IMBALANCE-AWARE EVALUATION OF STROKE RISK PREDICTION: A METHODOLOGICAL BENCHMARK USING CALIBRATION AND DECISION CURVE

Dewi Juliah Ratnaningsih[1)]
Wisnu Aji Pamungkas[2)]
[1,2)] Statistics Study Program, Faculty of Science and Technology, Universitas Terbuka, Tangerang Selatan, Indonesia
e-mail: djuli@ecampus.ut.ac.id

ABSTRACT

Stroke prediction models are often evaluated using accuracy and ROC-AUC, although these metrics can be misleading when the outcome is rare. This study presents a methodological benchmark—not a clinical validation study—for imbalance-aware evaluation of stroke risk prediction, emphasizing probability reliability, threshold performance, and decision-analytic utility. Logistic Regression (LR), Random Forest (RF), and Gradient Boosting (GB) were evaluated using the Kaggle Stroke Prediction Dataset (n = 5,110; 249 stroke cases; prevalence = 4.87%). Training folds were balanced by random oversampling, and an analytical prior correction was applied to adjust the class-prior shift introduced by oversampling. Evaluation included ROC-AUC, precision-recall AUC, Brier score, expected calibration error, calibration slope and intercept, exploratory F2-optimal thresholds, and decision curve analysis. GB achieved the highest discrimination (AUC = 0.8415; PR-AUC = 0.2126), whereas LR showed the best calibration (ECE = 0.0048; slope = 0.967). RF produced high accuracy but very low sensitivity and severe miscalibration (slope = 0.556), making its probabilities unsuitable for threshold-based interpretation without further recalibration. LR and GB generated comparable net benefit within a prespecified exploratory threshold range of 1–10%. These findings show that calibration-aware and decision-analytic evaluation is necessary for imbalanced prediction tasks. Because the dataset has uncertain clinical provenance and no external validation was performed, the results are not intended for direct clinical implementation.

Keywords: calibration, class imbalance, decision curve analysis, machine learning, stroke prediction.

INTRODUCTION

Stroke remains a major public health problem because it contributes substantially to mortality, disability, and long-term care burden. Global and regional epidemiological studies indicate that stroke is one of the leading causes of death and long-term disability, while Indonesian national health reporting has also identified stroke as a major non-communicable disease burden (Feigin et al., 2015; Hankey, 2017; Kementerian Kesehatan Republik Indonesia, 2018). Established clinical and epidemiological studies show that age, hypertension, heart disease, glucose-related factors, and lifestyle-related variables are consistently associated with stroke risk (Gorelick, 2020; Rothwell et al., 2005). These variables are often available in routine clinical records, which makes them attractive predictors for early risk stratification.

Machine learning has become increasingly popular for stroke prediction because it can model nonlinear relationships and interactions among demographic, clinical, and lifestyle variables. Several recent studies using the widely used Kaggle Stroke Prediction Dataset have reported acceptable to high discrimination for logistic regression, random forest, boosting methods, support vector machines, neural networks, and explainable machine learning pipelines (Biswas et al., 2022; Dritsas & Trigka, 2022; Fang et al., 2022; Hassan et al., 2024; Kokkotis et al., 2022; Shoaib et al., 2023). More recent studies continue to explore imbalanced stroke prediction and key risk factor identification (Akinwumi et al., 2025; Melnykova et al., 2025). A systematic review also indicates

that the stroke prediction literature is methodologically heterogeneous and often lacks complete reporting of calibration and decision-analytic utility (Asadi et al., 2024).

The central problem is that many stroke prediction studies emphasize accuracy or ROC-AUC. These indicators are useful but incomplete. When the event rate is low, a model can achieve high accuracy by predicting the majority class and can achieve reasonable ROC-AUC while still failing to identify clinically important positive cases. Precision-recall curves are more sensitive to minority-class performance because their baseline is the event prevalence rather than the overall distribution of both classes (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015). Therefore, PR-AUC is a more informative complement to ROC-AUC when stroke cases represent a small minority of observations.

Probability reliability is equally important. Clinical prediction models are not only ranking tools; they are often used to support decisions at specific risk thresholds. A predicted risk of 8%, for example, should be interpretable as a risk estimate rather than merely as a score. Calibration measures whether predicted probabilities agree with observed event rates, and poor calibration can make a model unsafe even when discrimination appears acceptable (Steyerberg, 2009; Van Calster et al., 2016, 2019). The TRIPOD+AI statement and recent guidance for evaluating clinical prediction models emphasize that discrimination, calibration, and clinical consequences should be reported together (Collins, Dhiman, et al., 2024; Collins, Moons, et al., 2024; Riley et al., 2024).

Class imbalance handling further complicates probability interpretation. Random oversampling and related procedures are frequently applied to improve sensitivity for the minority class. However, oversampling modifies the class prior in the training data and can inflate predicted probabilities when these probabilities are interpreted on the original population scale (Fernández et al., 2018; Johnson & Khoshgoftaar, 2019; Pozzolo et al., 2015). If calibration metrics are computed directly on uncorrected probabilities from oversampled models, the resulting Brier score, ECE, and decision curves may reflect the artificial training prior rather than the original event prevalence. Analytical prior correction can address this sampling-induced prior shift, but it does not guarantee calibration or replace empirical recalibration. This distinction is especially important when thresholds are defined on the probability scale.

The present study addresses these issues by applying an imbalance-aware evaluation framework to three commonly used classifiers: Logistic Regression (LR), Random Forest (RF), and Gradient Boosting (GB). The framework combines analytical prior correction, calibration slope/intercept, Brier score, expected calibration error (ECE), PR-AUC, exploratory F2-threshold optimization, and decision curve analysis (DCA). It does not propose a new algorithm or validate a clinical tool; instead, it demonstrates how established evaluation methods can be assembled into a coherent benchmark protocol for imbalanced binary prediction. The objective is to show that models with similar ROC-AUC can have very different probability reliability and decision-curve behavior.

This study also responds to recent concerns about benchmark datasets and reporting quality. Public datasets can support reproducible methodological comparisons, but their provenance and representativeness must be examined carefully (Gibson et al., 2026; Soriano, 2021). Accordingly, all results in this paper are interpreted strictly as benchmark findings rather than as evidence of clinical validity, transportability, or readiness for implementation. The contribution is a methodological article that clarifies how statistical evaluation can be conducted when a binary outcome is severely imbalanced and predicted probabilities are examined under threshold-based decision rules.

METHOD

Study design and dataset

This study used a quantitative methodological benchmark design and did not aim to develop or validate a clinical tool. The dataset was the Kaggle Stroke Prediction Dataset, which contains 5,110 individual records and a binary stroke outcome (Soriano, 2021). It includes 249 stroke cases (4.87%) and 4,861 non-stroke cases (95.13%), yielding an imbalance ratio of 19.52. Predictors include age, average glucose level, body mass index (BMI), hypertension, heart disease, gender, ever-married status, work type, residence type, and smoking status. The public dataset contains no direct personal identifiers; it was used solely to compare evaluation procedures, and no participant contact or intervention occurred.

The clinical provenance and validity of this dataset are uncertain, and recent provenance research has raised serious concerns about its reliability (Gibson et al., 2026). It was therefore treated only as a methodological benchmark, not as a source of validated clinical evidence. The analysis was not intended to produce a deployable stroke screening or risk-stratification tool; its purpose was to evaluate how discrimination, calibration, threshold metrics, and decision curves behave under severe class imbalance.

BMI had 201 missing values (3.93%). Missing values were imputed using the median computed within the training fold to avoid information leakage. Median imputation is simple and transparent, but it assumes that missingness does not strongly distort the relationship between BMI and stroke risk. More sophisticated missing-data procedures could be considered in future work when richer data and stronger assumptions are available (Little & Rubin, 2019).

Preprocessing and resampling

Categorical variables were one-hot encoded with reference-level dropping, producing 15 model features. Continuous predictors were standardized for LR using z-score transformation based only on training-fold statistics. Tree-based models were evaluated without standardization because their splitting rules are not sensitive to monotone scaling of numeric predictors.

Random oversampling with replacement was applied exclusively within training folds. This step was designed to reduce the dominance of the non-stroke class during model training while avoiding leakage from validation folds. Random undersampling and no-resampling protocols were also evaluated for LR as a robustness check. The resampling strategy was implemented using standard imbalanced-learning practice, while acknowledging that resampling can change model calibration and should be interpreted carefully (Chawla et al., 2002; Lemaître et al., 2017).

Prior correction of oversampled probabilities

Random oversampling changes the apparent class prior in the training folds from the original prevalence of 4.87% to approximately 50%. Because probabilities from a model trained on this altered distribution are not automatically probabilities under the original prevalence, an analytical prior correction was applied before probability-based evaluation. The correction factor was alpha = (0.0487/0.9513)/(0.50/0.50) = 0.0512 (Pozzolo et al., 2015). This adjustment addresses only the class-prior shift introduced by oversampling, under the assumption that the class-conditional feature distributions are otherwise unchanged. It does not correct model misspecification, guarantee calibration, or replace empirical recalibration methods such as logistic/Platt scaling or isotonic regression (Ojeda et al., 2023; Van Calster et al., 2016).

The mathematical notation used in the evaluation framework is defined as follows. Let y_i indicate the observed class for individual i, where $y_i = 1$ denotes stroke and $y_i = 0$ denotes non-stroke. Let p_i be the raw probability produced by a model trained on the oversampled training fold,

and let p_i* be the probability after correction to the original event prevalence. The true event prevalence and imbalance ratio were first defined using Equations (1) and (2).

$$\pi = n_1 / (n_0 - n_1) \quad (1)$$

where π is the event prevalence, n_1 is the number of stroke cases, and n_0 is the number of non-stroke cases.

$$IR = n_0 / n_1 \quad (2)$$

where IR is the imbalance ratio. In this dataset, IR = 4861/249 = 19.52, which confirms severe class imbalance.

$$\alpha = [\pi\_real / (1 - \pi\_real)] / [\pi\_train / (1 - \pi\_train)] \quad (3)$$

where α is the prior-correction factor, π_real is the original dataset prevalence, and π_train is the apparent prevalence after oversampling.

$$p_i^* = \alpha p_i / [\alpha p_i + (1 - p_i)] \quad (4)$$

Equation (4) maps raw oversampled probabilities back to the original prevalence scale. In this study, π_real = 0.0487, π_train = 0.5000, and α = 0.0512.

This correction was applied before computing Brier score, ECE, calibration slope/intercept, exploratory F2-optimal thresholds, and decision curves. It was not used to alter the observed labels or the cross-validation structure. Its purpose was to return probabilities to the original prevalence scale. Crucially, prior correction was treated as a prevalence adjustment rather than a complete calibration procedure: post-correction calibration was still evaluated, and no additional recalibration model was fitted in this benchmark.

Model development and evaluation

Three classifiers were evaluated: LR, RF, and GB. Hyperparameters were selected using five-fold stratified inner cross-validation, and primary performance estimation used ten-fold stratified outer cross-validation with out-of-fold probability aggregation. A nested cross-validation check was conducted for LR to examine whether the two-stage procedure introduced noticeable bias. This design follows established recommendations that model selection and model evaluation should be separated when estimating generalization performance (Kohavi, 1995). For the threshold analysis, the F2-optimal cutoff for each model was selected by maximizing F2 on the pooled prior-corrected out-of-fold predictions. The same pooled predictions were then used to summarize sensitivity, specificity, precision, and F2 at that cutoff. Consequently, these threshold-specific estimates are exploratory and may be optimistic; they are not independently validated operating points. A deployment-oriented analysis would select the cutoff within training data and evaluate it in a separate validation or external dataset.

The evaluation framework included discrimination, calibration, threshold performance, and decision-analytic utility. Discrimination was evaluated with ROC-AUC and PR-AUC. PR-AUC was

emphasized because stroke prevalence was low and precision-recall analysis is especially informative when the positive class is rare (Davis & Goadrich, 2006; Saito & Rehmsmeier, 2015). Classification performance at selected thresholds was summarized using sensitivity, specificity, precision, accuracy, F1 score, and F2 score. F2 was used as an illustrative recall-oriented criterion because it weights recall more strongly than precision (Powers, 2020); it should not be interpreted as establishing a clinically validated cost ratio.

Calibration was assessed using the Brier score, expected calibration error (ECE), calibration intercept, and calibration slope. These metrics were computed after prior correction because raw probabilities from oversampled models are not on the original prevalence scale. The Brier score in Equation (5) summarizes the mean squared distance between the observed class and the corrected probability (Brier, 1950).

$$Brier = (1 / n) \sum_{i=n} (y_i - p_i) \tag{5}$$

A lower Brier score indicates better overall probabilistic accuracy. For an uninformative model, the approximate no-skill Brier score is $\pi(1 - \pi)$.

$$ECE = \sum_{m=1} M (|B_m|/n) |acc(B_m) - conf(B_m)| \tag{6}$$

In Equation (6), B_m denotes the m-th probability bin, $acc(B_m)$ is the observed event rate in that bin, and $conf(B_m)$ is the mean predicted probability in the bin. Smaller ECE values indicate closer agreement between predicted and observed risks.

$$logit\{Pr(Y_i = 1)\} = \beta_0 + \beta_1 logit(p_i^*) \tag{7}$$

Equation (7) defines the calibration model. The intercept β_0 assesses calibration-in-the-large, while the slope β_1 assesses whether predictions are too extreme or too compressed. The ideal values are β_0 = 0 and β_1 = 1.

Threshold-based performance was evaluated using sensitivity, specificity, precision, F2 score, and decision curve analysis. Sensitivity and specificity describe the true-positive and true-negative rates, whereas precision is especially important under class imbalance because it reflects the proportion of predicted positive cases that are truly positive. Decision-curve interpretation was restricted a priori to an exploratory threshold interval of 0.01–0.10. This interval brackets the sample prevalence of 4.87% and illustrates low-threshold decision settings with a relatively high tolerance for false positives: based on threshold odds, t = 0.01 and t = 0.10 assign one false positive approximately 0.0101 and 0.1111 times the weight of one true positive, respectively. Equivalently, about 99 and 9 false positives have the same total weight as one true positive at the two endpoints. This range is operationally illustrative, not a guideline-endorsed clinical cutoff, and would need to be re-specified for a defined target population and action (Kerr et al., 2016; Vickers & Elkin, 2006). The main threshold formulas are shown in Equations (8) to (11).

$$Sensitivity = TP/(TP + FN), Specificity = TN/(TN + FP) \tag{8}$$

TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

$$Precision = TP/(TP + FP) \tag{9}$$

Precision was interpreted together with recall because a model may achieve high sensitivity by classifying many non-stroke cases as high risk.

$$F_\beta = (1 + \beta^2)(Precision \times Recall)/[(\beta^2 \times Precision) + Recall], \qquad \beta = 2 \tag{10}$$

The F2 score places greater weight on recall than precision. In this study it was used to illustrate a recall-oriented operating point under class imbalance; it does not by itself establish the relative clinical costs of false negatives and false positives.

$$NB(t) = (TP/n) - (FP/n)[t/(1 - t)] \tag{11}$$

Equation (11) defines net benefit at decision threshold t. The term t/(1 − t) weights false positives according to the decision exchange rate implied by the threshold probability.

$$AP = \sum_k (R_k - R_{k-1})P_k \tag{12}$$

Average precision summarizes the precision-recall curve, where P_k and R_k are precision and recall at the k-th operating point. This is useful because PR-AUC is more sensitive than ROC-AUC when the event class is rare.

RESULTS AND DISCUSSION

Descriptive context and discrimination performance

The dataset contained a small minority of stroke cases, making it an appropriate benchmark for examining the limitations of accuracy and ROC-AUC under imbalance. Age, average glucose level, hypertension, heart disease, BMI, marital status, work type, and smoking status showed meaningful bivariate associations with stroke in the source analysis. These associations are plausible because they align with the known epidemiology of vascular events and stroke risk factors (Gorelick, 2020; Rothwell et al., 2005).

Table 1 presents classification performance at the default threshold. GB achieved the highest AUC (0.8415) and PR-AUC (0.2126), and LR achieved the highest sensitivity (0.8073). RF, however, had an apparently high accuracy of 0.9415 but an extremely low sensitivity of 0.0320. Its accuracy was lower than the 0.9513 accuracy of a trivial always-negative rule, and its sensitivity indicates that it missed 96.8% of stroke cases. This is a direct illustration of the accuracy paradox: when the positive outcome is rare, accuracy can make a practically ineffective classifier appear strong. Accuracy should therefore not be interpreted without sensitivity, PR-AUC, and probability calibration.

Table 1. Classification performance (mean ± SD, 10-fold CV; threshold 0.50 applied to raw probabilities)

Model	AUC	PR-AUC	Sensitivity	Specificity	Precision	Accuracy	F1
Logistic Regression	0.8371 ± 0.0238	0.2089 ± 0.0387	0.8073 ± 0.0852	0.7322 ± 0.0173	0.1338 ± 0.0118	0.7358 ± 0.0151	0.2294 ± 0.0200
Random Forest	0.8045 ± 0.0343	0.1543 ± 0.0247	0.0320 ± 0.0299	0.9881 ± 0.0042	0.1111 ± 0.1054	0.9415 ± 0.0039	0.0492 ± 0.0462
Gradient Boosting	0.8415 ± 0.0288	0.2126 ± 0.0501	0.7665 ± 0.0997	0.7657 ± 0.0191	0.1440 ± 0.0196	0.7658 ± 0.0185	0.2423 ± 0.0319

Based on Table 1, LR and GB provide more informative minority-class trade-offs than RF because they identify a substantially larger proportion of stroke cases. The small difference between LR and GB in AUC should not be overinterpreted because the value of a model under threshold-based use depends on calibration and net benefit, not discrimination alone. The result is consistent with recent stroke prediction studies that report good discrimination but often leave calibration underdeveloped (Akinwumi et al., 2025; Hassan et al., 2024; Melnykova et al., 2025).

Calibration after prior correction

Table 2 shows calibration metrics computed after prior correction. LR had the lowest ECE (0.0048) and a calibration slope close to 1.0 (0.967). GB also had good calibration, with slope 0.929 and corrected Brier score 0.0424. RF was the weakest model by calibration, with slope 0.556 and intercept 0.423. A slope far below 1.0 indicates that predicted probabilities do not vary appropriately with observed risk and are not reliable for threshold-based use.

This finding is important because RF still had an acceptable ROC-AUC. The discrepancy between discrimination and calibration demonstrates why calibration should be evaluated separately. A model may rank some high-risk and low-risk individuals correctly while producing probability values that are too compressed or otherwise misaligned with observed event rates. This is the problem described by Van Calster et al. (2019) as a central weakness of predictive analytics when calibration is neglected.

The added equations clarify why the same model can appear acceptable under one metric and problematic under another. ROC-AUC evaluates ranking over all possible thresholds, whereas Equations (5) to (7) evaluate whether the numerical probabilities themselves are trustworthy. For any threshold-based application, this distinction is essential: a predicted probability of 0.10 should correspond approximately to a 10% observed event rate among comparable observations. When the calibration slope is far below one, as observed for RF, predictions are excessively compressed and cannot be interpreted directly as individualized probabilities.

Table 2. Prior-corrected calibration metrics and calibration slope/intercept

Model	Brier (raw)	Brier (corr.)	ECE (corr.)	Calib. Slope	Calib. Intercept	Slope 95% CI	Brier 95% CI	Interpretation
Logistic Regression	0.1724	0.0426	0.0048	0.9670	-0.0543	[0.861–1.079]	[0.0382–0.0474]	Well-calibrated
Random Forest	0.0488	0.0476	0.0438	0.5563	0.4233	[0.485–0.639]	[0.0421–0.0536]	Severely miscalibrated
Gradient Boosting	0.1503	0.0424	0.0088	0.9291	0.0586	[0.821–1.055]	[0.0380–0.0473]	Well-calibrated

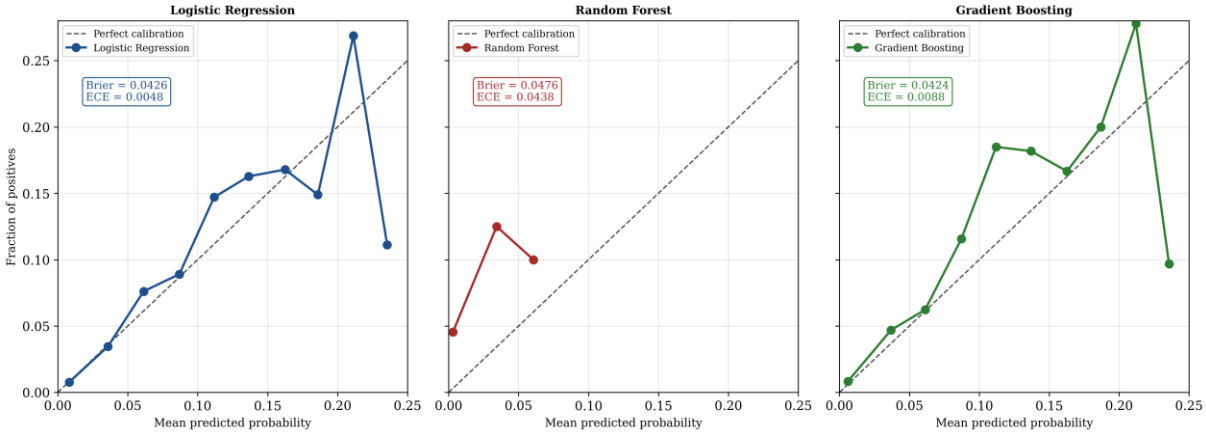

Figure 1. Calibration curves based on prior-corrected probabilities

Based on Figure 1, LR and GB track the diagonal calibration line more closely than RF. The RF curve is confined to a narrow low-probability range, explaining why RF provides little value within the prespecified exploratory threshold range after prior correction. This result also confirms that raw Brier scores from oversampled models can be misleading. Prior correction is necessary to return probabilities to the original prevalence scale, but it remains only a first step; calibration must still be assessed and, when necessary, improved through a separate recalibration procedure (Ojeda et al., 2023; Pozzolo et al., 2015).

The calibration results are also compatible with the broader calibration literature. Logistic regression is often a strong baseline for probability estimation when the functional form is adequate, whereas more flexible machine learning models may require recalibration or careful probability assessment (Ojeda et al., 2023; Van Calster et al., 2016). GB performed well in this benchmark, but RF probabilities were unsuitable for direct threshold-based interpretation. This suggests that model choice should be guided by probability reliability, not only by ranking performance.

Threshold optimization and decision utility

Table 3 compares performance at the default cutoff and at exploratory F2-optimal cutoffs. The default threshold of 0.5 is poorly aligned with an event prevalence of approximately 5%. After prior correction, LR reached its highest pooled out-of-fold F2 score at 0.070, whereas GB reached its highest score at 0.080. These low cutoffs should not be interpreted as screening recommendations; they are data-dependent operating points obtained from the same pooled out-of-fold predictions used for the threshold summary. RF improved when the threshold was lowered to 0.010, but its performance remained constrained by poor probability separation.

The threshold analysis shows that cutoff selection is not a secondary technical detail because it determines the resulting sensitivity, specificity, and precision. Reporting only the default cutoff can misrepresent performance and favor models that are not useful under the intended decision rule. F2 optimization provides a transparent recall-oriented summary, but a final cutoff should be linked to a clearly defined action, relative consequences, resource constraints, and independent validation (Powers, 2020; Riley et al., 2024).

Table 3. Performance at default and F2-optimal thresholds

Setting	Threshold	Sensitivity	Specificity	Precision	F_2
LR (default 0.5, raw)	0.50	0.8073	0.7322	0.1338	0.4023
LR (F_2-optimal, corrected)	0.070	0.7189	0.8000	0.1555	0.4169
RF (default 0.5, raw)	0.50	0.0320	0.9881	0.1111	0.0373
RF (F_2-optimal, corrected)	0.010	0.4538	0.8733	0.1550	0.3275
GB (default 0.5, raw)	0.50	0.7665	0.7657	0.1440	0.4111
GB (F_2-optimal, corrected)	0.080	0.6667	0.8424	0.1781	0.4305

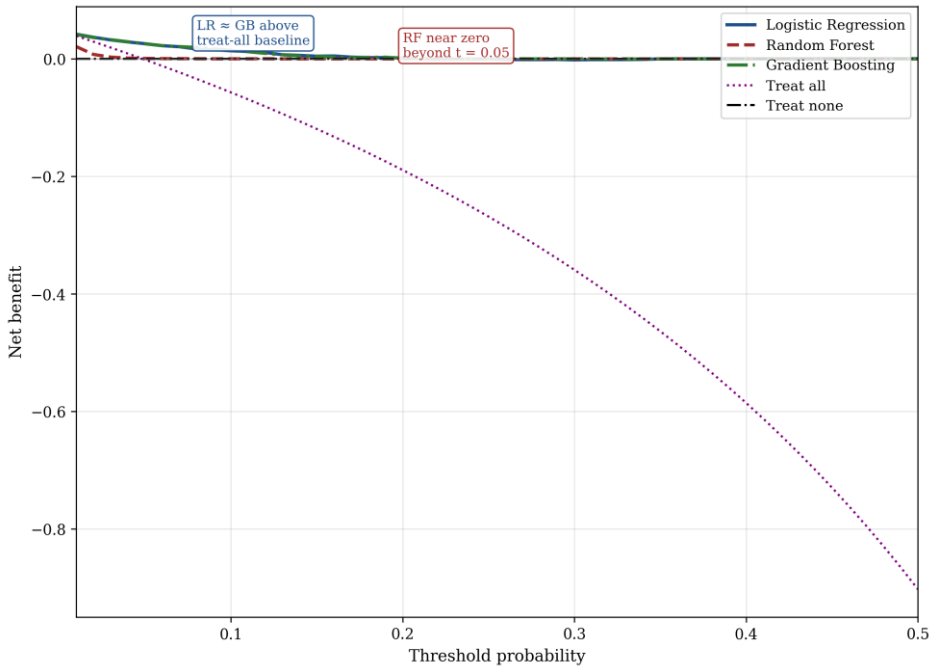

Figure 2. Decision curve analysis based on prior-corrected probabilities

Figure 2 presents the decision curve analysis. Interpretation is restricted to the prespecified exploratory threshold range of 1-10%. Within this interval, LR and GB produced positive estimated net benefit relative to the treat-all and treat-none strategies under the decision weights encoded by the selected thresholds. This is a methodological comparison of decision-curve behavior, not evidence of clinical usefulness or readiness for implementation. RF produced near-zero net benefit above approximately 5%, consistent with its weak calibration and probability separation. DCA therefore reinforces the conclusion that RF should not be preferred merely because its accuracy appears high.

DCA is useful because it translates predictions into a net-benefit scale under explicit threshold assumptions. A small difference in AUC may not matter if two models produce similar net benefit over the threshold range relevant to a defined decision. Conversely, a model with acceptable AUC may offer little estimated benefit if its net benefit is near zero. This study therefore follows the recommendation that prediction models be evaluated not only by discrimination but also

by calibration and potential decision impact (Huber et al., 2023; Kerr et al., 2016; Vickers & Elkin, 2006).

Equation (11) makes the decision weighting explicit. At t = 0.05, for example, a false positive receives a weight of 0.05/0.95 = 0.0526 relative to a true positive. DCA therefore asks whether the gain from identifying additional positive cases exceeds the threshold-weighted cost of false positives. Under the illustrative assumptions used here, LR and GB performed more favorably than RF. This result does not establish that 5%, or any other cutoff in the plotted range, is an appropriate clinical action threshold.

Feature importance and robustness

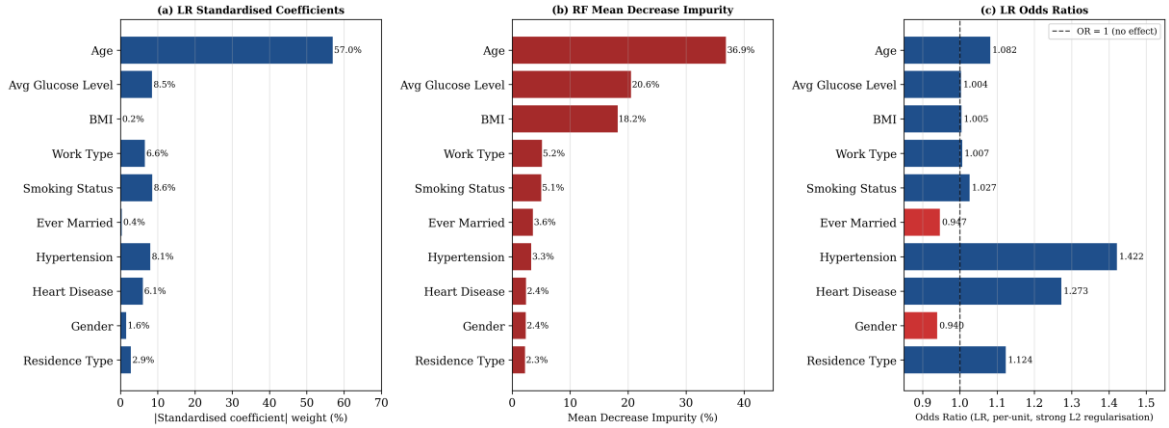

Figure 3. Feature importance proxies for LR, RF, and GB

Figure 3 shows that age was the dominant predictor across model-specific importance measures. Hypertension and heart disease were the strongest modifiable risk factors in the LR model. These findings are clinically plausible and consistent with the established role of age and cardiovascular risk factors in stroke occurrence (Hankey, 2017; Rothwell et al., 2005). They also align with recent machine learning studies that identify age, hypertension, glucose-related measures, and cardiovascular history as important predictors (Hassan et al., 2024).

However, feature importance should be interpreted cautiously. The LR coefficient-based summary is model-specific and affected by regularization. RF and GB used mean decrease impurity, which can be biased toward certain feature structures. A model-agnostic explanation method, such as SHAP, could provide a more consistent comparison of feature contributions across algorithms (Lundberg & Lee, 2017). Therefore, the feature importance results are best viewed as supportive evidence rather than causal conclusions.

The robustness check showed that LR discrimination was relatively stable across random oversampling, no resampling, and random under sampling. Random oversampling achieved the highest PR-AUC, but the AUC differences were small. This suggests that for this dataset, resampling changed the sensitivity and probability scale more strongly than the overall ranking ability of LR. It reinforces the need to report both resampling protocol and probability correction when evaluating imbalanced binary classifiers (Fernández et al., 2018; Johnson & Khoshgoftaar, 2019).

Table 4. LR robustness across three resampling protocols

Protocol	AUC (Mean+/-SD)	PR-AUC	Rank
ROS (train-only)	0.8371 ± 0.0238	0.2089	1
No Resampling	0.8333 ± 0.0213	0.2013	2
RUS (train-only)	0.8268 ± 0.0202	0.1892	3

Comparison with previous stroke prediction studies

The present results should be interpreted in relation to prior stroke prediction studies using similar or identical benchmark data. Earlier studies reported AUC values ranging from acceptable to very high, but many did not report PR-AUC, calibration slope, prior correction, or decision curves (Biswas et al., 2022; Dritsas & Trigka, 2022; Fang et al., 2022; Kokkotis et al., 2022; Shoaib et al., 2023). More recent work has improved model complexity and explainability, yet complete probability evaluation remains inconsistent (Akinwumi et al., 2025; Hassan et al., 2024; Melnykova et al., 2025).

The main contribution of this paper is therefore not a higher AUC. Instead, it demonstrates how prior-corrected calibration, PR-AUC, exploratory F2-threshold optimization, and DCA reveal different aspects of model behavior under probability-based decision rules. This contribution is aligned with TRIPOD+AI, which emphasizes transparent reporting of model development, evaluation, calibration, and intended use (Collins et al., 2024). It is also aligned with contemporary guidance that external evaluation and model updating are required before any consideration of clinical implementation (Collins, Dhiman, et al., 2024; Riley et al., 2024).

Table 5 positions the present study relative to selected prior work. Previous studies often emphasized discrimination and did not consistently include calibration or decision analysis. By adding prior correction and calibration slope, this study examines whether probability estimates remain reliable after resampling. By adding DCA, it estimates relative net benefit under explicitly stated threshold assumptions. These additions strengthen methodological interpretation but do not overcome the dataset's provenance limitations or provide clinical validation.

Table 5. Comparison with selected prior studies on the Kaggle Stroke Prediction Dataset

Study	Year	Models	AUC (best)	PR-AUC	Calibration Reported	DCA	Prior Corr.	Nested CV
Kokkotis et al.	2022	XGBoost, AdaBoost, SVM, KNN	0.921	NR	NR	No	No	No
Dritsas & Trigka	2022	Rotation Forest, Extra Trees, SGD	0.990	NR	NR	No	No	No
Biswas et al.	2022	LR, NB, DT, RF, KNN	0.870	NR	NR	No	No	No
Fang et al.	2022	LR, SVM, DT, RF, GB	0.840	NR	NR	No	No	No
Shoaib et al.	2023	RF, LR, NB, SVM, DT	0.940	NR	NR	No	No	No
Hassan et al.	2024	LR, RF, XGBoost, GB	0.890	NR	NR	No	No	No
Melnykova et al.	2025	RF, GB, LR, SVM	0.880	0.320	NR	No	No	No
Present study	2026	LR, RF, GB	0.837–0.842	0.154–0.213	Brier, ECE, slope, CI	Yes	Yes (α=0.051)	Yes (LR)

Implications for statistical and machine learning practice

For statistical practice, the findings emphasize the value of simple but complete evaluation. LR was not the most complex model, yet it produced the best calibration and decision-curve performance comparable to GB within the exploratory range. This result supports the view that interpretable models can remain competitive when the outcome is rare and the objective is reliable probability estimation. Complex models may still be valuable, but their probability outputs should not be assumed to be valid without calibration assessment (Ojeda et al., 2023; Van Calster et al., 2019).

For machine learning practice, the results show that resampling should be followed by explicit correction for sampling-induced prior shift and by calibration assessment. Prior correction can restore the prevalence scale, but residual miscalibration may still require empirical recalibration. In threshold-based decision settings, the probability scale matters because cutoffs are directly connected to actions and relative consequences. Reporting discrimination, prior-corrected calibration, and decision curves provides a clearer picture of what a model can and cannot support (Ojeda et al., 2023; Pozzolo et al., 2015).

For JMST readers, the methodological lesson is transferable beyond stroke prediction. Any binary classification problem with a rare positive class may face similar evaluation challenges. Examples include early warning systems, fraud detection, rare disease screening, equipment failure detection, and educational risk prediction. In all such cases, a model should be evaluated by its ability to rank cases, estimate reliable probabilities, and support decisions under relevant thresholds.

This study has several limitations. First, the dataset has uncertain provenance and should not be treated as a validated clinical cohort (Gibson et al., 2026; Soriano, 2021). Second, no external evaluation was performed, so transportability to other populations is unknown. Third, feature importance was based on model-specific proxies rather than a unified explanation framework. Fourth, confidence intervals for decision curves were not computed. Fifth, the prior-correction parameter is prevalence-specific and must be recalculated when the target prevalence changes. Sixth, the F2-optimal cutoffs were selected and summarized using the same pooled out-of-fold predictions, so threshold-specific performance is exploratory and may be optimistic. Finally, the 1-10% DCA interval was chosen as an operational illustration around the sample prevalence, not as a clinically validated action range.

CONCLUSION

This methodological benchmark demonstrates that evaluating stroke risk prediction under severe class imbalance requires more than accuracy and ROC-AUC. Using the Kaggle Stroke Prediction Dataset, GB achieved the highest discrimination, whereas LR achieved the best probability calibration and comparable decision-curve performance within the exploratory threshold range. RF appeared strong by accuracy but failed by sensitivity, calibration slope, and estimated net benefit. Prior correction, calibration slope, ECE, PR-AUC, exploratory F2-threshold analysis, and decision curve analysis therefore provide a more informative evaluation framework for imbalanced binary classifiers. Prior correction should not be confused with full recalibration, and data-dependent cutoffs should not be treated as validated operating points. Because the dataset has uncertain provenance and no external evaluation was conducted, none of the results should be interpreted as a clinical recommendation or as evidence that the models are ready for implementation. Future research should apply the framework to verified clinical cohorts, select thresholds within training data, evaluate them externally, compare recalibration methods, report decision-curve uncertainty, and use model-agnostic explanation methods such as SHAP.

REFERENCE

Akinwumi, P. O., Ojo, S., Nathaniel, T. I., Wanliss, J., Karunwi, O., & Sulaiman, M. (2025). Evaluating machine learning models for stroke prediction based on clinical variables. *Frontiers in Neurology*, *16*, 1668420. https://doi.org/10.3389/fneur.2025.1668420

Asadi, F., Rahimi, M., Daeechini, A. H., & Paghe, A. (2024). The most efficient machine learning algorithms in stroke prediction: A systematic review. *Health Science Reports*, *7*(10), e70062. https://doi.org/10.1002/hsr2.70062

Biswas, N., Uddin, K. M. M., Rikta, S. T., & Dey, S. K. (2022). A comparative analysis of machine learning classifiers for stroke prediction: A predictive analytics approach. *Healthcare Analytics*, *2*, 100116. https://doi.org/10.1016/j.health.2022.100116

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Collins, G. S., Dhiman, P., Ma, J., Schlussel, M. M., Archer, L., Van Calster, B., Harrell, F. E., Martin, G. P., Moons, K. G. M., van Smeden, M., Sperrin, M., Bullock, G. S., & Riley, R. D. (2024). Evaluation of clinical prediction models (part 1): from development to external validation. *BMJ*, *384*, e074819. https://doi.org/10.1136/bmj-2023-074819

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., … Logullo, P. (2024). TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, *385*, e078378. https://doi.org/10.1136/bmj-2023-078378

Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. *Proceedings of the 23rd International Conference on Machine Learning - ICML '06*, 233–240. https://doi.org/10.1145/1143844.1143874

Dritsas, E., & Trigka, M. (2022). Stroke Risk Prediction with Machine Learning Techniques. *Sensors*, *22*(13), 4670. https://doi.org/10.3390/s22134670

Fang, Y., Zhao, X., Gong, Z., & He, J. (2022). Comparison of machine learning methods for stroke risk prediction. *Journal of Stroke and Cerebrovascular Diseases*, *31*.

Feigin, V. L., Krishnamurthi, R. V., Parmar, P., Norrving, B., Mensah, G. A., Bennett, D. A., Barker-Collo, S., Moran, A. E., Sacco, R. L., Truelsen, T., Davis, S., Pandian, J. D., Naghavi, M., Forouzanfar, M. H., Nguyen, G., Johnson, C. O., Vos, T., Meretoja, A., Murray, C. J. L., & Roth, G. A. (2015). Update on the Global Burden of Ischemic and Hemorrhagic Stroke in 1990-2013: The GBD 2013 Study. *Neuroepidemiology*, *45*(3), 161–176. https://doi.org/10.1159/000441085

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., & Herrera, F. (2018). *Learning from Imbalanced Data Sets*. Springer International Publishing. https://doi.org/10.1007/978-3-319-98074-4

Gibson, A. D., White, N. M., Collins, G. S., & Barnett, A. G. (2026). Evidence of unreliable data and poor data provenance in clinical prediction model research and clinical practice. *BMC Medicine*, *24*(1), 386. https://doi.org/10.1186/s12916-026-04981-y

Gorelick, P. B. (2020). Risk factors for stroke. *Stroke*, *51*(3).

Hankey, G. J. (2017). Stroke. *The Lancet*, *389*(10069), 641–654. https://doi.org/10.1016/S0140-6736(16)30962-X

Hassan, A., Gulzar Ahmad, S., Ullah Munir, E., Ali Khan, I., & Ramzan, N. (2024). RETRACTED

ARTICLE: Predictive modelling and identification of key risk factors for stroke using machine learning. *Scientific Reports*, *14*(1), 11498. https://doi.org/10.1038/s41598-024-61665-4

Huber, M., Schober, P., Petersen, S., & Luedi, M. M. (2023). Decision curve analysis confirms higher clinical utility of multi-domain versus single-domain prediction models in patients with open abdomen treatment for peritonitis. *BMC Medical Informatics and Decision Making*, *23*(1), 63. https://doi.org/10.1186/s12911-023-02156-w

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 27. https://doi.org/10.1186/s40537-019-0192-5

Kementerian Kesehatan Republik Indonesia. (2018). *Laporan Nasional RISKESDAS 2018*.

Kerr, K. F., Brown, M. D., Zhu, K., & Janes, H. (2016). Assessing the Clinical Impact of Risk Prediction Models With Decision Curves: Guidance for Correct Interpretation and Appropriate Use. *Journal of Clinical Oncology*, *34*(21), 2534–2540. https://doi.org/10.1200/JCO.2015.65.5654

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1137–1143.

Kokkotis, C., Giarmatzis, G., Giannakou, E., Moustakidis, S., Tsatalas, T., Tsiptsios, D., Vadikolias, K., & Aggelousis, N. (2022). An Explainable Machine Learning Pipeline for Stroke Prediction on Imbalanced Data. *Diagnostics*, *12*(10), 2392. https://doi.org/10.3390/diagnostics12102392

Lemaître, G., Nogueira, F., & Aridas, C. K. (2017). Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research*, *18*(17), 1–5.

Little, R., & Rubin, D. (2019). *Statistical Analysis with Missing Data, Third Edition*. Wiley. https://doi.org/10.1002/9781119482260

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 4765–4774.

Melnykova, N., Patereha, Y., Skopivskyi, S., Farion, M., Fedushko, S., & Drohomyretska, K. (2025). Machine learning for stroke prediction using imbalanced data. *Scientific Reports*, *15*(1), 33773. https://doi.org/10.1038/s41598-025-01855-w

Ojeda, F. M., Jansen, M. L., Thiéry, A., Blankenberg, S., Weimar, C., Schmid, M., & Ziegler, A. (2023). Calibrating machine learning approaches for probability estimation: A comprehensive comparison. *Statistics in Medicine*, *42*(29), 5451–5478. https://doi.org/10.1002/sim.9921

Powers, D. M. W. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *International Journal of Machine Learning Technology*, *2*(1), 37–63.

Pozzolo, A. D., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating Probability with Undersampling for Unbalanced Classification. *2015 IEEE Symposium Series on Computational Intelligence*, 159–166. https://doi.org/10.1109/SSCI.2015.33

Riley, R. D., Archer, L., Snell, K. I. E., Ensor, J., Dhiman, P., Martin, G. P., Bonnett, L. J., & Collins, G. S. (2024). Evaluation of clinical prediction models (part 2): how to undertake an external validation study. *BMJ*, *384*, e074820. https://doi.org/10.1136/bmj-2023-074820

Rothwell, P., Coull, A., Silver, L., Fairhead, J., Giles, M., Lovelock, C., Redgrave, J., Bull, L., Welch, S., Cuthbertson, F., Binney, L., Gutnikov, S., Anslow, P., Banning, A., Mant, D., & Mehta, Z. (2005). Population-based study of event-rate, incidence, case fatality, and mortality for all acute vascular events in all arterial territories (Oxford Vascular Study). *The Lancet*, *366*(9499), 1773–1783. https://doi.org/10.1016/S0140-6736(05)67702-1

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLOS ONE*, *10*(3),

e0118432. https://doi.org/10.1371/journal.pone.0118432

Shoaib, M., Ali, N., Ali, S., Khattak, A., & Islam, N. (2023). Stroke prediction using machine learning with oversampling. *Computers in Biology and Medicine*, *161*, 107021.

Soriano, F. (2021). *Healthcare dataset: Stroke prediction*. https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset

Steyerberg, E. W. (2009). *Clinical Prediction Models*. Springer New York. https://doi.org/10.1007/978-0-387-77244-8

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, *17*(1), 230. https://doi.org/10.1186/s12916-019-1466-7

Van Calster, B., Nieboer, D., Vergouwe, Y., De Cock, B., Pencina, M. J., & Steyerberg, E. W. (2016). A calibration hierarchy for risk models was defined: from utopia to empirical data. *Journal of Clinical Epidemiology*, *74*, 167–176. https://doi.org/10.1016/j.jclinepi.2015.12.005

Vickers, A. J., & Elkin, E. B. (2006). Decision Curve Analysis: A Novel Method for Evaluating Prediction Models. *Medical Decision Making*, *26*(6), 565–574. https://doi.org/10.1177/0272989X06295361