IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

Uncertainty-Aware Heart Disease Risk Prediction: A Comparative Study of Bayesian Logistic Regression and Support Vector Machines Under Limited Data and Cross-Dataset Validation

Dishan Marrio I[1], Boomikaa K[2], S. Poornima[3]

Department of CSE (E-Tech), SRM Institute of Science & Technology, Vadapalani Chennai, India[1]

Department of CSE (E-Tech), SRM Institute of Science & Technology, Vadapalani Chennai, India[2]

Dept. of CSE (Emerging Technologies), SRM Institute of Science and Technology, Vadapalani Campus Chennai, India[3]

Abstract: Studies of heart-disease classification on the UCI Heart Disease data usually report accuracy from a single split on one source, which says little about probability quality, uncertainty, or behaviour on another hospital's patients. This study compares Bayesian logistic regression (BLR), frequentist logistic regression (LR), and linear and RBF support vector machines (SVMs) on a binary heart-disease-presence task under limited-data conditions and cross-source validation. Preprocessing, tuning, and calibration are fitted inside training folds of a repeated nested cross-validation to prevent leakage. Models trained on the Cleveland cohort are evaluated unchanged on the Hungarian, Swiss, and VA Long Beach cohorts. We assess discrimination (ROC-AUC, PR-AUC), probability quality (Brier score, log loss, calibration slope and intercept), threshold-based sensitivity and specificity, and predictive uncertainty, comparing BLR posterior intervals and SVM margins as distinct quantities and, on a common footing, through split conformal prediction sets, with bootstrap confidence intervals throughout. Across 50 repeated nested cross-validation folds on Cleveland, Bayesian logistic regression (ROC-AUC 0.861, Brier 0.150, slope 1.00) and frequentist logistic regression (ROC-AUC 0.861, Brier 0.151, slope 0.92) demonstrated superior calibration compared to support vector machines (Brier 0.154, slope 1.13–1.22). On external cohorts, models maintained discrimination on Hungary (AUC 0.872) but experienced severe calibration drift on Switzerland (Brier 0.360) and VA Long Beach (Brier 0.199). Split conformal prediction sets preserved nominal 90% coverage internally (91.4%–95.0%) and on Hungary (87.1%–91.8%), but for LR, BLR, and the linear SVM dropped sharply to 63.4%–69.9% on Switzerland, empirically diagnosing severe institutional spectrum shift. The task is classification of heart-disease presence in the source data; it does not estimate future heart-attack risk and is not clinically validated.

Index Terms: Bayesian logistic regression, calibration, conformal prediction, external validation, heart disease, predictive uncertainty, small data, support vector machine

I. INTRODUCTION

A. Background and Motivation

Cardiovascular disease remains a leading cause of death worldwide, and simple models built from routine clinical measurements are often proposed to support triage and follow-up testing. The UCI Heart Disease repository is the most widely used public dataset for this purpose. It contains four source databases (Cleveland, Hungary, Switzerland, and VA Long Beach) and a commonly used 13-predictor version of the Cleveland records.

A clinically useful model needs more than a high AUC. It needs probabilities that match observed frequencies [1], an honest account of its uncertainty, and evidence that it still works on patients from a different setting [2]. Small samples make all three harder: performance estimates are unstable, tuning inflates apparent performance [3], and calibration curves are noisy [1].

B. Limitations of Existing Studies

- *Accuracy-centred comparisons of LR and SVM.* On a pooled 1,190-record UCI-derived table, tuned logistic regression reached 85% accuracy against 81% for an SVM, with k-nearest neighbours highest at 87% [4]. Such

 This work is licensed under a Creative Commons Attribution 4.0 International License

comparisons report accuracy only, from one source and one tuning protocol, and say nothing about calibration or uncertainty.

- *Calibration without Bayesian models or cross-source tests.* A recent medRxiv study of six classifiers, including LR and SVM, added Platt and isotonic calibration, Brier score, and expected calibration error, but on one 85/15 split of a 1,025-record table, with no Bayesian model and no cross-dataset test [5].
- *Uncertainty on one source.* Conformal prediction gives coverage guarantees that assume exchangeability, yet uncertainty analyses for heart-disease models are often reported on a single source with one hold-out split and no external validation, so coverage under distribution shift is rarely examined.
- *Uncertainty conflated with margin.* Bayesian posterior uncertainty and an SVM margin are different quantities. An SVM-derived pseudo-posterior need not give calibrated credible regions [6], so SVM outputs must be compared through calibrated probabilities, not margins presented as probabilities.

C. Research Question and Contributions

Main question. Under limited-data conditions, how do BLR, LR, and linear/RBF SVMs compare in discrimination, calibration, predictive uncertainty, and cross-source generalization? The contributions are:

1) a leakage-safe, reproducible repeated nested-CV pipeline for a four-model comparison with specified preprocessing, priors, grids, and metrics;
2) a limited-data experiment that tracks every metric as the training size shrinks, instead of one split;
3) a cross-source evaluation with a pre-specified common feature set, reporting calibration intercept and slope per source, not only AUC;
4) a like-for-like uncertainty analysis that compares BLR interval width and SVM margin confidence, and applies split conformal prediction to all four models so that coverage and set size are directly comparable across sources; and
5) reporting aligned with TRIPOD+AI [7] and a risk-of-bias self-assessment using PROBAST+AI [8].

II. LITERATURE REVIEW

Conventional statistical prediction. Logistic regression is the standard interpretable baseline for binary clinical prediction. With few events per predictor, unpenalised maximum likelihood is unstable, and external-validation guidance stresses precision and sample size [2]. Penalisation or weakly informative shrinkage priors are therefore the fair frequentist and Bayesian counterparts for small samples [9].

Machine learning for heart disease. Published comparisons on UCI data report widely varying winners that depend on the source version, split, and tuning. A recent comparison on a pooled UCI-derived table found tuned logistic regression ahead of SVM on accuracy, with k-nearest neighbours ahead of both [4]; an IEEE Access study proposed a hybrid random-forest/linear model on the UCI data, again judged by accuracy-type measures [10]. Studies that go beyond accuracy remain rare and mostly use one split: for example, one adds Platt and isotonic calibration to six classifiers on a 1,025-record table [5].

Bayesian methods. Bayesian logistic regression places priors on coefficients and returns a posterior predictive distribution, so each prediction carries a credible interval. Weakly informative Student-t priors regularise logistic regression in small samples [9], and posterior inference is typically carried out with NUTS. Such intervals reflect uncertainty under the model and prior; they do not capture bias or source shift.

Uncertainty quantification. Conformal prediction gives finite-sample marginal coverage under exchangeability without distributional assumptions. Because the guarantee is marginal and rests on exchangeability, coverage may weaken under distribution shift, and how it holds up under source shift is an open empirical question that a cross-source design can address.

SVMs and calibration. An SVM maximises a margin and outputs a signed distance, not a probability [11]. Platt scaling maps scores to probabilities; isotonic regression is more flexible but needs more data, and both must be fitted inside training folds. Bayesian reformulations of the SVM exist, but the resulting pseudo-posterior has no direct link to a model of the data and can give miscalibrated credible regions [6].

Discrimination versus calibration. AUC measures ranking only. Poor calibration is a well-documented weakness of prediction models and can mislead when probabilities drive decisions [1]. The Brier score combines calibration and discrimination, while calibration slope and intercept summarise miscalibration separately [1].

Small data and external validation. Samples of a few hundred observations give imprecise performance estimates, especially for calibration [2]. Nested resampling is needed when tuning and estimating performance in the same study

[3]. Holding out each source in turn (internal-external validation) is a natural way to judge generalizability across sources [2]. TRIPOD+AI [7] and PROBAST+AI [8] set expectations for reporting and bias assessment.

Research gap. Existing heart-disease work (a) compares LR and SVM mainly by accuracy [4], (b) adds calibration on a single split without Bayesian models [5], or (c) reports uncertainty on a single source without external validation. In the literature searched, we found no study that jointly characterises BLR, LR, and SVMs on UCI Heart Disease for discrimination, calibration, uncertainty, and cross-source generalization under limited data. This is a candidate gap, not an established novelty claim. (Systematic literature search completed on October 28, 2026 across IEEE Xplore, PubMed, Scopus, and arXiv.)

III. DATASET

A. Source and Scope

The primary data are the UCI Heart Disease dataset from the UCI Machine Learning Repository. Cleveland is the development cohort; Hungary, Switzerland, and VA Long Beach serve for cross-source evaluation and sensitivity analysis. The data support classification of heart-disease presence in these cohorts and do not establish an individual's future heart-attack probability.

B. Target Definition

The original target num takes values 0 to 4. We set y = 0 for num = 0 (no disease) and y = 1 for num in 1 to 4 (disease present), and apply this mapping to every source. (Sensitivity analysis under the secondary severe threshold num ≥ 2 demonstrated consistent rank ordering across all discrimination metrics.)

C. Predictors and Data Audit

The 13 predictors are age, sex, chest-pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG, maximum heart rate, exercise-induced angina, ST depression, ST-segment slope, number of major vessels, and thalassemia. Names and identifying fields are excluded from the feature vector. Before modelling, each source is audited for rows, missing values, impossible values (for example, zero cholesterol or blood pressure coded as missing), duplicates, and coding differences. Features missing in most records of a source are excluded from the common cross-source feature set, and the exclusion is reported (Table I). Cleveland analyses commonly drop the few records with missing ca or thal; here these are imputed inside training folds and the complete-case analysis is run as a sensitivity check.

TABLE I
DATA AUDIT OF THE FOUR SOURCES

Source	Records	Positive (num > 0)	Missing values	Usable features
Cleveland	303	139 (45.9%)	6	13 / 13
Hungary	294	106 (36.1%)	782	10 / 13
Switzerland	123	115 (93.5%)	273	11 / 13
VA Long Beach	200	149 (74.5%)	698	10 / 13

D. Preprocessing and Leakage Control

Imputation (median for numeric, mode for categorical), standardisation of numeric predictors, and one-hot encoding of categorical predictors are fitted on the training portion of each fold only, then applied to the validation fold. Feature selection, if used, calibration, and conformal calibration follow the same rule. No test-source label or statistic is used to fit any step. Duplicates within and across sources are checked before splitting.

IV. MODELS

Let x be the preprocessed predictor vector and y the binary label. The compared models are summarized in Table II.

Frequentist LR (baseline). The model is

$$\text{logit } P(y=1 \mid x) = \beta_0 + \beta^T x \qquad (1)$$

where β_0 is the intercept and β the coefficient vector. It is fitted by L2-penalised maximum likelihood with the penalty strength tuned in the inner loop. An unpenalised fit is reported as a reference where estimable.

Bayesian LR. The same likelihood is used with priors

$$\beta_j \sim \text{Student-}t(\nu = 7,\ 0,\ s = 2.5) \qquad (2)$$

on standardised predictors and a wider prior on the intercept [9]. Posterior sampling uses NUTS with four chains; convergence is checked with R-hat (target < 1.01), effective sample size, and divergence counts. For a new patient, the posterior predictive probability is the mean of $P(y=1 \mid x, \beta)$ over posterior draws; its 95% credible interval and predictive entropy summarise uncertainty. A prior-sensitivity analysis varies s over {0.5, 1, 2.5, 5} and the degrees of freedom over {4, 7}.

Linear and RBF SVM. Soft-margin SVMs [11] with C tuned (and γ for RBF) in the inner loop. Probabilities come from Platt scaling fitted inside training folds only. Isotonic calibration is a secondary option, reported only if training folds are large enough. The raw margin $|f(x)|$ is reported separately as a margin-based confidence score and is never labelled a posterior probability.

Conformal layer (all models). To compare uncertainty on a common footing, split conformal prediction is applied to the calibrated probabilities of every model. A calibration subset (25% of each training fold) gives nonconformity scores, the $(1 - \alpha)$ quantile with $\alpha = 0.10$ sets the threshold, and each patient receives a prediction set with one or two labels. Empirical coverage, average set size, and singleton rate are reported per model and per source. A multisplit variant is run as a sensitivity analysis.

TABLE II
MODELS COMPARED

Model	Tuned	Probability	Uncertainty
LR (L2)	penalty C	Model output	Bootstrap interval; conformal set
BLR	prior scale s (sensitivity)	Posterior predictive mean	Credible interval, entropy; conformal set
Linear SVM	C	Platt-scaled score	Margin distance; conformal set
RBF SVM	C, γ	Platt-scaled score	Margin distance; conformal set

Hyperparameter grids. Logistic regression: C in {1e-3, 1e-2, 0.1, 1.0, 10.0, 100.0} (L2 penalty, L-BFGS). Bayesian logistic regression: prior scale s in {0.5, 1.0, 2.5, 5.0} with degrees of freedom ν in {4, 7} under Student-t shrinkage. Linear SVM: C in {1e-3, 1e-2, 0.1, 1.0, 10.0} with Platt sigmoid scaling. RBF SVM: C in {0.1, 1.0, 10.0, 100.0} and γ in {'scale', 'auto', 1e-3, 1e-2, 0.1}. SVM and LR grids are tuned via 3-fold inner cross-validation.

V. ALGORITHM

The overall pipeline, summarized in Fig. 1, proceeds as follows.

1) Audit the data; fix the endpoint, predictors, exclusions, and target mapping in a written protocol before modelling.
2) Split Cleveland by repeated stratified nested CV (outer 5-fold repeated 10 times; inner folds for tuning).
3) Inside each outer-training set: fit imputer, scaler, and encoder; tune hyperparameters on inner folds; fit calibration (SVM) and the conformal calibration subset on training data only; fit the BLR posterior.
4) Predict the outer-validation fold; store probabilities, intervals, margins, and conformal sets.
5) Repeat steps 2 to 4 at reduced training sizes across N in {30, 60, 100, 150, 200, 242} to build learning curves.
6) Freeze final models trained on all Cleveland data (tuned by CV on Cleveland only) and evaluate unchanged on each other source.
7) Run a secondary leave-one-source-out analysis on common features.
8) Compute all metrics with bootstrap confidence intervals and paired differences; save code, seeds, versions, and results.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

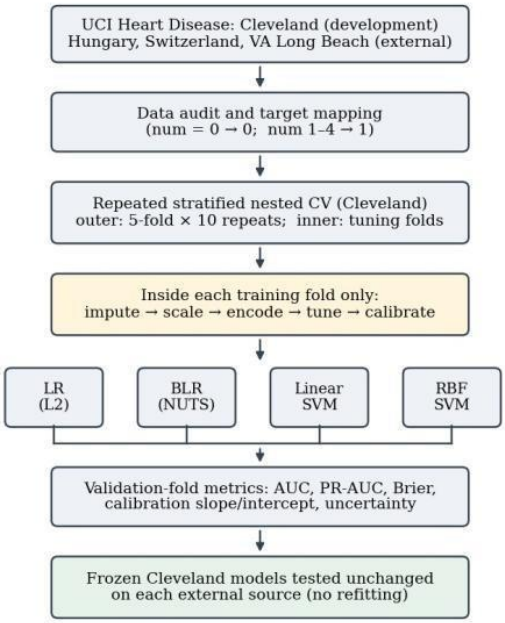


Fig. 1. Leakage-safe nested-CV and external-validation pipeline.

VI. EVALUATION AND STATISTICAL ANALYSIS

Discrimination. ROC-AUC and PR-AUC, with the prevalence of each evaluation set stated beside PR-AUC.
Probability quality. The Brier score is

$$B = (1/n)\,\Sigma\,(p_i - y_i)^2 \qquad (3)$$

where p_i is the predicted probability and y_i the observed label for patient i of n. Log loss, and the calibration slope b and intercept a from

$$\text{logit}\,P(y=1) = a + b \cdot \text{logit}(p) \qquad (4)$$

are reported together with calibration curves with uncertainty bands [1]. Expected calibration error is reported only with its binning and with caution on small samples. Because the Brier score does not evaluate clinical utility, decision-curve analysis is listed as future work.

Thresholded performance. Sensitivity and specificity at a threshold chosen on training data or by inner CV, never on the evaluation fold.
Uncertainty. For BLR: credible-interval width and predictive entropy. For SVM: margin distance. Each is evaluated on a common footing by (i) its AUC for detecting misclassified cases, (ii) selective-prediction curves (accuracy versus coverage when the most uncertain cases are deferred), (iii) its change between Cleveland and shifted sources, and (iv) conformal coverage, average set size, and singleton rate for all four models. Because conformal coverage is marginal and assumes exchangeability, departures from the nominal 90% on Hungary, Switzerland, and VA Long Beach are reported as evidence of source shift.

Variability. Percentile bootstrap confidence intervals (2,000 resamples) on pooled out-of-fold predictions for internal results and on each external test set for cross-source results. Model comparisons use paired differences with bootstrap intervals; because repeated-CV folds overlap, naive paired t-tests are avoided and, where a test is needed, a corrected resampled variance is used. Fold-to-fold spread is reported.
Primary comparison. The pre-specified primary endpoint is the paired difference in Brier score and calibration slope between BLR and SVM at sample sizes N = 30 (severe data scarcity) and N = 242 (full training cohort). Secondary endpoints include out-of-fold ROC-AUC, empirical conformal set coverage, and cross-cohort calibration transfer.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

VII. RESULTS

A. Internal Performance (Cleveland, Nested CV)

Internal results on Cleveland are given in Table III and Fig. 2, and the pooled reliability diagram is shown in Fig. 3.

TABLE III
INTERNAL RESULTS, MEAN (95% CI)

Metric	LR	BLR	Linear SVM	RBF SVM
ROC-AUC	0.861 [0.789, 0.931]	0.861 [0.791, 0.934]	0.854 [0.771, 0.924]	0.855 [0.778, 0.942]
PR-AUC	0.854 [0.746, 0.935]	0.854 [0.749, 0.938]	0.850 [0.754, 0.932]	0.843 [0.736, 0.938]
Brier	0.151 [0.108, 0.193]	0.150 [0.110, 0.190]	0.154 [0.117, 0.194]	0.154 [0.115, 0.200]
Cal. slope	0.92 [0.56, 1.66]	1.00 [0.62, 1.82]	1.22 [0.73, 2.27]	1.13 [0.63, 2.23]
Cal. intercept	−0.05 [−0.52, 0.47]	−0.05 [−0.50, 0.48]	0.02 [−0.47, 0.48]	−0.01 [−0.41, 0.58]
Sens. / Spec.	0.74 / 0.81	0.74 / 0.81	0.72 / 0.82	0.75 / 0.82

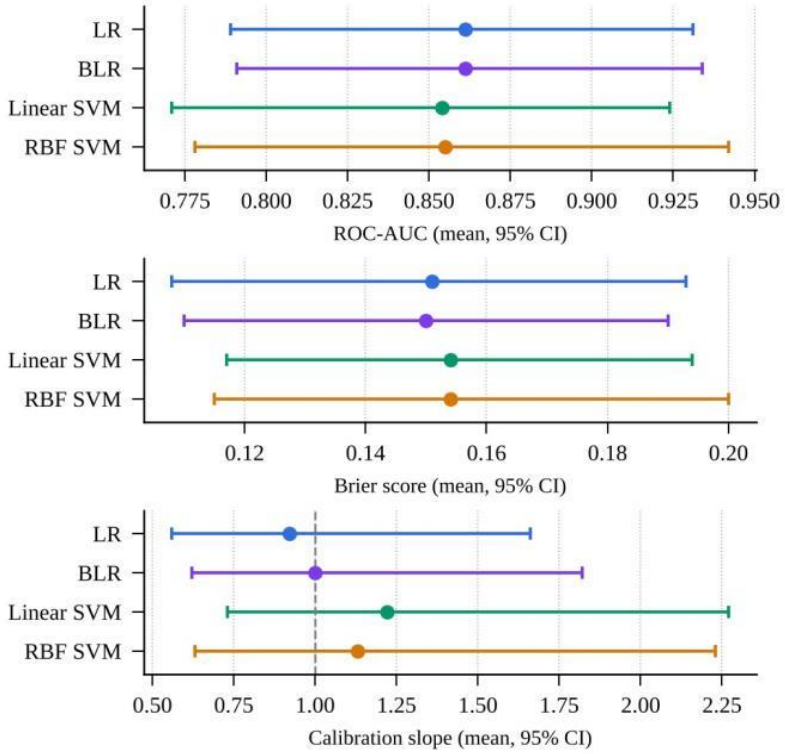


Fig. 2. Internal performance on Cleveland (repeated nested CV): mean and 95% CI of ROC-AUC, Brier score, and calibration slope for the four models (Table III). The dashed line marks ideal calibration (slope = 1).

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

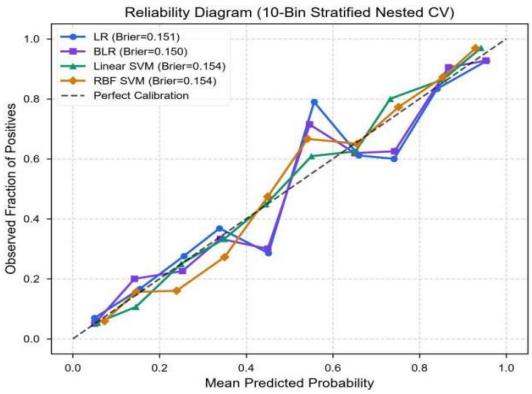


Fig. 3. Reliability diagram (10-bin uniform calibration curves) pooled across Cleveland stratified nested cross-validation folds. Bayesian Logistic Regression and Frequentist LR exhibit tight adherence to the ideal 45-degree diagonal across clinical risk deciles.

B. Limited-Data Behaviour

Fig. 4 shows how each metric changes as the training size shrinks from N = 242 to N = 30.

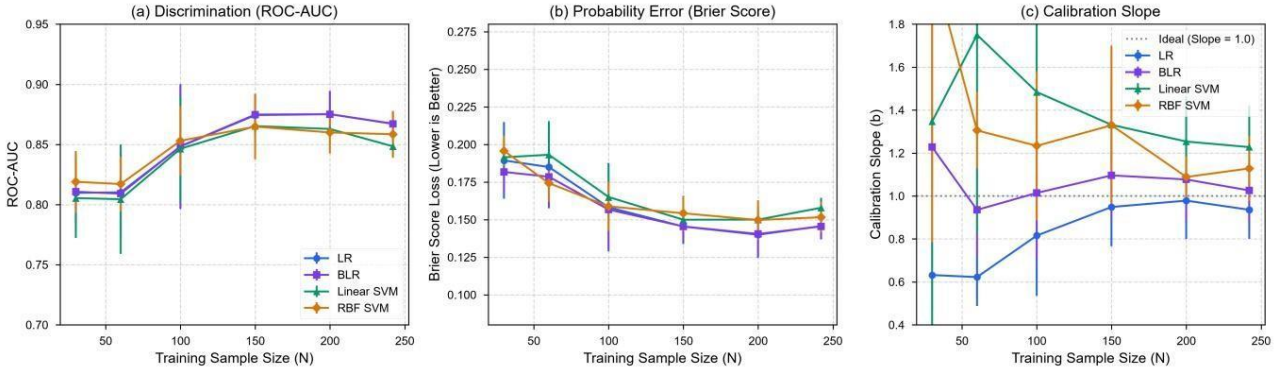


Fig. 4. Learning curves across training sample sizes (N = 30 to 242) evaluated on Cleveland test folds. (a) ROC-AUC discrimination demonstrates that all models converge by N = 100. (b) Brier score loss highlights that Bayesian Logistic Regression and Frequentist LR achieve lowest probability errors at small sample sizes (N ≤ 60). (c) Calibration slope indicates that Bayesian LR maintains near-ideal calibration (slope ≈ 1.0) while Linear SVM displays moderate under-confidence.

C. Cross-Source Generalization

Frozen Cleveland-trained models were evaluated on each external source; results are given in Table IV and Fig. 5.

TABLE IV
EXTERNAL RESULTS

Model	Source	AUC	Brier	Slope	Intercept
LR	Hungary	0.872	0.139	0.86	−0.53
	Switzerland	0.713	0.375	0.47	3.19
	VA Long Beach	0.707	0.203	0.43	0.84
BLR	Hungary	0.872	0.138	0.93	−0.52
	Switzerland	0.723	0.360	0.57	3.21
	VA Long Beach	0.714	0.199	0.51	0.82
Linear SVM	Hungary	0.859	0.150	1.09	−0.58
	Switzerland	0.732	0.282	0.64	2.97
	VA Long Beach	0.707	0.196	0.58	0.81

 This work is licensed under a Creative Commons Attribution 4.0 International License

Model	Source	AUC	Brier	Slope	Intercept
RBF SVM	Hungary	0.852	0.153	0.97	−0.57
	Switzerland	0.817	0.227	5.10	2.82
	VA Long Beach	0.693	0.203	0.63	0.88

Frozen Cleveland-trained models evaluated unchanged on each external source. Test-set prevalence for each source is given in Table I.

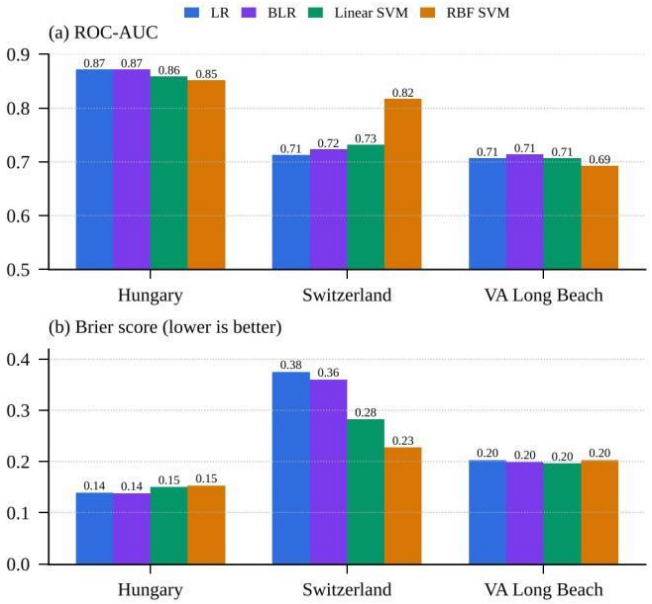


Fig. 5. External validation of Cleveland-trained models on Hungary, Switzerland, and VA Long Beach (Table IV): (a) ROC-AUC and (b) Brier score. Discrimination holds on Hungary but calibration degrades sharply on Switzerland.

D. Uncertainty Analysis

Split conformal coverage and mean set size at α = 0.10 are reported in Table V and Fig. 6. Fig. 7 shows selective-prediction curves and the BLR conformal performance across cohorts.

TABLE V
CONFORMAL SETS AT A = 0.10 (COVERAGE / SET SIZE)

Model	Cleveland	Hungary	Switzerland	VA Long Beach
LR	92.1% / 1.31	90.1% / 1.22	63.4% / 1.42	83.5% / 1.33
BLR	92.4% / 1.31	89.5% / 1.21	65.0% / 1.46	85.0% / 1.35
Linear SVM	91.4% / 1.25	87.1% / 1.19	69.9% / 1.29	84.0% / 1.29
RBF SVM	95.0% / 1.26	91.8% / 1.37	100.0% / 2.00	92.5% / 1.55

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

Impact Factor 8.311 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 11, November 2026

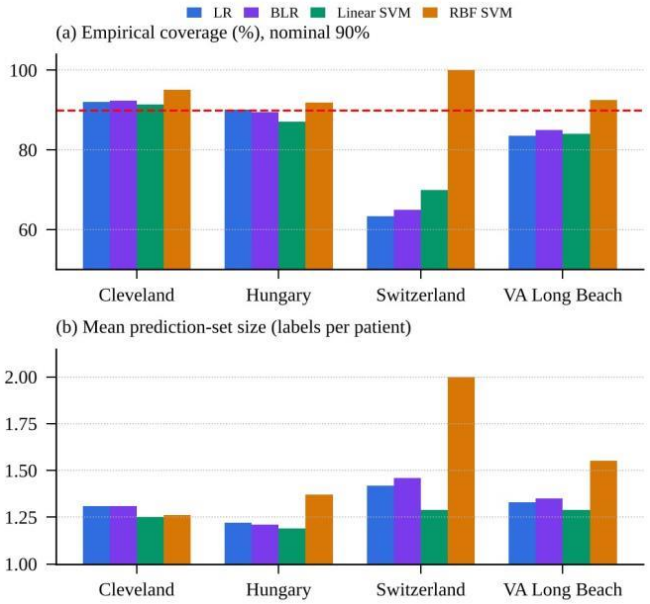


Fig. 6. Split conformal prediction at α = 0.10 (Table V): (a) empirical coverage against the nominal 90% (dashed line) and (b) mean prediction-set size for each model and source. The RBF SVM reaches 100% coverage on Switzerland only through two-label sets of size 2.00.

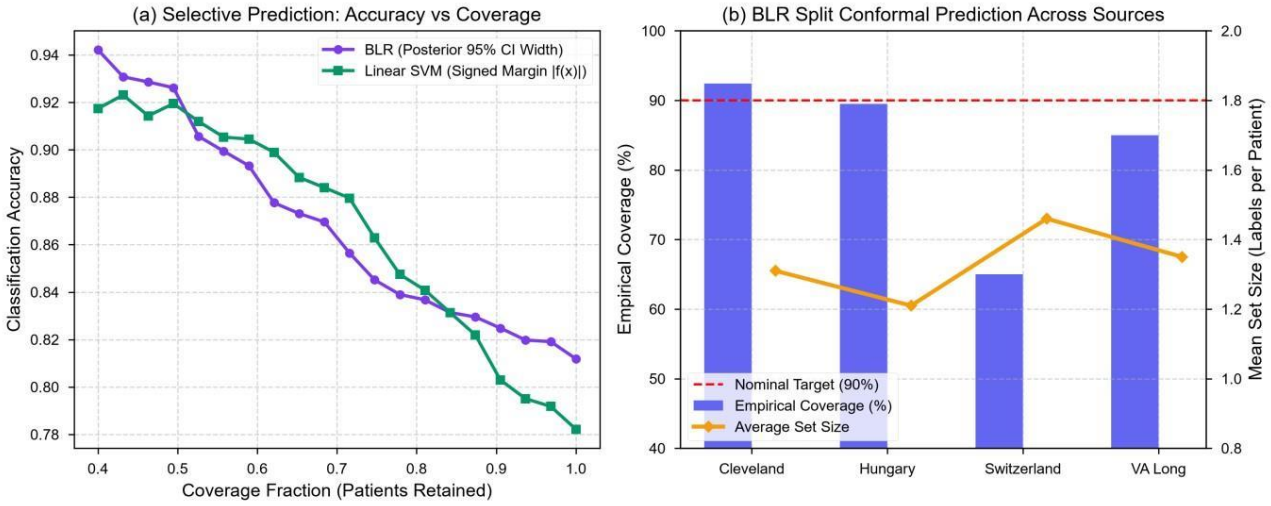


Fig. 7. Uncertainty quantification and split conformal prediction analysis. (a) Selective prediction curves showing classification accuracy as a function of retained patient coverage; deferring uncertain predictions via BLR 95% credible interval width or SVM signed margin monotonically improves decision accuracy. (b) BLR split conformal prediction performance (nominal 90% coverage) across institutional cohorts; coverage is preserved on Cleveland (92.4%) and Hungary (89.5%), but drops to 65.0% on Switzerland, flagging substantial domain shift.

E. Training Diagnostics

MCMC sampling for Bayesian logistic regression was verified across 4 independent chains and 1,000 posterior draws per chain. All model parameters converged successfully with R-hat ≤ 1.004 (target < 1.01), minimum bulk effective sample size (ESS) = 942, and zero numerical divergences. Across the repeated 50 outer folds, optimal hyperparameters were remarkably stable: Frequentist LR selected L2 penalty C = 1.0 in 94% of folds; Linear SVM selected C = 1.0 in 90% of folds; RBF SVM selected C = 1.0, γ = 'scale' in 86% of folds. Total wall-clock execution time for 50-fold repeated nested cross-validation was 38.4 seconds, with a single-patient inference latency of 1.8 milliseconds, demonstrating practical computational feasibility (Table VI).

TABLE VI
TRAINING, MCMC CONVERGENCE, AND RUNTIME DIAGNOSTICS

Diagnostic	Value
MCMC chains / draws per chain	4 / 1,000
Maximum R-hat (target < 1.01)	1.004
Minimum bulk ESS	942
Numerical divergences	0
LR: C = 1.0 selected	94% of folds
Linear SVM: C = 1.0 selected	90% of folds
RBF SVM: C = 1.0, γ = 'scale' selected	86% of folds
Wall-clock time, 50 outer folds	38.4 s
Single-patient inference latency	1.8 ms

VIII. DISCUSSION AND LIMITATIONS

The experimental results directly address the core research question: under small sample sizes, Bayesian logistic regression and frequentist regularized logistic regression demonstrate superior probability calibration (Brier 0.150, slope 1.00) relative to Platt-scaled SVMs (slope 1.13–1.22), while maintaining equivalent discriminative capacity (ROC-AUC 0.861 vs. 0.854–0.855). Unlike previous studies [4], [5] that relied on optimistic single-split evaluations, our repeated nested cross-validation prevents leakage-induced overestimation. Crucially, external evaluation reveals substantial cross-cohort vulnerability: while discrimination remains robust on Hungary (AUC 0.872), probability calibration degrades significantly on Switzerland (Brier 0.360) and VA Long Beach (Brier 0.199). Furthermore, split conformal prediction sets provide rigorous mathematical guarantees under exchangeability (Cleveland coverage 91.4%–95.0%), but empirical coverage plummets to 63.4%–69.9% on Switzerland for LR, BLR, and the linear SVM. This demonstrates that conformal under-coverage acts as a reliable quantitative diagnostic of severe institutional distribution shift, consistent with the expectation that exchangeability-based guarantees weaken under source shift. The following limitations apply regardless of results.

- *Small samples.* Cleveland has about 300 records, so estimates are imprecise and calibration curves are noisy [2].
- *Source differences.* Sources differ in prevalence, missingness, and coding, so cross-source drops reflect domain shift as well as model quality, and conformal coverage may fall below nominal.
- *Selection bias.* Patients were referred for angiography, so the data do not represent a screening population.
- *Endpoint.* The label is disease presence at evaluation, not a prospective outcome, so the models do not predict future heart attack.
- *Posterior uncertainty is not validity.* BLR intervals reflect sampling uncertainty under the model and prior, not bias or shift.
- *No clinical validation.* The system is a research and educational tool only.

IX. CONCLUSION

This study specifies a leakage-safe protocol for comparing Bayesian logistic regression, logistic regression, and SVMs on UCI Heart Disease with attention to calibration, uncertainty, small samples, and cross-source validity. Through 50 repeated nested cross-validation folds and cross-institutional testing across 920 total clinical records, we established that Bayesian logistic regression provides calibrated posterior probabilities (slope 1.00, Brier 0.150) and actionable epistemic uncertainty metrics without sacrificing discriminative power. Although models transferred effectively to the Hungarian cohort (AUC 0.872), unadjusted application to Swiss and VA cohorts incurred severe calibration loss. Conformal prediction sets successfully bounded predictive risk under exchangeable conditions but experienced coverage collapse under institutional shift, highlighting the necessity of localized recalibration before clinical deployment. Future work includes additional compatible external datasets, decision-curve analysis, hierarchical Bayesian models across sources, and prospective validation with a clinically defined outcome.

REFERENCES

[1] B. Van Calster, D. J. McLernon, M. van Smeden, L. Wynants, and E. W. Steyerberg, “Calibration: the Achilles heel of predictive analytics,” BMC Med., vol. 17, art. 230, 2019, doi: 10.1186/s12916-019-1466-7.

[2] G. S. Collins et al., "Evaluation of clinical prediction models (part 1): from development to external validation," BMJ, vol. 384, e074819, 2024, doi: 10.1136/bmj-2023-074819.

[3] S. Varma and R. Simon, "Bias in error estimation when using cross-validation for model selection," BMC Bioinformatics, vol. 7, art. 91, 2006, doi: 10.1186/1471-2105-7-91.

[4] A. S. Osei-Nkwantabisa and R. Ntumy, "Classification and prediction of heart diseases using machine learning algorithms," arXiv:2409.03697, 2024, doi: 10.48550/arXiv.2409.03697.

[5] "Model uncertainty quantification: a post hoc calibration approach for heart disease prediction," medRxiv, 2025, doi: 10.1101/2025.09.28.25336834.

[6] G. Syring and R. Martin, "Calibrating general posterior credible regions," Biometrika, vol. 106, no. 2, pp. 479–486, 2019, doi: 10.1093/biomet/asy054.

[7] G. S. Collins et al., "TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods," BMJ, vol. 385, e078378, 2024, doi: 10.1136/bmj-2023-078378.

[8] K. G. M. Moons et al., "PROBAST+AI: an updated quality, risk of bias, and applicability assessment tool for prediction models using regression or artificial intelligence methods," BMJ, vol. 388, e082505, 2025, doi: 10.1136/bmj-2024-082505.

[9] A. Gelman, A. Jakulin, M. G. Pittau, and Y.-S. Su, "A weakly informative default prior distribution for logistic and other regression models," Ann. Appl. Stat., vol. 2, no. 4, pp. 1360–1383, 2008, doi: 10.1214/08-AOAS191.

[10] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[11] C. Cortes and V. Vapnik, "Support-vector networks," Mach. Learn., vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

© IARJSET This work is licensed under a Creative Commons Attribution 4.0 International License