE
SVM	UNCAL	0.0845	0.1408
SVM	TEMP	0.1193	0.1149
SVM	PLATT	0.0840	0.1373
SVM	ISO	0.0963	0.1401
SVM	BETA	0.0904	0.1336
KNN	UNCAL	0.2770	0.3288
KNN	TEMP	0.2464	0.2884
KNN	PLATT	0.1813	0.1728
KNN	ISO	0.1937	0.1977
KNN	BETA	0.1958	0.2057

Neural Networks (MLP). MLP models were moderately calibrated in their uncalibrated form (Brier ≈ 0.118). TEMP scaling increased Brier score, while PLATT and ISO achieved minor improvements. Table 9 provides detailed metrics.

Table 9. Brier score and ECE for MLP.

Method	Brier	ECE
UNCAL	0.1184	0.1619
TEMP	0.1453	0.1642
PLATT	0.1335	0.1557
ISO	0.1200	0.1612
BETA	0.1204	0.1646

4.2.2. Average Ranks Across Models

Table 10 shows the average ranks obtained from Friedman tests across all classifiers. PLATT consistently achieved the best rank in the majority of models, followed by BETA and ISO. TEMP scaling exhibited inconsistent behavior, ranking lower for tree ensembles and neural networks.

Table 10. Average ranks of calibration methods (Lower = Better).

Model	UNCAL	TEMP	PLATT	ISO	BETA
XGB	1.6333	4.4667	3.3667	2.9000	2.6333
ET	2.9500	3.0000	1.1000	2.9500	5.0000
RF	4.1000	3.9667	2.3000	2.2333	2.4000
DT	4.1500	3.7000	1.0000	4.1500	2.0000
LDA	3.8667	2.8667	1.7333	3.5000	3.0333
SVM	2.3000	4.2000	2.1000	3.6333	2.7667
KNN	4.7667	3.9333	1.1000	2.5333	2.6667
GNB	4.1333	2.6667	3.3667	2.6333	2.2000
MLP	1.2000	4.8667	3.8000	2.3667	2.7667
LR	2.7000	4.4000	1.6333	2.6667	3.6000

4.2.3. Discussion Summary

- Platt Scaling emerged as the most reliable calibration method for ensemble and non-linear models (ET, DT, KNN, LR), consistently reducing Brier score and ECE.
- Beta Calibration and Isotonic Regression showed moderate gains for probabilistic and linear models, while TEMP scaling had inconsistent effects, sometimes improving ECE but worsening Brier score.
- Neural networks (MLP) required minor calibration, with PLATT and ISO offering marginal improvements.
- Tree-based ensembles (XGB, RF) showed sensitivity to TEMP scaling; careful method selection is recommended depending on deployment requirements.

Overall, model-specific calibration is essential to achieve reliable probabilistic predictions in clinical decision-support systems. Among the evaluated methods, Platt Scaling provided a strong and comparatively consistent calibration approach across the evaluated classifiers; however, the observed model-dependent differences indicate that calibration selection should remain empirically driven.

4.2.4. Confidence–Accuracy Consistency and Calibration Stability Analysis for ET Classifier

We present the probabilistic behavior of the Extra Trees (ET) classifier as an example, demonstrating notable variability across calibration analyses, as illustrated in Figure 3. The Critical Difference diagram (Figure 3a) based on the Brier score indicates that Platt scaling achieves the best average rank, significantly outperforming Beta calibration and marginally surpassing temperature scaling, isotonic regression, and the uncalibrated model. This suggests that a parametric sigmoid-based recalibration is most effective for stabilizing ET's probability estimates. The reliability diagram (Figure 3b) further reveals sub-

stantial bin-level volatility: low-confidence predictions (≈ 0.15 – 0.35) exhibit near-zero empirical accuracy, indicating localized overconfidence in incorrect predictions, while mid-confidence regions show large oscillations around the ideal diagonal, reflecting unstable confidence–accuracy alignment. In contrast, high-confidence predictions (≈ 0.80 and above) approach perfect empirical accuracy, suggesting that extreme-confidence outputs are comparatively trustworthy. The confidence histogram (Figure 3c) indicates that ET predictions are predominantly distributed in low-to-mid confidence regions rather than saturating at extremes, which partially explains the observed instability in intermediate bins. Collectively, these findings indicate that while ET is not globally overconfident, its probabilistic outputs lack smooth calibration consistency, particularly in moderate-confidence regions, and benefit substantially from appropriate parametric calibration such as Platt scaling. All such plots for the other classifiers are presented in Supplementary Material File S2.

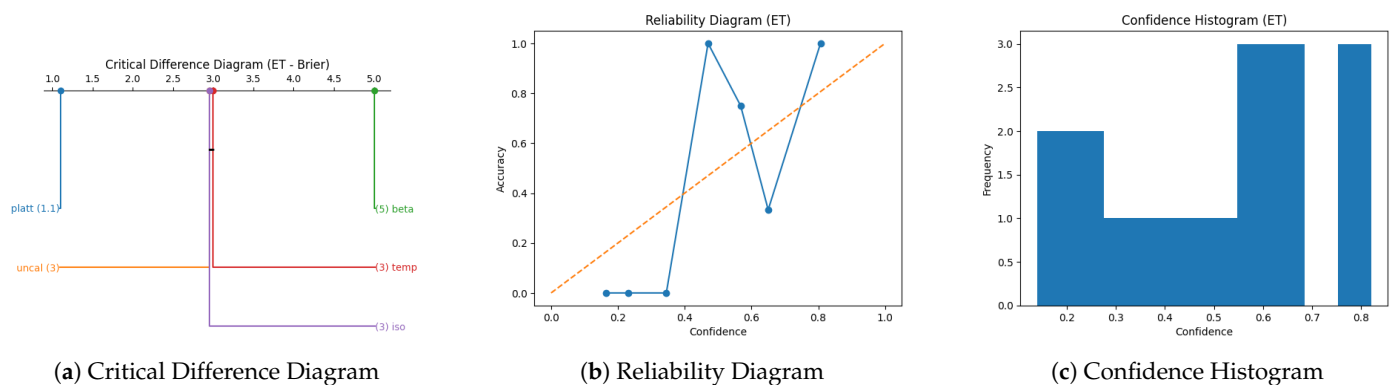


Figure 3. Comprehensive calibration analysis of the Extra Trees (ET) classifier. The Critical Difference diagram presents statistical ranking across calibration methods, the reliability diagram evaluates confidence–accuracy alignment, and the histogram illustrates the distribution of predicted probabilities.

4.3. Confidence Behavior Analysis

To further assess probabilistic reliability beyond global calibration metrics, we performed a confidence–accuracy consistency analysis aggregated across all 50 repeated stratified cross-validation folds. For each of the ten classifiers, predicted probabilities were partitioned into fixed confidence intervals (0.5–0.6, 0.6–0.7, 0.7–0.8, 0.8–0.9, and 0.9–1.0), and empirical accuracy was computed within each bin. This allowed direct evaluation of conditional calibration, specifically testing whether high-confidence predictions corresponded to proportionally high empirical correctness (e.g., whether 90% confidence yielded $\approx 90\%$ accuracy). The analysis was performed for both uncalibrated outputs and beta-calibrated probabilities, with predictions aggregated across all cross-validation folds to ensure statistical robustness. Additionally, a Wilcoxon signed-rank test was conducted on predictions within the highest-confidence bin (0.9–1.0) to statistically assess whether calibration significantly improved reliability in clinically critical high-certainty regions. This bin-wise reliability evaluation provides a more interpretable and clinically meaningful assessment of model trustworthiness than global metrics such as Brier score or Expected Calibration Error alone.

The predictive confidence of ten classifiers—XGB, ET, RF, DT, LDA, SVM, KNN, GNB, MLP, and LR—was evaluated using 10-fold repeated stratified cross-validation (5 repeats). The study compared uncalibrated probabilities with Beta-calibrated probabilities (as an example) across three representative confidence bins: $[0.5, 0.6)$, $[0.7, 0.8)$, and $[0.9, 1.0)$. The aggregated results are summarized in Table 11, while the bin-wise confidence–accuracy relationships for some selected classifiers are illustrated in Figure 4. The complete results are presented in Supplementary Material File S3.

Table 11. Aggregated confidence vs empirical accuracy for uncalibrated and Beta-calibrated models (representative bins). N is the total number of out-of-fold predictions across all folds and repeats. If N = 0, then its corresponding values in other columns become ‘nan’. For instance, the N value for bin 0.5–0.6 for DT is 0 and hence its corresponding UncalConf, UncalAcc, BetaConf and BetaAcc become ‘nan’.

Model	Bin	N	Uncal Conf	Uncal Acc	Beta Conf	Beta Acc
XGB	0.5–0.6	15	0.5633	0.8000	0.5399	0.7143
	0.7–0.8	29	0.7566	0.6207	0.7523	0.7000
	0.9–1.0	182	0.9711	0.9505	0.9823	0.9343
ET	0.5–0.6	75	0.5472	0.6800	0.5445	0.8257
	0.7–0.8	61	0.7411	0.9508	0.7168	1.0000
	0.9–1.0	4	0.9250	1.0000	nan	nan
RF	0.5–0.6	76	0.5541	0.5658	0.5565	0.5862
	0.7–0.8	56	0.7450	0.9286	0.7451	0.7955
	0.9–1.0	10	0.9130	1.0000	0.9617	0.9717
DT	0.5–0.6	0	nan	nan	nan	nan
	0.7–0.8	0	nan	nan	nan	nan
	0.9–1.0	244	1.0000	0.7746	0.9999	0.7746
LDA	0.5–0.6	1	0.5146	0.0000	nan	nan
	0.7–0.8	4	0.7481	0.5000	0.7805	0.4000
	0.9–1.0	220	0.9954	0.7455	0.9879	0.7500
SVM	0.5–0.6	26	0.5562	0.3846	0.5507	0.9000
	0.7–0.8	27	0.7528	0.7778	0.7423	0.8571
	0.9–1.0	155	0.9635	0.9419	0.9727	0.9467
KNN	0.5–0.6	0	nan	nan	nan	nan
	0.7–0.8	0	nan	nan	nan	nan
	0.9–1.0	20	1.0000	1.0000	0.9633	0.9716
GNB	0.5–0.6	1	0.5038	1.0000	0.5546	0.6957
	0.7–0.8	3	0.7648	0.6667	0.7586	0.4211
	0.9–1.0	283	0.9981	0.8127	0.9214	0.8744
MLP	0.5–0.6	17	0.5512	0.7059	0.5544	0.7500
	0.7–0.8	12	0.7563	0.4167	0.7486	0.3636
	0.9–1.0	171	0.9821	0.9181	0.9833	0.9050
LR	0.5–0.6	18	0.5473	0.6667	0.5508	0.6875
	0.7–0.8	11	0.7632	0.8182	0.7440	0.6000
	0.9–1.0	184	0.9840	0.8967	0.9868	0.8794

4.3.1. Calibration Effectiveness

Beta calibration generally adjusts overconfident predictions in the low- to mid-confidence bins (0.5–0.8), aligning predicted probabilities more closely with observed accuracies (Table 11). In high-confidence bins (0.9–1.0), ensemble and tree-based models (XGB, RF) already provide highly reliable predictions, resulting in minor adjustments from calibration.

4.3.2. High-Confidence Bin Analysis

Wilcoxon signed-rank tests were performed on samples in the 0.9–1.0 bin to compare uncalibrated and Beta-calibrated accuracies. Most models show no significant difference (e.g., XGB: $p = 0.637$, SVM: $p = 0.808$), confirming that Beta calibration preserves the reliability of high-confidence predictions. Models with insufficient high-confidence samples (ET, KNN) were excluded from statistical testing.

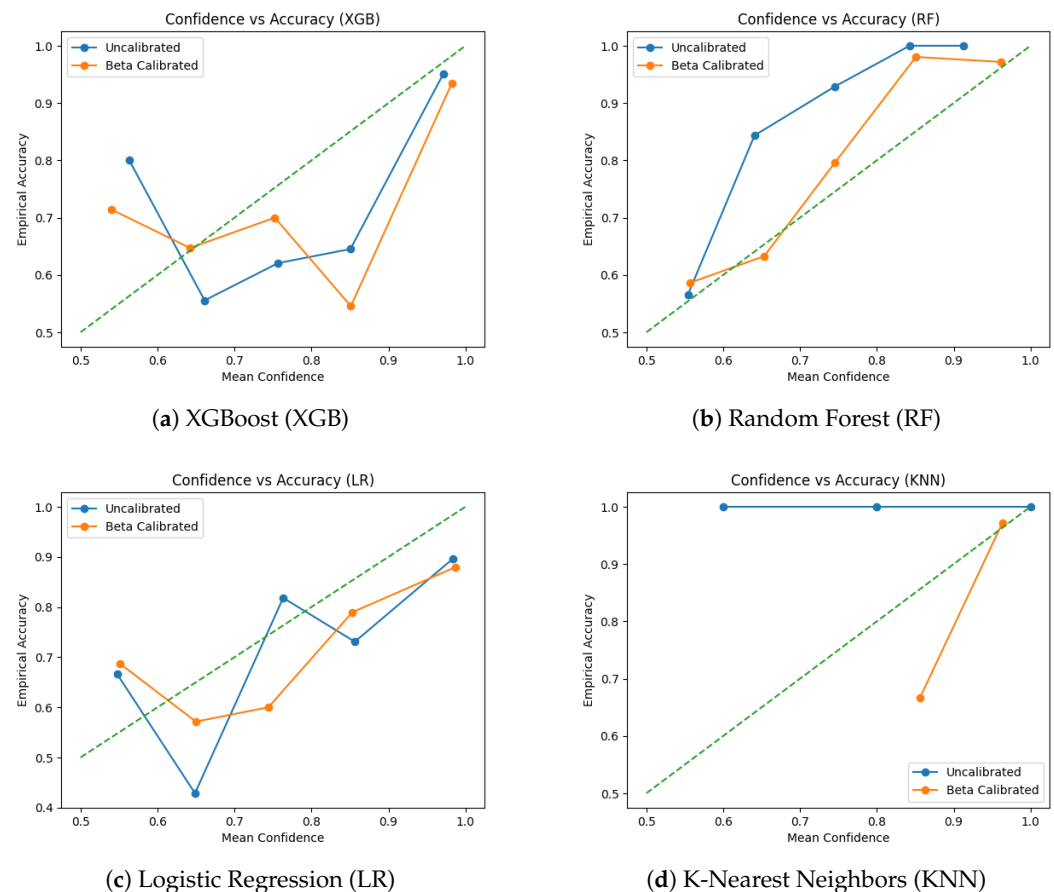


Figure 4. Confidence vs. empirical accuracy curves for selected classifiers. Solid lines represent uncalibrated predictions, whereas dashed lines represent Beta-calibrated predictions. The diagonal reference line ($y = x$) indicates perfect calibration.

4.3.3. Confidence–Accuracy Plot Analysis

The confidence–accuracy plots (Figure 4) provide a bin-wise assessment of probabilistic reliability by comparing mean predicted confidence with empirical accuracy for some selected classifiers; such plots for all other classifiers are presented in Supplementary Metrical File S3. The diagonal reference line denotes perfect calibration, where predicted probability equals observed outcome frequency. Points below the diagonal indicate overconfidence, whereas points above the diagonal reflect underconfidence. These plots allow direct visualization of how Beta calibration modifies probability estimates across different confidence regions.

For XGBoost (XGB), the uncalibrated model exhibits noticeable variability across intermediate confidence bins. Several mid-range points deviate substantially from the diagonal, indicating inconsistent probability estimation and localized overconfidence. High-confidence predictions approach better alignment but remain imperfect. After applying Beta calibration, the curve becomes more stable, particularly in moderate confidence regions, and alignment with the diagonal improves. Nevertheless, minor volatility persists, suggesting that while calibration mitigates systematic bias, intrinsic variance in ensemble-based boosting models is not entirely eliminated.

In the case of Random Forest (RF), the uncalibrated model demonstrates systematic underconfidence, especially in mid-confidence regions where empirical accuracy exceeds predicted confidence. High-confidence bins approach near-perfect discrimination but still show calibration mismatch. Beta calibration substantially improves alignment across most bins, shifting the curve closer to the diagonal and correcting the underconfidence bias. The

improvement is more uniform compared to XGB, indicating that RF benefits strongly from post hoc probabilistic recalibration.

Logistic Regression (LR), being inherently probabilistic, shows comparatively stable behavior even before calibration. Although some mid-confidence bins display mild deviations and localized underconfidence, the overall curve remains relatively smooth. Beta calibration refines these deviations, producing a more monotonic and diagonally aligned relationship between confidence and accuracy. The magnitude of improvement is moderate, reflecting the fact that linear probabilistic models generally require less aggressive recalibration than ensemble methods.

For K-Nearest Neighbors (KNN), the uncalibrated model presents a step-like pattern with extreme underconfidence in several bins, where empirical accuracy reaches very high levels despite lower predicted confidence. This behavior arises from the discrete nature of neighbor-based probability estimation. Beta calibration improves high-confidence alignment to some extent; however, instability persists in intermediate regions. The results suggest that calibration can partially smooth discrete probability artifacts but cannot fully compensate for structural limitations of instance-based probability estimation.

Overall, the plots demonstrate that Beta calibration consistently improves probability reliability across models, with the most pronounced gains observed in tree-based ensembles. High-confidence predictions become more trustworthy after calibration, and systematic biases in mid-confidence regions are reduced. These findings reinforce the importance of post hoc calibration when deploying complex machine learning models in high-stakes decision environments where probabilistic reliability is critical.

4.3.4. Model-Specific Observations

Tree ensembles (RF, XGB) exhibit only minor improvements in low-confidence bins while maintaining high accuracy in top bins. Linear models (LDA, LR) benefit more from calibration in mid-confidence bins, reducing overconfident predictions. KNN and GNB demonstrate moderate improvements, though variability in small bins ($N < 20$) limits robustness.

4.3.5. Practical Implications

Beta calibration effectively mitigates overconfidence in probabilistic predictions, particularly in mid-confidence ranges. For high-confidence predictions, ensemble methods already provide reliable estimates, indicating that calibration is most beneficial for models prone to probability misestimation.

4.4. Risk-Coverage Analysis for Uncertainty-Aware Deployment

To evaluate the reliability of the predictive models under uncertainty, we conducted a Risk-Coverage experiment. In this procedure, predictions were ranked by confidence, defined as the absolute distance of the predicted probability from 0.5. The lowest-confidence samples were progressively rejected, and accuracy was recomputed for the retained subset of predictions. This allows assessment of model performance as a function of coverage (the fraction of samples retained) and demonstrates how selective deployment can improve reliability. The procedure was applied to both uncalibrated predictions (UNCAL) and calibrated outputs, including Temperature Scaling (TEMP), Platt Scaling (PLATT), Isotonic Regression (ISO), and Beta Calibration (BETA).

Figure 5 illustrates the Risk-Coverage curves for some selected models and all calibration methods, showing the trade-off between coverage (fraction of samples retained) and accuracy. Across most models, accuracy increases monotonically as low-confidence predictions are rejected. For example, in XGB, RF, ET, and SVM, rejecting the 20% least confident predictions raised accuracy above 0.94 for high-confidence subsets (Table 12).

Calibration methods generally provide additional gains in selective prediction, with TEMP and BETA slightly outperforming others in post-rejection accuracy, highlighting the benefit of probability calibration in clinical deployment scenarios. Complete results are presented in Supplementary Material File S4.

Table 12. Accuracy improvement after rejecting 20% least confident predictions for all models and calibration methods. ‘→’ indicates the previous values x become the new vales y.

Model	UNCAL	TEMP	PLATT	ISO	BETA
XGB	0.854 → 0.911	0.854 → 0.911	0.854 → 0.911	0.854 → 0.911	0.854 → 0.909
ET	0.877 → 0.935	0.877 → 0.935	0.877 → 0.931	0.877 → 0.935	0.877 → 0.931
RF	0.862 → 0.935	0.862 → 0.940	0.860 → 0.935	0.862 → 0.933	0.858 → 0.931
DT	0.745 → nan	0.745 → 0.719	0.745 → 0.746	0.745 → nan	0.745 → 0.749
LDA	0.701 → 0.729	0.701 → 0.736	0.701 → 0.736	0.699 → nan	0.701 → 0.734
SVM	0.876 → 0.945	0.876 → 0.945	0.876 → 0.947	0.881 → 0.926	0.887 → 0.931
KNN	0.617 → 0.770	0.617 → 0.688	0.793 → 0.819	0.793 → nan	0.793 → 0.809
GNB	0.830 → 0.902	0.830 → 0.902	0.828 → 0.867	0.835 → 0.900	0.845 → 0.902
MLP	0.831 → 0.907	0.831 → 0.907	0.831 → 0.907	0.831 → 0.907	0.831 → 0.907
LR	0.814 → 0.880	0.814 → 0.883	0.814 → 0.883	0.816 → 0.883	0.814 → 0.883

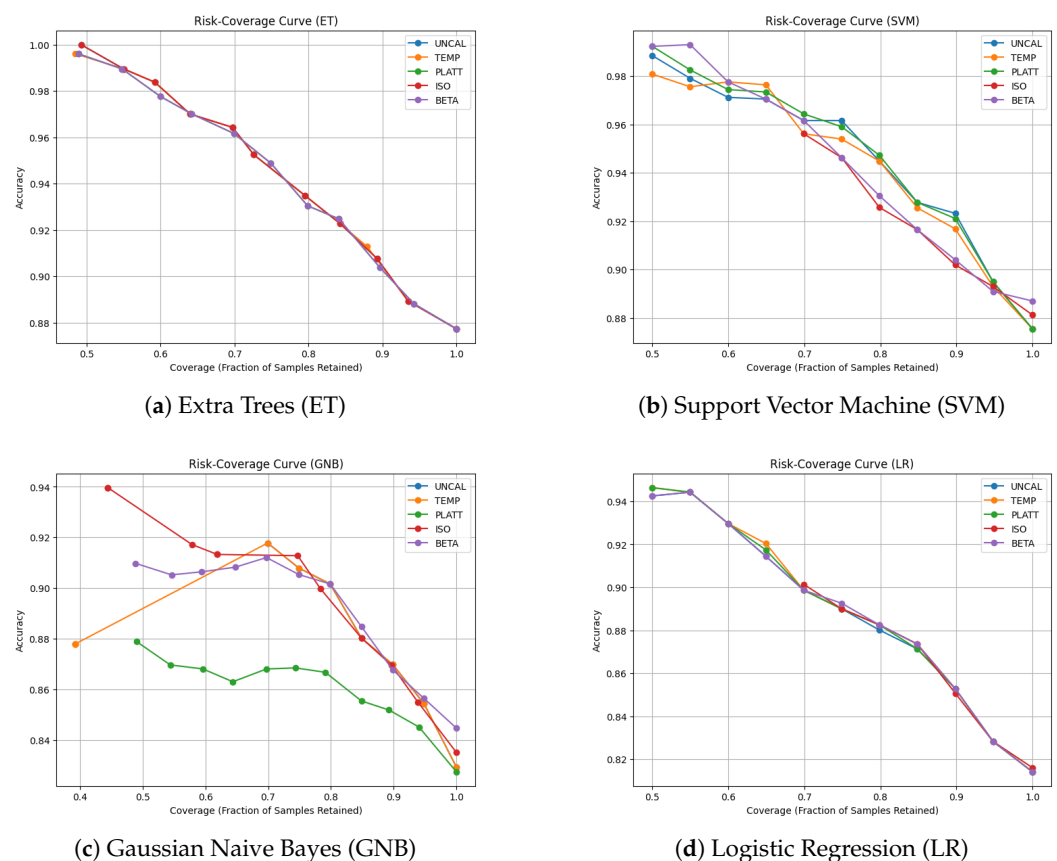


Figure 5. Risk–coverage curves for selected classifiers: Extra Trees (ET), Support Vector Machine (SVM), Gaussian Naive Bayes (GNB), and Logistic Regression (LR). Accuracy improves as low-confidence predictions are rejected, demonstrating the potential of selective deployment for safe clinical application.

These results confirm that uncertainty-based selective prediction is an effective strategy to improve model reliability. Rejecting the 20% least confident predictions consistently enhances accuracy across classifiers, offering a practical approach for safe clinical deployment where high-confidence predictions are prioritized.

4.5. Robustness and Stability Analysis: Noise, Small Data, and Feature Perturbation

To evaluate the robustness and stability of the proposed classifiers and their calibration methods, we conducted three complementary experiments. First, a noise injection experiment was performed by adding Gaussian noise to the input features at levels of 5%, 10%, and 20%, simulating measurement or sensor perturbations, and measuring the resulting degradation in AUC and Expected Calibration Error (ECE) for both calibrated and uncalibrated models. Second, a small data simulation assessed calibration stability under limited data scenarios by training models on 100%, 75%, 50%, and 25% of the dataset while evaluating performance on a held-out test set. Third, a feature perturbation sensitivity analysis was conducted by systematically dropping the top- k most important features ($k = 1, 5, 10$) based on feature importance, and examining the impact on discrimination (AUC) and calibration (ECE). Across all experiments, we compared multiple calibration strategies—including temperature scaling, Platt scaling, isotonic regression, and beta calibration—against uncalibrated models to quantify the relative improvements in robustness and stability under realistic perturbations. Complete experimental results are presented in Supplementary Material File S5.

4.5.1. Noise Injection

As shown in Table 13, noise injection slightly reduces AUC across all models. Tree-based models (XGB, RF, ET) remain robust, maintaining high discrimination. Calibration methods such as TEMP and PLATT sometimes worsen ECE under noise, while BETA shows more stable calibration.

Table 13. Noise injection performance (AUC/ECE) for different models and calibration methods.

Model	Method	Noise 5% AUC/ECE	Noise 10% AUC/ECE	Noise 20% AUC/ECE
XGB	UNCAL	0.928/0.395	0.922/0.394	0.910/0.376
	TEMP	0.928/0.475	0.921/0.476	0.910/0.470
	PLATT	0.928/0.438	0.922/0.444	0.910/0.431
	ISO	0.921/0.419	0.911/0.420	0.899/0.402
	BETA	0.928/0.415	0.922/0.419	0.911/0.406
RF	UNCAL	0.939/0.357	0.937/0.355	0.929/0.348
	TEMP	0.939/0.471	0.937/0.472	0.929/0.475
	PLATT	0.940/0.375	0.938/0.363	0.930/0.354
	ISO	0.938/0.389	0.935/0.374	0.928/0.369
	BETA	0.940/0.371	0.939/0.370	0.931/0.358
ET	UNCAL	0.950/0.383	0.947/0.374	0.943/0.360
	TEMP	0.950/0.383	0.947/0.374	0.943/0.360
	PLATT	0.950/0.370	0.947/0.360	0.942/0.352
	ISO	0.950/0.383	0.947/0.374	0.943/0.360
	BETA	0.950/0.382	0.947/0.374	0.942/0.360

4.5.2. Small Data Simulation

As summarized in Table 14, reducing data fraction decreases AUC for all models, especially simpler models. Ensemble models (RF, ET) are more resilient. Calibration affects ECE more than AUC; BETA provides the lowest ECE under limited data.

Table 14. Performance under reduced data fractions (AUC/ECE).

Model	Method	100% Data AUC/ECE	50% Data AUC/ECE	25% Data AUC/ECE
XGB	UNCAL	0.901/0.430	0.885/0.396	0.859/0.373
	TEMP	0.901/0.538	0.883/0.479	0.860/0.450
	PLATT	0.901/0.502	0.885/0.430	0.859/0.406
	ISO	0.901/0.432	0.878/0.435	0.850/0.424
	BETA	0.901/0.487	0.885/0.426	0.859/0.392
RF	UNCAL	0.901/0.299	0.937/0.366	0.867/0.360
	TEMP	0.901/0.509	0.937/0.468	0.867/0.498
	PLATT	0.901/0.473	0.937/0.374	0.869/0.375
	ISO	0.901/0.531	0.936/0.408	0.867/0.435
	BETA	0.901/0.421	0.937/0.393	0.866/0.372
ET	UNCAL	0.852/0.325	0.944/0.369	0.867/0.352
	TEMP	0.852/0.325	0.944/0.369	0.867/0.352
	PLATT	0.852/0.368	0.947/0.352	0.866/0.349
	ISO	0.852/0.325	0.944/0.369	0.867/0.352
	BETA	0.852/0.225	0.945/0.379	0.866/0.324

4.5.3. Feature Perturbation

Table 15 shows that dropping top features slightly reduces AUC, but tree-based models remain robust. Calibration affects ECE more than AUC; BETA and PLATT often provide better-calibrated predictions. Simpler models are more sensitive to feature perturbation in terms of probability reliability.

Table 15. Top-10 feature drop sensitivity (AUC/ECE).

Model	Method	AUC/ECE
XGB	UNCAL	0.934/0.400
	TEMP	0.934/0.478
	PLATT	0.934/0.445
	ISO	0.927/0.426
	BETA	0.934/0.421
RF	UNCAL	0.922/0.323
	TEMP	0.922/0.468
	PLATT	0.922/0.345
	ISO	0.920/0.358
	BETA	0.922/0.337
ET	UNCAL	0.936/0.346
	TEMP	0.936/0.346
	PLATT	0.936/0.364
	ISO	0.936/0.346
	BETA	0.936/0.344

4.6. Threshold and Clinical Utility Analysis

To assess the clinical applicability of the proposed models, we extended conventional probability calibration experiments by performing threshold optimization and decision curve analysis. For each calibrated classifier, optimal decision thresholds were identified by maximizing the Youden index, F1 score, and balanced accuracy. Subsequently, decision curve analysis was conducted to compute the net benefit across a range of thresholds, allowing quantification of clinical utility. This approach enables identification of the models and thresholds that offer the highest potential benefit in real-world decision-making scenarios, complementing standard discrimination and calibration metrics.

To evaluate the clinical applicability of the calibrated classifiers, threshold optimization and decision curve analysis (DCA) were performed. For each classifier, optimal decision thresholds were determined by maximizing the Youden index, F1 score, and balanced accuracy. Subsequently, DCA was conducted to compute the net benefit across a range of thresholds, highlighting the threshold ranges where each model provides positive clinical utility. Complete results are presented in Supplementary Material File S6.

Table 16 summarizes the optimal thresholds and corresponding metric values for all classifiers. The results indicate substantial variation in optimal thresholds depending on the evaluation metric and model characteristics. XGB, RF, and GNB achieved moderate optimal thresholds (0.32–0.42) for Youden index and balanced accuracy, reflecting a balanced trade-off between sensitivity and specificity. ET and SVM favored thresholds closer to 0.49–0.51, suggesting more conservative positive predictions. In contrast, DT, LDA, and KNN exhibited very low optimal thresholds (0.01–0.15), highlighting the need for aggressive threshold adjustment to improve positive prediction performance. Across all models, F1-optimized thresholds were generally lower than those based on Youden or balanced accuracy, emphasizing the trade-off between precision-recall and global accuracy.

Table 16. Threshold optimization results across classifiers.

Model	Youden Thresh	Youden Val	F1 Thresh	F1 Val	Balanced Thresh	Balanced Val
XGB	0.32	0.6930	0.19	0.8630	0.32	0.8465
ET	0.49	0.7318	0.49	0.8802	0.49	0.8659
RF	0.35	0.7029	0.34	0.8682	0.35	0.8515
DT	0.01	0.4646	0.01	0.7424	0.01	0.7323
LDA	0.15	0.4635	0.08	0.7526	0.15	0.7318
SVM	0.51	0.7772	0.20	0.8942	0.51	0.8886
KNN	0.01	0.6290	0.01	0.8060	0.01	0.8145
GNB	0.42	0.7034	0.41	0.8668	0.42	0.8517
MLP	0.18	0.7206	0.18	0.8743	0.18	0.8603
LR	0.33	0.6833	0.18	0.8474	0.33	0.8417

To visualize clinical utility, DCA plots (Figure 6) were generated for four representative models (XGB, ET, DT, and SVM), showing net benefit across threshold probabilities. In addition to the model curves, the standard treat-all and treat-none reference strategies were included to facilitate direct interpretation of clinical utility. The treat-none strategy corresponds to a net benefit of zero across all threshold probabilities, whereas the net benefit of the treat-all strategy varies according to the observed disease prevalence and decreases as the threshold probability increases. Thus, a model was considered to provide clinical utility when its net benefit exceeded both reference strategies over a meaningful threshold range. Vertical lines indicate thresholds maximizing the Youden index, F1 score, and balanced accuracy, providing a clear visualization of the corresponding optimal decision points.

XGB and ET demonstrated positive net benefit across a broad range of threshold probabilities and generally remained above both the treat-all and treat-none reference strategies, indicating comparatively favorable clinical utility. ET showed a relatively stable net-benefit profile across the evaluated threshold range, whereas XGB maintained positive net benefit over much of the range before declining at very high thresholds. SVM similarly maintained a small positive net benefit across a broad range of thresholds, while DT demonstrated positive net benefit primarily at low-to-intermediate thresholds and showed a progressive decline at higher thresholds, eventually becoming negative. These findings indicate that the clinical utility of the models is dependent on the selected decision threshold and that the benefit of model-based decision-making should be evaluated relative to the default treat-all and treat-none strategies rather than from the model net-benefit curve alone.

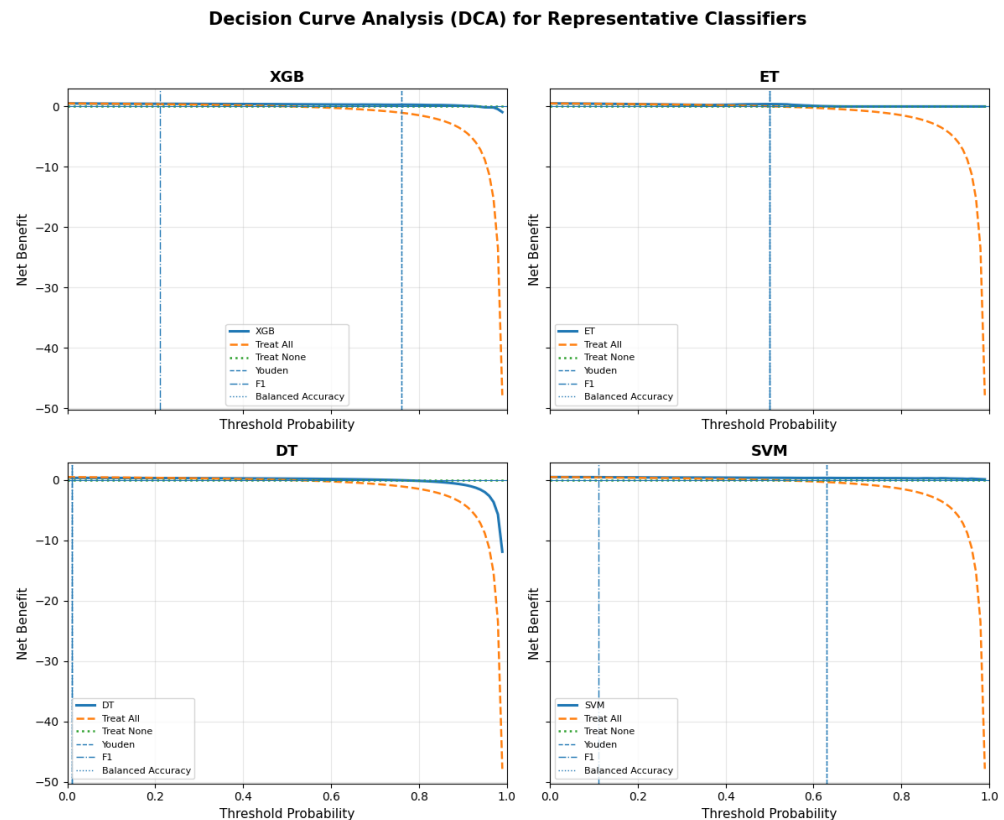


Figure 6. Decision Curve Analysis (DCA) for four representative classifiers (XGB, ET, DT, and SVM). The curves show clinical net benefit across a range of threshold probabilities together with the standard treat-all and treat-none reference strategies. The treat-none strategy corresponds to a net benefit of zero, whereas the treat-all curve represents the net benefit obtained by classifying all individuals as positive. A model provides clinical benefit when its net benefit exceeds both reference strategies over a meaningful threshold range. Vertical dashed lines denote the optimal operating thresholds obtained using the Youden index, F1-score, and balanced accuracy criteria.

These results highlight the critical role of threshold optimization in improving clinical utility, complementing conventional calibration and discrimination metrics, and provide guidance for selecting thresholds that maximize net benefit in real-world applications.

4.7. Comprehensive Evaluation of Ensemble Reliability Study

To systematically investigate the effect of calibration on ensemble reliability, a structured experimental protocol was conducted using repeated stratified 10-fold cross-validation with five repetitions (50 folds total) to ensure robust and unbiased performance estimation. Five configurations were evaluated: (i) uncalibrated base learners aggregated via probability averaging (BASE_UNCAL); (ii) individually calibrated base learners using Platt scaling (BASE_CAL); (iii) stacking ensemble constructed from uncalibrated base predictions (STACK_UNCAL); (iv) stacking ensemble built on calibrated base learners (STACK_CAL); (v) meta-level calibrated stacking (META_CAL), where sigmoid calibration was applied to the stacking meta-classifier. Statistical tests are presented in Tables 17–20.

Table 17. Performance comparison of ensemble configurations (Mean $\pm$ Std over 50 folds). Lower values are better for Brier, LogLoss, and ECE.

Model	Accuracy	F1-Score	AUC	MCC	Brier $\downarrow$	LogLoss $\downarrow$	ECE $\downarrow$
BASE_UNCAL	0.8611 $\pm$ 0.0776	0.8643 $\pm$ 0.0764	0.9310 $\pm$ 0.0566	0.7301 $\pm$ 0.1522	0.1088 $\pm$ 0.0346	0.3610 $\pm$ 0.0825	0.1860 $\pm$ 0.0392
BASE_CAL	0.8718 $\pm$ 0.0764	0.8797 $\pm$ 0.0694	0.9575 $\pm$ 0.0392	0.7524 $\pm$ 0.1490	0.1203 $\pm$ 0.0208	0.4116 $\pm$ 0.0443	0.2357 $\pm$ 0.0474
STACK_UNCAL	0.8646 $\pm$ 0.0792	0.8685 $\pm$ 0.0774	0.9230 $\pm$ 0.0601	0.7377 $\pm$ 0.1557	0.1030 $\pm$ 0.0491	0.3375 $\pm$ 0.1399	0.1523 $\pm$ 0.0518

Table 17. *Cont.*

Model	Accuracy	F1-Score	AUC	MCC	Brier ↓	LogLoss ↓	ECE ↓
STACK_CAL	0.8752 ± 0.0741	0.8819 ± 0.0676	0.9621 ± 0.0355	0.7586 ± 0.1449	0.0824 ± 0.0415	0.2677 ± 0.1129	0.1363 ± 0.0425
META_CAL	0.8657 ± 0.0809	0.8701 ± 0.0788	0.9233 ± 0.0601	0.7398 ± 0.1596	0.1021 ± 0.0449	0.3358 ± 0.1180	0.1553 ± 0.0422

Table 18. Friedman test for multiple model comparison (ECE).

Number of Models (<i>k</i>)	5
Number of Folds (<i>N</i>)	50
<i>p</i> -value	<0.001

Table 19. Average ranks (lower is better, based on ECE).

STACK_CAL	1.60
STACK_UNCAL	2.36
META_CAL	2.42
BASE_UNCAL	3.02
BASE_CAL	4.60

Table 20. Nemenyi Critical Difference (CD) at $\alpha = 0.05$.

q_α	2.728
Critical Difference (CD)	0.863

Performance was assessed using both discrimination metrics (Accuracy, F1-score, AUC, and MCC) and probabilistic reliability metrics (Brier score, LogLoss, and Expected Calibration Error). Mean ± standard deviation values were computed across all folds to quantify performance stability.

4.7.1. Discrimination Performance

STACK_CAL achieved the highest discrimination performance across all metrics, including Accuracy (0.8752), F1-score (0.8819), AUC (0.9621), and MCC (0.7586). BASE_CAL improved AUC relative to BASE_UNCAL (0.9575 vs. 0.9310), confirming enhanced ranking ability after sigmoid calibration. However, improvements in ranking did not consistently translate into improved probability calibration.

4.7.2. Calibration Performance

From a reliability perspective, STACK_CAL achieved the lowest Brier score (0.0824), LogLoss (0.2677), and ECE (0.1363), indicating superior alignment between predicted probabilities and observed outcomes. In contrast, BASE_CAL exhibited degraded calibration compared to BASE_UNCAL, with increased Brier score (0.1203), LogLoss (0.4116), and ECE (0.2357). This suggests that independently calibrating heterogeneous base learners may distort aggregated probability distributions. META_CAL provided moderate improvements over STACK_UNCAL but did not surpass STACK_CAL.

4.7.3. Statistical Significance Analysis

The Friedman test confirmed statistically significant differences among the five configurations.

The Critical Difference was computed as:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} = 0.863.$$

Average rank differences exceeding 0.863 are statistically significant under the Nemenyi criterion. STACK_CAL significantly outperformed BASE_CAL and BASE_UNCAL, while differences between STACK_CAL, STACK_UNCAL, and META_CAL did not exceed the CD threshold.

Kendall's $W = 0.326$ indicates a moderate effect size, suggesting meaningful agreement in ranking across folds (Table 21).

Table 21. Kendall's coefficient of concordance (effect size).

Kendall's W	0.326
Effect Size Interpretation	Moderate

4.7.4. Structural Interpretation

Calibration effectiveness depends on structural placement within the ensemble. Calibration integrated prior to stacking enables improved propagation of uncertainty information, whereas post hoc meta-level calibration provides limited correction. These findings emphasize that probability correction should be structurally aligned with ensemble architecture rather than applied as an isolated post-processing step.

4.7.5. Clinical Implications

In clinical decision-support systems, predicted probabilities serve as risk scores guiding triage and intervention thresholds. The superior calibration performance of STACK_CAL supports more reliable risk estimation. Conversely, the deterioration observed for BASE_CAL demonstrates that improvements in AUC alone do not guarantee trustworthy probability estimates, underscoring the importance of evaluating both discrimination and calibration before deployment in high-stakes environments.

4.8. Cross-Dataset Applicability Assessment Using the Alzheimer's Disease Clinical Dataset

The reliability analysis performed on the DARWIN handwriting dataset identified the STACK_CAL ensemble configuration as the most reliable candidate, demonstrating the best overall balance between discrimination performance, probability calibration, and uncertainty estimation. To assess the cross-dataset applicability of the proposed reliability-aware strategy beyond handwriting-derived biomarkers, the selected STACK_CAL configuration was independently evaluated on the Alzheimer's Disease Clinical Dataset. Because the two datasets have substantially different feature spaces and data modalities, this experiment does not constitute external validation of a model trained on the DARWIN dataset. Instead, the STACK_CAL architecture and its associated calibration and ensemble procedures were independently re-trained from scratch on the clinical dataset to determine whether the proposed reliability strategy remains effective when reapplied to a different clinical data modality.

Unlike the DARWIN dataset, which consists of handwriting-derived kinematic features, the Alzheimer's Disease Clinical Dataset comprises demographic, lifestyle, medical, and cognitive attributes representing a substantially different clinical data modality. Since the two datasets share no common feature space, no feature-level transfer was performed. Instead, the same STACK_CAL architecture, including the same calibration strategy and ensemble construction procedure, was re-trained from scratch on the clinical dataset. Consequently, this experiment evaluates the applicability of the calibration-aware stacking strategy under a different feature representation and independent clinical data distribution, rather than the direct generalization or external validation of the handwriting-trained model.

The independently retrained STACK_CAL configuration was evaluated using repeated stratified 10-fold cross-validation with five repetitions, resulting in 50 held-out test-fold evaluations. Within each training fold, four heterogeneous base classifiers (Extra Trees, Random Forest, Extreme Gradient Boosting, and Support Vector Machine) were trained using embedded feature selection through Recursive Feature Elimination (RFE) or SelectKBest. The calibrated probability estimates were then stacked to train a Logistic Regression meta-classifier, forming the final calibrated stacking ensemble.

Performance was evaluated on the held-out test folds using both discrimination metrics (Accuracy, F1-score, ROC-AUC, and Matthews Correlation Coefficient) and probabilistic reliability metrics (Brier Score, LogLoss, and Expected Calibration Error). Reliability diagrams, risk–coverage curves, and uncertainty-aware selective prediction were additionally employed to evaluate the consistency of probabilistic predictions under deployment-oriented clinical decision-support scenarios. The lowest 20% of predictions, ranked according to the confidence margin $m_i = |\hat{p}_i - 0.5|$, were referred for expert review, and the corresponding referral rate was recorded.

Table 22 summarizes the performance of the independently retrained STACK_CAL configuration on the Alzheimer’s Disease Clinical Dataset. The model achieved an Accuracy of 0.8245 ± 0.0201 and a ROC-AUC of 0.8582 ± 0.0249 , indicating stable discriminative capability across the 50 held-out test-fold evaluations despite the substantially different feature space. The corresponding F1-score (0.7110 ± 0.0341) and MCC (0.6076 ± 0.0483) further demonstrate balanced predictive performance under class imbalance.

Table 22. Performance of the independently retrained STACK_CAL configuration on the Alzheimer’s Disease Clinical Dataset (Mean $\pm$ Std over 50 held-out test folds).

Metric	Mean $\pm$ Std
Accuracy	0.8245 ± 0.0201
F1-score	0.7110 ± 0.0341
ROC-AUC	0.8582 ± 0.0249
MCC	0.6076 ± 0.0483
Brier Score	0.1470 ± 0.0161
Log-Loss	0.5698 ± 0.0721
ECE	0.1324 ± 0.0188

From a probabilistic reliability perspective, the independently retrained calibration-aware stacking configuration achieved a Brier Score of 0.1470 ± 0.0161 , LogLoss of 0.5698 ± 0.0721 , and an Expected Calibration Error of 0.1324 ± 0.0188 . These results indicate that the reliability-aware strategy can provide calibrated probability estimates when independently reapplied to the clinical dataset. Because the model was retrained from scratch and the feature space differs substantially from that of DARWIN, these findings should be interpreted as evidence of cross-dataset applicability of the reliability strategy rather than preservation or transfer of the reliability characteristics of a handwriting-trained model.

In the selective prediction analysis, the lowest 20% of predictions ranked according to the confidence margin $m_i = |\hat{p}_i - 0.5|$ were designated for referral to expert assessment. Here, 20% denotes the rejection proportion rather than a numerical confidence threshold. This conservative referral strategy limits automated prediction coverage and provides a mechanism for directing low-confidence cases toward human review. The result therefore illustrates how uncertainty-aware decision support can be incorporated into the independently evaluated clinical setting, rather than demonstrating readiness for autonomous clinical deployment. Overall, the findings indicate that the proposed reliability-aware calibration and stacking strategy can be independently reapplied to a substantially different clinical dataset and retain useful discrimination and probabilistic reliability characteristics.

These results support the cross-dataset applicability of the proposed reliability strategy, while further prospective and independent clinical validation would be required before real-world deployment.

5. Conclusions and Future Work

This study introduced STACK_CAL, a calibrated stacking framework for Alzheimer's disease (AD) detection derived using the DARWIN handwriting dataset. Under a rigorous repeated stratified cross-validation protocol, ensemble and margin-based base learners were integrated through calibrated probabilistic stacking, yielding strong and statistically validated discrimination performance. The results confirm that combining heterogeneous classifiers with post hoc probability calibration can enhance both predictive performance and probabilistic reliability. In particular, Platt scaling consistently improved Brier score and Expected Calibration Error, demonstrating the importance of calibrated probabilities for clinically meaningful risk estimation.

A key contribution of this work lies in the cross-dataset applicability assessment of the proposed reliability-aware strategy. After identifying STACK_CAL through the reliability evaluation on the DARWIN handwriting dataset, the same architecture and calibration strategy were independently retrained and evaluated on an Alzheimer's disease clinical dataset with a substantially different feature space and data modality. No feature-level transfer was performed between the two datasets. The second-dataset experiment therefore does not constitute external validation or direct generalization of a handwriting-trained model; rather, it assesses whether the proposed calibration-aware stacking and reliability evaluation strategy can be independently reapplied to a different clinical data representation. The observed discrimination and probabilistic reliability provide evidence supporting the cross-dataset applicability of the proposed strategy.

Furthermore, the integration of an uncertainty-aware selective prediction mechanism demonstrated measurable safety-oriented benefits. By flagging low-confidence predictions for referral, predictive accuracy improved within the retained high-confidence subsets, reinforcing the potential value of uncertainty-aware decision support in medical AI systems. Reliability diagrams and risk–coverage analysis further characterized the relationship between prediction confidence and empirical correctness, highlighting the importance of calibration and selective prediction when designing systems intended to support, rather than replace, clinical decision-making.

Despite these contributions, several limitations must be acknowledged. First, although the DARWIN handwriting dataset provides a structured benchmark for model derivation, there is currently no large-scale publicly available handwriting dataset for AD detection. This restricts external reproducibility and limits large-cohort validation. Second, although the proposed reliability-aware strategy was independently evaluated on a second symptom-based clinical dataset, the different feature spaces and independent retraining mean that this experiment should be interpreted as a cross-dataset applicability assessment rather than external validation of the handwriting-trained model. Broader validation across multi-center, multilingual, and demographically diverse populations remains necessary. The present study deliberately focuses on evaluating the reliability of handwriting-derived biomarkers and subsequently assessing the applicability of the proposed reliability-aware strategy on a structurally different clinical tabular dataset; therefore, multimodal fusion of handwriting, speech, neuroimaging, and cognitive assessments was beyond the scope of the current work. Future research will investigate multimodal, reliability-aware fusion strategies to determine whether complementary digital biomarkers can further improve the robustness and clinical utility of AD detection. Furthermore, the present evaluation is retrospective and dataset-based; therefore, the observed reliability and clinical utility

characteristics should not be interpreted as evidence of prospective clinical effectiveness or deployment readiness. Third, the stacking architecture relies on logistic regression as the meta-learner; alternative meta-model configurations or adaptive weighting mechanisms may further enhance performance and reliability. Finally, prospective clinical validation and real-world workflow integration were beyond the scope of this study.

Future research should prioritize the development or release of larger standardized handwriting datasets for AD detection to promote reproducibility and benchmarking. Multi-institutional validation studies are needed to assess generalizability under distribution shifts. Methodologically, exploring adaptive stacking strategies, Bayesian ensembling, and conformal prediction frameworks may provide stronger uncertainty guarantees. Additionally, integrating multimodal data—including handwriting, cognitive assessments, speech, and digital biomarkers—could further enhance robustness. Prospective clinical studies evaluating cost-sensitive decision thresholds and referral strategies will be essential for determining the practical utility of calibrated AI systems in routine screening practice.

In conclusion, this study demonstrates that a reliability-aware calibrated stacking strategy can provide strong discrimination, calibrated probability estimates, and uncertainty-aware decision support when evaluated on the DARWIN handwriting dataset. Its independent reapplication to a heterogeneous clinical dataset with a substantially different feature representation further provides evidence of the cross-dataset applicability of the proposed reliability strategy. However, because the model was retrained from scratch on the second dataset, these findings should not be interpreted as evidence of direct transferability or external validation of the handwriting-trained model. Further prospective, multi-center, and independent clinical evaluation will be required before the proposed approach can be considered for real-world clinical deployment.

Supplementary Materials: The following supporting information can be downloaded at <https://www.mdpi.com/article/10.3390/info17090876/s1>.

Author Contributions: Conceptualization, U.M. and S.U.R.; methodology, U.M. and S.U.R.; software, U.M. and S.U.R.; validation, U.M. and S.U.R.; formal analysis, U.M. and S.U.R.; investigation, U.M. and S.U.R.; resources, U.M. and S.U.R.; data curation, U.M.; writing—original draft preparation, U.M. and S.U.R.; writing—review and editing, U.M. and S.U.R.; visualization, U.M. and S.U.R.; supervision, S.U.R.; project administration, S.U.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by Kingdom University, Bahrain, under Grant KU-SRU-2026-13.

Institutional Review Board Statement: This study involved secondary analysis of the publicly available DARWIN dataset, originally collected by Cilia et al. (2022) [26] and subsequently deposited in the UCI Machine Learning Repository (DOI: 10.24432/C55D0K). The dataset comprises handwriting data from 174 participants and is publicly available under a Creative Commons Attribution 4.0 International (CC BY 4.0) license. No participants were recruited or contacted, and no new human-subject data were collected by the present authors. The ethical considerations for research involving secondary use of publicly available data are addressed in the Indian Council of Medical Research (ICMR) National Ethical Guidelines for Biomedical and Health Research Involving Human Participants (2017), Section 4 (Ethical Review Procedures). These guidelines recognize research involving less-than-minimal risk and data available in the public domain without linked identifiers among the circumstances that may qualify for exemption from review. The present study was limited to secondary computational analysis of the existing publicly available dataset and involved no direct interaction with human participants.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

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

References

1. Prince, M.; Wimo, A.; Guerchet, M.; Ali, G.C.; Wu, Y.T.; Prina, M. World Alzheimer Report 2015. The Global Impact of Dementia: An Analysis of Prevalence, Incidence, Cost and Trends. Ph.D. Thesis, Alzheimer's Disease International, London, UK, 2015.
2. Sperling, R.A.; Aisen, P.S.; Beckett, L.A.; Bennett, D.A.; Craft, S.; Fagan, A.M.; Iwatsubo, T.; Jack, C.R., Jr.; Kaye, J.; Montine, T.J.; et al. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimer's Dement.* **2011**, *7*, 280–292. [[CrossRef](#)]
3. Jack, C.R., Jr.; Bennett, D.A.; Blennow, K.; Carrillo, M.C.; Dunn, B.; Haeberlein, S.B.; Holtzman, D.M.; Jagust, W.; Jessen, F.; Karlawish, J.; et al. NIA-AA research framework: Toward a biological definition of Alzheimer's disease. *Alzheimer's Dement.* **2018**, *14*, 535–562. [[CrossRef](#)]
4. Impedovo, D.; Pirlo, G. Dynamic handwriting analysis for the assessment of neurodegenerative diseases: A pattern recognition perspective. *IEEE Rev. Biomed. Eng.* **2018**, *12*, 209–220. [[CrossRef](#)]
5. Miah, A.S.M.; Suzuki, T.; Shin, J. A methodological and structural review of Parkinson's disease detection across diverse data modalities. *IEEE Access* **2025**, *13*, 98931–98975. [[CrossRef](#)]
6. Ehrman, R. *Engineered Biomaterials for Adjuvanting Vaccine Depots, Photoimmunotherapy, and Fundamental Radiation Studies*; The University of Texas at Dallas: Richardson, TX, USA, 2025.
7. Mitra, U.; Rehman, S.U. ML-powered handwriting analysis for early detection of Alzheimer's disease. *IEEE Access* **2024**, *12*, 69031–69050. [[CrossRef](#)]
8. Ul Rehman, S.; Mitra, U. Cognitive handwriting insights for Alzheimer's diagnosis: A hybrid framework. *Information* **2025**, *16*, 249. [[CrossRef](#)]
9. Demircioglu Diren, D. Design and validation of a hybrid machine learning model for Alzheimer's detection using handwriting data. *Am. J. Alzheimer's Dis. Other Dement.* **2025**, *40*, 15333175251374913. [[CrossRef](#)]
10. Zhang, R.; Gatsonis, C.; Steingrimsson, J.A. Role of calibration in uncertainty-based referral for deep learning. *Stat. Methods Med. Res.* **2023**, *32*, 927–943. [[CrossRef](#)]
11. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On calibration of modern neural networks. In Proceedings of the International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017; pp. 1321–1330.
12. Marconi, L.; Cabitza, F. Show and tell: A critical review on robustness and uncertainty for a more responsible medical AI. *Int. J. Med. Inform.* **2025**, *202*, 105970. [[CrossRef](#)]
13. Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U.R.; et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Inf. Fusion* **2021**, *76*, 243–297. [[CrossRef](#)]
14. Gawlikowski, J.; Tassi, C.R.N.; Ali, M.; Lee, J.; Humt, M.; Feng, J.; Kruspe, A.; Triebel, R.; Jung, P.; Roscher, R.; et al. A survey of uncertainty in deep neural networks: J. Gawlikowski et al. *Artif. Intell. Rev.* **2023**, *56*, 1513–1589.
15. Platt, J. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv. Large Margin Classif.* **1999**, *10*, 61–74.
16. Zadrozny, B.; Elkan, C. Transforming classifier scores into accurate multiclass probability estimates. In Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, AB, Canada, 23–26 July 2002; pp. 694–699.
17. Kull, M.; Silva Filho, T.; Flach, P. Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In *Artificial Intelligence and Statistics*; PMLR: Cambridge, MA, USA, 2017; pp. 623–631.
18. Geifman, Y.; El-Yaniv, R. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*; Curran Associates: Red Hook, NY, USA, 2017; Volume 30.
19. Vickers, A.J.; Elkin, E.B. Decision curve analysis: A novel method for evaluating prediction models. *Med. Decis. Mak.* **2006**, *26*, 565–574. [[CrossRef](#)]
20. Calster, B. Calibration: The Achilles heel of predictive analytics. *BMC Med.* **2019**, *17*, 230. [[CrossRef](#)]
21. Minderer, M.; Djolonga, J.; Romijnders, R.; Hubis, F.; Zhai, X.; Houlsby, N.; Tran, D.; Lucic, M. Revisiting the calibration of modern neural networks. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 15682–15694.
22. Schwabe, D.; Becker, K.; Seyferth, M.; Klaw, A.; Schaeffter, T. The METRIC-framework for assessing data quality for trustworthy AI in medicine: A systematic review. *npj Digit. Med.* **2024**, *7*, 203. [[CrossRef](#)]
23. Šuster, S.; Baldwin, T.; Verspoor, K. Analysis of predictive performance and reliability of classifiers for quality assessment of medical evidence revealed important variation by medical area. *J. Clin. Epidemiol.* **2023**, *159*, 58–69. [[CrossRef](#)]
24. Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.; Lakshminarayanan, B.; Snoek, J. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*; Curran Associates: Red Hook, NY, USA, 2019; Volume 32.

25. Finlayson, S.G.; Bowers, J.D.; Ito, J.; Zittrain, J.L.; Beam, A.L.; Kohane, I.S. Adversarial attacks on medical machine learning. *Science* **2019**, *363*, 1287–1289. [[CrossRef](#)]
26. Cilia, N.D.; De Gregorio, G.; De Stefano, C.; Fontanella, F.; Marcelli, A.; Parziale, A. Diagnosing Alzheimer’s disease from on-line handwriting: A novel dataset and performance benchmarking. *Eng. Appl. Artif. Intell.* **2022**, *111*, 104822. [[CrossRef](#)]

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.