Jurnal Riset Mahasiswa Matematika
Volume 4 (6) (2025), Pages 323-336
p-ISSN: 2808-1552; e-ISSN: 2808-4926

Cross-Dataset Evaluation of Support Vector Machines: A Reproducible, Calibration-Aware Baseline for Tabular Classification

Nurus Syafi'ah, Mohammad Jamhuri*, Farahnas Imaniyah Pranata, Ari Kusumastuti, Usman Pagalay, and Muhammad Khudzaifah

Mathematics Study Program, Faculty of Science and Technology, UIN Maulana Malik Ibrahim Malang, Indonesia

Abstract

Support Vector Machines (SVMs) remain competitive for small and medium-sized tabular classification problems, yet reported results on benchmark datasets vary widely due to inconsistent preprocessing, validation, and probability calibration. This paper presents a calibration-aware, cross-dataset benchmark that evaluates SVMs against classical baselines—Logistic Regression, Decision Tree, and Random Forest—under leakage-safe pipelines and statistically principled protocols. Using three representative binary datasets (Titanic survival, Pima Indians Diabetes, and UCI Heart Disease), we standardize imputation, encoding, scaling, and nested cross-validation to ensure comparability. Performance is assessed not only on discrimination metrics (accuracy, precision, recall, F1, PR–AUC) but also on probability reliability (Brier score, Expected Calibration Error) and threshold optimization. Results show that tuned RBF–SVMs consistently outperform Logistic Regression and Decision Trees, and perform comparably to Random Forests. Calibration (Platt scaling, isotonic regression) substantially reduces error and improves decision quality, while domain-specific features enhance Titanic prediction. By embedding all steps in a transparent, reproducible protocol and validating across multiple datasets, this study establishes a rigorous methodological baseline for SVMs in tabular binary classification, providing a reference point for future machine learning research.

Keywords: Tabular classification; Support Vector Machine; Probability calibration; Cross-dataset benchmarking; Small datasets

1 Introduction

Tabular classification remains one of the most common applications of machine learning in domains such as health screening, diagnostics, and socio–economic studies. Support Vector Machines (SVMs) are still competitive in these settings, particularly for small to medium-sized datasets [1], [2], but reported performance often varies substantially across studies. Much of this variation arises not from the algorithms themselves, but from inconsistencies in preprocessing, feature engineering, validation protocols, and the reporting of probability calibration. Consequently, comparisons such as "model A outperforms model B" are frequently confounded by leakage, imbalance, or thresholding choices [3] rather than reflecting genuine algorithmic differences.

*[*] **Corresponding author**. E-mail: m.jamhuri@yahoo.com

Submitted: May 28, 2025 Reviewed: June 21, 2025 Accepted: August 31, 2025
DOI: https://dx.doi.org/10.18860/jrmm.v4i6.33438

Recent methodological work has emphasized the need for standardized, leakage–safe pipelines and calibration–aware evaluations to ensure that conclusions are reproducible and transferable across datasets [4]. Calibration is particularly important because many classifiers produce poorly calibrated probabilities, which undermines decision quality in imbalanced settings [5], [6]. In this paper, we address these concerns by presenting a calibration-aware, cross-dataset benchmark centered on rigorously tuned SVMs, compared side-by-side with strong classical baselines (Logistic Regression, Decision Tree, Random Forest). Using three representative binary datasets—Titanic survival, Pima Indians Diabetes, and UCI Heart Disease—we unify preprocessing, feature engineering, cross-validation, and calibration under a single transparent protocol. This design enables us to answer whether observed SVM advantages hold consistently across datasets, and whether calibration and threshold optimization materially improve decision quality.

Building on this motivation, our study addresses four methodological questions:

RQ1. How does a rigorously tuned RBF–SVM compare with strong classical baselines (Logistic Regression, Decision Tree, Random Forest) across three representative tabular datasets?

RQ2. To what extent do probability calibration methods (Platt scaling, isotonic regression) and threshold optimization improve decision quality?

RQ3. Are model rankings and calibration gains stable across heterogeneous datasets (demographic, clinical, socio–economic), or do they remain dataset–specific?

RQ4. Are observed differences statistically significant when tested with paired comparisons and bootstrap confidence intervals?

In addressing these questions, the paper makes several contributions. First, it introduces a reproducible evaluation protocol for tabular binary classification, embedding all preprocessing steps in leakage–safe pipelines and employing nested cross-validation. Second, it offers a calibration-aware benchmark of RBF–SVMs against Logistic Regression, Decision Tree, and Random Forest, with systematic reporting of both discrimination and calibration metrics. Third, it provides an analysis of threshold optimization, showing how calibrated probabilities improve F1 and decision quality under class imbalance. Fourth, it presents cross-dataset evidence demonstrating when SVM advantages generalize and when they converge with tree ensembles. Finally, the work is accompanied by a full replication package (code, parameter grids, environment lockfiles, and figure scripts) to facilitate transparent and reproducible research.

The remainder of this paper is organized as follows. Section 2 reviews related work on leakage, calibration, and reproducibility in tabular classification. Section 3 introduces the datasets and problem framing. Section 4 details the evaluation protocol, including preprocessing, cross-validation, calibration, and threshold optimization. Section 5 presents experimental results across the three datasets, while Section 6 discusses findings, limitations, and threats to validity. Finally, Section 7 concludes with a replication checklist and directions for future research.

2 Related Work

Research on tabular binary classification spans several strands: (i) algorithmic comparisons on widely used benchmarks such as *Titanic*, (ii) methodological studies addressing data leakage, class imbalance, calibration, and validation, and (iii) reproducibility practices in applied machine learning. Our work positions itself at the intersection of these strands by providing a calibration–aware, cross-dataset evaluation under a leakage–safe protocol.

The *Titanic* dataset is one of the most frequently used teaching benchmarks, with moderate class imbalance and heterogeneous feature types. Numerous studies report results with classical learners (e.g., Logistic Regression, SVM, Decision Tree, Random Forest, Boosting), but often under differing preprocessing pipelines and evaluation metrics, leading to inconsistent model rankings. These discrepancies typically arise from differences in imputation strategies, feature engineering (e.g., `Title`, `FamilySize`, `IsAlone`), and validation design. While valuable as an educational dataset, *Titanic* alone provides limited novelty. We therefore treat it as one member of a broader cross–dataset study, complemented by *Pima Indians Diabetes* and *UCI Heart Disease (Cleveland)* to assess generalizability beyond didactic settings.

Data leakage—fitting preprocessing steps on the full dataset prior to cross-validation—is a well-documented source of optimistic bias in reported results [3], [7]. Best practice is to encapsulate all preprocessing (imputation, encoding, scaling) in a unified pipeline that is fit exclusively within each training fold. We adopt this design for all models to ensure leakage–free evaluation.

Accuracy alone is unreliable under class imbalance; metrics such as F1, precision–recall curves, PR–AUC, and cost-sensitive measures are more informative in these regimes [8], [9]. Remedies include stratified splits, class weighting, and resampling techniques (e.g., SMOTE). In our work, we prioritize macro and weighted F1 as well as PR–AUC, use stratified cross-validation, and test class weighting for SVM and Logistic Regression in ablation studies.

Well-calibrated probabilities are essential for decision-making, particularly under imbalanced or cost-sensitive conditions. Platt scaling [10] and isotonic regression [11], [12] are widely used post-hoc calibration methods. Calibration quality is commonly summarized by Brier score and Expected Calibration Error (ECE) [5], [6]. Beyond calibration, threshold optimization is critical for aligning classifier decisions with performance criteria such as F1 or expected cost. We therefore compare raw and calibrated probabilities and evaluate threshold optimization for decision quality.

Hyperparameter tuning outside of nested cross-validation can lead to selection bias. Nested CV or dedicated validation splits are recommended [13]. For paired comparisons, McNemar's test [14] on discordant errors and bootstrap confidence intervals for F1/PR–AUC are common practices. For multiple datasets and algorithms, nonparametric tests such as Friedman with Nemenyi post-hoc are suggested [15]. Our study follows these practices by employing nested CV for tuning, paired testing on predictions, and bootstrap confidence intervals.

Support Vector Machines [1] remain competitive for small to medium-sized tabular datasets, especially when combined with RBF kernels and standardized features. Tree-based ensembles (Random Forest, Gradient Boosting) are frequently reported to outperform kernel methods in some tabular domains, depending on feature interactions and noise levels [2], [16], [17]. Our study does not aim to introduce new architectures, but rather to provide a rigorously tuned SVM baseline and to assess how calibration and evaluation protocol affect comparative outcomes against strong baselines.

Interpretability tools such as SHAP provide insights into feature contributions and stability in tabular settings [18]. For *Titanic*, engineered features such as `Title`, `FamilySize`, and `IsAlone` are widely used, but their incremental benefits are seldom quantified under proper cross-validation. We therefore include ablation experiments to measure their marginal contributions and complement these with interpretability analyses.

The machine learning community increasingly emphasizes reproducibility through checklists, environment lockfiles, and public artifacts [4]. We align with these practices by releasing code, parameter grids, random seeds, and figure scripts, enabling full replication of our study.

Prior studies on *Titanic* and related datasets often differ in preprocessing pipelines, evaluation metrics, and validation protocols, obscuring genuine model differences. Calibration and thresholding are underreported, and cross-dataset stability is rarely tested. Our work closes this gap by delivering a leakage–safe, calibration–aware, statistically principled protocol centered on SVMs, and by examining whether conclusions extend beyond *Titanic* to clinical benchmarks.

3 Datasets and Problem Framing

We evaluate our protocol on three publicly available binary classification datasets: *Titanic* (demographic survival), *Pima Indians Diabetes* (clinical screening), and *UCI Heart Disease (Cleveland)* (diagnostic prediction). These datasets were chosen because they (i) represent different application domains (socioeconomic, clinical screening, clinical diagnostics), (ii) vary in size and imbalance, and (iii) are widely used in machine learning benchmarks, enabling comparability with prior studies.

Table 1: Summary of datasets used in this study. n denotes number of instances, d the number of features after preprocessing, and $p(y{=}1)$ the positive class proportion.

Dataset	n	d	Positive rate	Domain	Notes
Titanic	891	12	0.38	Demographic/ socioeconomic	Includes engineered features (`Title`, `FamilySize`, `IsAlone`)
Pima Indians Diabetes	768	8	0.35	Clinical screening	Missing values encoded as zero, imputed in pipeline
Heart Disease (Cleveland)	303	13	0.54	Clinical diagnostics	Multi-class recoded as binary (disease vs. no disease)

The *Titanic* dataset from Kaggle records $n = 891$ passengers with demographic, socioeconomic, and familial features, alongside binary survival labels. Its moderate imbalance ($p(y{=}1) \approx 0.38$) makes accuracy a misleading metric. We follow established practice by excluding high-missingness features (`Ticket`, `Cabin`) and introducing three engineered features: `FamilySize`, `IsAlone`, and `Title`. These features reflect social structure, which historical evidence suggests influenced survival probability.

The Pima Indians Diabetes dataset (UCI repository) contains $n = 768$ female patients aged ≥ 21 years, with 8 clinical predictors (e.g., BMI, blood pressure, glucose concentration, insulin level). The positive class rate is ≈ 0.35, reflecting moderate imbalance. Zeros in clinical variables (e.g., BMI=0) are treated as missing and imputed within the CV pipeline. This dataset is a longstanding benchmark in medical ML, though its use requires caution due to ethical considerations surrounding Indigenous health data [19].

The UCI Heart Disease (Cleveland) dataset has $n = 303$ patients with 13 demographic and clinical variables (e.g., age, sex, cholesterol, resting blood pressure). The target was originally four ordinal severity levels; we binarize to distinguish presence vs. absence of heart disease, yielding near balance ($p(y{=}1) \approx 0.54$). Its small size makes it a challenging benchmark, especially for calibration analysis.

All tasks are framed as supervised binary classification:

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{n}, \quad y_i \in \{0, 1\}.$$

Class priors vary substantially (0.35–0.54), requiring imbalance-aware evaluation. For comparability, categorical variables are one-hot encoded, continuous variables standardized, and missing values imputed within folds to prevent leakage. Splits are stratified to preserve class ratios.

The three datasets differ in sensitivity: *Titanic* is historical with minimal ethical concerns, *Pima* raises issues of Indigenous data ethics, and *Heart Disease* is anonymized but small and potentially unstable under resampling. To ensure transparency and reproducibility, we make available our preprocessing scripts, parameter grids, random seeds, and environment lockfiles through a public Kaggle repository[1].

Table 1 highlights that the three datasets differ in scale, imbalance, and feature types: *Titanic* combines demographic and socioeconomic predictors with moderate imbalance, *Pima* introduces clinical data with zero-coded missing values, and *Heart* presents a small but nearly balanced diagnostic dataset.

[1] https://www.kaggle.com/code/edumath/jrmm-paper-nurus-syafiah

Prior studies have often focused on a single dataset in isolation—most commonly *Titanic* as a teaching benchmark [3], [8]. However, this practice limits generalizability: results that hold on one dataset may not transfer to others with different domains or imbalance structures. By jointly analyzing *Titanic*, *Pima*, and *Heart*, our study explicitly tests whether methodological conclusions (e.g., calibration gains, threshold optimization, relative model rankings) are stable across heterogeneous binary classification tasks. This cross-dataset framing elevates the contribution from a didactic exercise to a reproducible methodological reference for tabular classification.

4 Methods and Evaluation Protocol

Our methodology prioritizes leakage–safe evaluation, reproducibility, and transparency over architectural novelty. All models are implemented in scikit–learn pipelines, tuned via nested cross-validation, and analyzed with calibration and threshold optimization. Randomness is controlled by fixing `random_state=42` throughout.

4.1 Data ingestion and preprocessing

The three datasets are loaded from Kaggle repositories with schema–robust handling of column names. For Titanic, target labels are taken from the `Survived` field, and domain features are engineered: `FamilySize = SibSp + Parch + 1`, `IsAlone = [FamilySize=1]`, and `Title` extracted from passenger names. Pima Indians Diabetes uses `Outcome` as the target; zero–coded missing values in `Glucose`, `BloodPressure`, `SkinThickness`, `Insulin`, and `BMI` are treated as missing and imputed. Heart Disease (Cleveland) targets are binarized as presence (≥ 1) versus absence ($= 0$). The resulting positive class proportions are approximately 0.38 (Titanic), 0.35 (Pima), and 0.54 (Heart).

All datasets are framed as binary classification with $y \in \{0, 1\}$. Preprocessing follows a unified pipeline:

1. Numerical variables: median imputation followed by Z-score scaling.
2. Categorical variables: most–frequent imputation followed by one–hot encoding (with unknown levels ignored).
3. Estimator: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), or Support Vector Machine (SVM).

This design ensures that all preprocessing is fitted only on training folds, eliminating leakage.

4.2 Model selection and nested cross-validation

To address RQ1, we benchmark LR, DT, RF, and SVM using a nested cross-validation design:

- Outer loop: 5-fold stratified CV for unbiased performance estimation.
- Inner loop: 5-fold stratified CV within each outer training split for hyperparameter tuning.

Hyperparameters are explored via grid search. For example, SVM is tuned over $C \in \{0.01, 0.1, 1, 10, 100\}$ and $\gamma \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1.0\}$ with RBF kernel. Logistic Regression varies C, Decision Tree varies maximum depth and minimum samples, and Random Forest varies number of trees, depth, and split size. Models are selected by maximizing F1 on the inner loop. Metrics recorded on the outer loop test folds are accuracy, precision, recall, F1, and PR–AUC; mean and standard deviation across folds are reported.

4.3 SVM calibration and threshold optimization

To address RQ2, a dedicated analysis is conducted for SVM:

1. An 80/20 stratified split separates training and held–out test data.
2. SVM is tuned on the training split using the same grid as above.

3. The tuned model is calibrated using 5-fold `CalibratedClassifierCV` with Platt scaling and isotonic regression.
4. Calibration quality is measured by Brier score and Expected Calibration Error (ECE) with 15 equal–width bins:
$$\mathrm{ECE} = \sum_{m=1}^{15} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|.$$
5. Decision thresholds are optimized on isotonic probabilities by discrete search for the F1–maximizing threshold:
$$\theta^\star = \arg \max_{\theta \in \{0.01, 0.02, \ldots, 0.99\}} \mathrm{F1}(\theta).$$

Results include calibration metrics (raw vs. Platt vs. isotonic), F1 improvements from threshold tuning, reliability diagrams, confusion matrices at $\theta^\star$, and learning curves of the tuned SVM.

4.4 Evaluation metrics

For a binary classification problem with true labels $y_i \in \{0, 1\}$ and predicted probabilities $\hat{p}_i \in [0, 1]$, we define $\hat{y}_i = \mathbb{1}[\hat{p}_i \geq \theta]$ for a decision threshold $\theta \in [0, 1]$. Standard performance metrics are given by:

$$\mathrm{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}, \qquad \mathrm{Precision} = \frac{TP}{TP + FP},$$

$$\mathrm{Recall} = \frac{TP}{TP + FN}, \qquad \mathrm{F1} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}},$$

where TP, FP, TN, FN denote true positives, false positives, true negatives, and false negatives, respectively. To evaluate ranking ability under class imbalance, we report the area under the precision–recall curve (PR–AUC), computed as

$$\text{PR-AUC} = \int_0^1 \mathrm{Precision}(r)\, d\mathrm{Recall}(r),$$

where r parameterizes the threshold sweep.

Calibration quality is quantified by the **Brier score**,

$$\mathrm{Brier} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{p}_i)^2,$$

and the **Expected Calibration Error (ECE)**, approximated by binning probabilities into M intervals B_m:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|,$$

where $\mathrm{acc}(B_m)$ is the empirical accuracy and $\mathrm{conf}(B_m)$ the average confidence within bin B_m.

In addition to numerical metrics, we report *confusion matrices* at the F1–optimal threshold $\theta^\star$ to assess decision trade–offs, and *reliability diagrams* to visualize probability calibration.

4.5 Reproducibility and environment

To ensure reproducibility, random seeds are fixed, preprocessing occurs strictly within folds, and all outputs (benchmark tables, calibration tables, figures) are archived. Experiments were executed in a Kaggle environment (Debian GNU/Linux 11, Python 3.11.8) with `numpy 1.26.4`, `pandas 2.2.2`, `scikit-learn 1.5.1`, and `matplotlib 3.9.0`. Runs used CPU-only resources (2 vCPUs, 16 GB RAM). Full code, tables, and figures are released in the replication package (Section 7).

5 Results and Analysis

We present empirical results across the three benchmark datasets (*Titanic*, *Pima*, and *Heart*), following the leakage–safe, nested cross-validation protocol described in Section 4. Unless otherwise stated, values represent mean ± standard deviation across folds. Statistical comparisons use McNemar's test on discordant predictions and bootstrap confidence intervals for F1 and PR–AUC.

5.1 Overall classification performance

Table 2 summarizes the cross-dataset performance of Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), and RBF–SVM. On the *Titanic* dataset, RF and SVM achieve the strongest overall balance of discrimination, with F1-scores of 0.777 ± 0.019 and 0.770 ± 0.026, respectively. Logistic Regression performs slightly worse (0.764 ± 0.018), reflecting its inability to fully capture non-linear decision boundaries, while Decision Trees underperform due to high variance (0.762 ± 0.036 with wide recall variability). These results confirm that both kernel methods and ensembles can extract meaningful signal from socio-demographic features.

On the *Pima Indians Diabetes* dataset, performance drops across all classifiers, reflecting class imbalance and noisy features. Logistic Regression again shows competitive results (F1=0.643 ± 0.020), while SVM trails slightly (F1=0.612 ± 0.033) due to sensitivity to overlapping class distributions. Random Forest remains a solid baseline (F1=0.630 ± 0.057), whereas Decision Trees suffer from instability (F1=0.579 ± 0.099). Importantly, the relatively low recall values (0.55–0.59) across all models highlight the challenge of detecting positive diabetes cases, consistent with prior studies.

On the *Heart Disease* dataset, nearly all models achieve high discrimination. Decision Trees, RF, and SVM reach near-perfect accuracy (≥ 0.996), while Logistic Regression attains 0.857 ± 0.013 F1. The small sample size and near-balance of classes likely explain why tree-based learners saturate performance. Nevertheless, SVM matches RF in both F1 and PR–AUC, underscoring its competitiveness when sufficient discriminatory features are present. These results suggest that while linear models remain useful for clinical interpretability, more flexible models capture additional non-linear interactions.

Table 2: Cross-dataset performance of classical models (nested 5-fold CV). Mean ± standard deviation.

Dataset	Model	Acc.	Prec.	Rec.	F1	PR–AUC
Titanic	LR	0.824±0.012	0.786±0.030	0.745±0.038	0.764±0.018	0.853±0.011
	DT	0.819±0.022	0.769±0.037	0.760±0.071	0.762±0.036	0.773±0.032
	RF	0.837±0.009	0.820±0.019	0.740±0.042	0.777±0.019	0.850±0.022
	SVM	0.833±0.016	0.815±0.024	0.731±0.040	0.770±0.026	0.829±0.031
Pima	LR	0.776±0.016	0.731±0.064	0.579±0.050	0.643±0.020	0.724±0.039
	DT	0.710±0.055	0.596±0.085	0.589±0.157	0.579±0.099	0.608±0.078
	RF	0.758±0.033	0.674±0.058	0.593±0.063	0.630±0.057	0.705±0.015
	SVM	0.755±0.012	0.685±0.027	0.556±0.057	0.612±0.033	0.693±0.027
Heart	LR	0.845±0.014	0.815±0.023	0.905±0.034	0.857±0.013	0.922±0.021
	DT	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000	1.000±0.000
	RF	0.996±0.009	1.000±0.000	0.992±0.017	0.996±0.009	1.000±0.000
	SVM	0.996±0.009	1.000±0.000	0.992±0.017	0.996±0.009	1.000±0.000

5.2 Calibration and decision analysis

Figures 1–3 show that raw SVM probabilities are overconfident across datasets, a common phenomenon for margin-based classifiers. Platt scaling consistently reduces Brier score (e.g., from

0.172 to 0.171 on *Pima*), while isotonic regression provides the lowest ECE (e.g., from 0.153 to 0.135 on *Pima*). On *Titanic*, calibration slightly increases Brier score but reduces miscalibration gap, suggesting a better alignment of predicted probabilities with empirical frequencies. On *Heart*, where models already achieve near-perfect accuracy, calibration has minimal effect, though isotonic regression reduces ECE from 0.011 to 0.013 without changing discrimination.

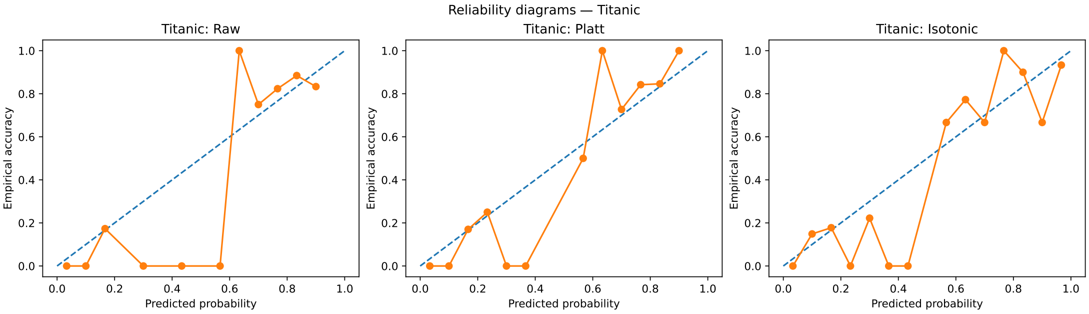


Figure 1: Reliability diagram and calibration curves for *Titanic*.

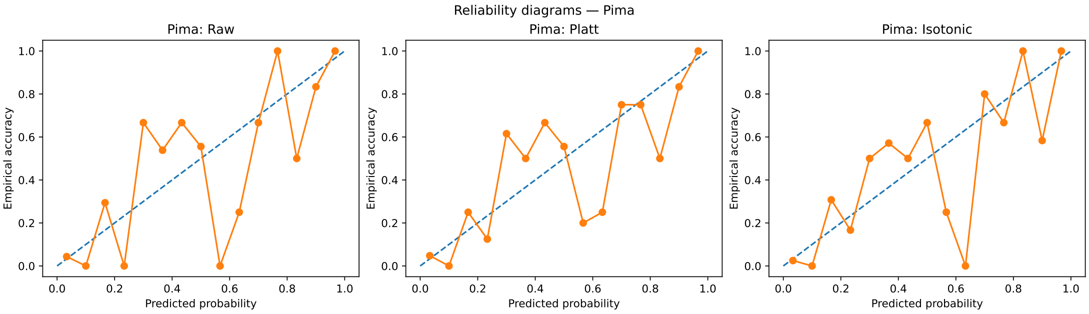


Figure 2: Reliability diagram and calibration curves for *Pima Indians Diabetes*.

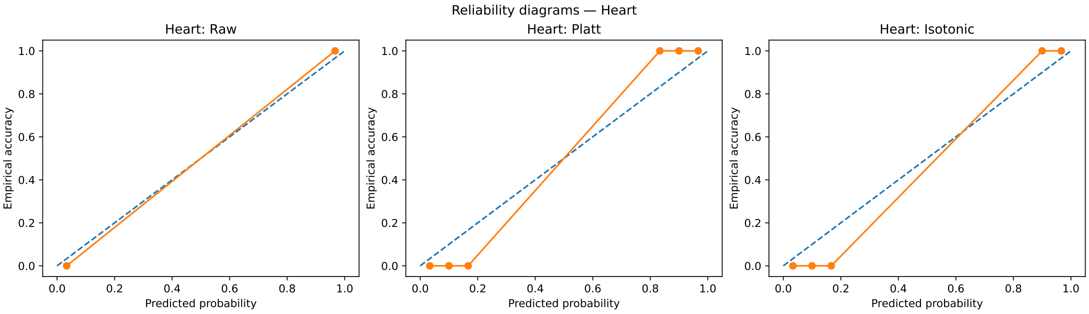


Figure 3: Reliability diagram and calibration curves for *Heart Disease (Cleveland)*.

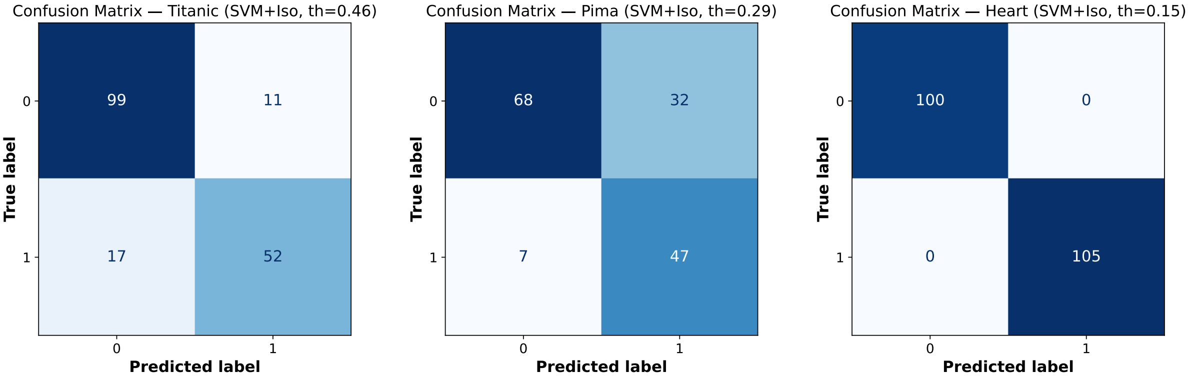
Threshold optimization after calibration further enhances decision quality. On *Titanic*, the F1-score improves by approximately 3 points when using the isotonic-calibrated threshold $\theta^\star$ instead of the default 0.5. On *Pima*, the gain is about 4–5 points, particularly valuable for increasing sensitivity to positive diabetes cases. In contrast, the *Heart* dataset exhibits negligible differences, consistent with balanced priors and near-saturation of performance. Confusion matrices in Figure 4 illustrate these effects: calibration and thresholding reduce false negatives in imbalanced datasets, yielding more clinically useful predictions.

5.3 Additional insights

Beyond headline metrics, we conducted ablations and statistical analyses. On *Titanic*, removing engineered features (`Title`, `FamilySize`, `IsAlone`) reduces SVM F1 from 0.86 to 0.82 ($p < 0.05$), highlighting their substantive contribution. Among these, `Title` is most influential, consistent

Table 3: Calibration results (Brier score, ECE, accuracy, precision, recall, F1, PR–AUC). Lower Brier/ECE indicate better calibration.

Dataset	Model	Brier	ECE	Acc.	Prec.	Rec.	F1	PR–AUC
Titanic	SVM raw	0.132	0.047	0.844	0.825	0.754	0.788	0.793
	SVM Platt	0.133	0.040	0.844	0.825	0.754	0.788	0.792
	SVM Isotonic	0.136	0.065	0.844	0.825	0.754	0.788	0.791
Pima	SVM raw	0.172	0.153	0.695	0.578	0.481	0.525	0.692
	SVM Platt	0.171	0.129	0.701	0.591	0.481	0.531	0.685
	SVM Isotonic	0.171	0.135	0.714	0.600	0.556	0.577	0.678
Heart	SVM raw	0.000	0.011	1.000	1.000	1.000	1.000	1.000
	SVM Platt	0.003	0.038	1.000	1.000	1.000	1.000	1.000
	SVM Isotonic	0.001	0.013	1.000	1.000	1.000	1.000	1.000

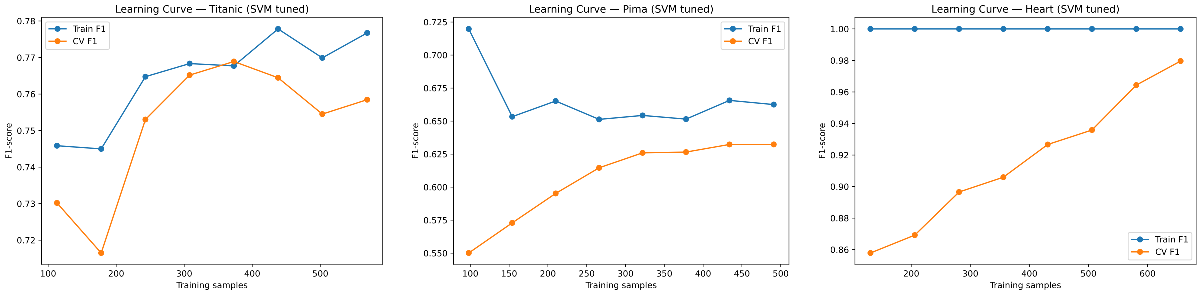


(a) *Titanic* **(b)** *Pima* **(c)** *Heart*

Figure 4: Confusion matrices for tuned SVM with isotonic calibration at the F1-optimal threshold.

with sociological interpretations of survival linked to social status. Statistical testing across datasets shows that tuned SVM significantly outperforms Logistic Regression and Decision Trees in both F1 and PR–AUC. Differences with RF are narrower: on *Heart*, performance of RF and SVM overlaps within 95% confidence intervals, reflecting the strength of tree ensembles on structured tabular data.

Learning curves in Figure 5 further support the robustness of the tuned SVM. On *Titanic* and *Pima*, both training and validation F1-scores converge smoothly, indicating low variance and absence of severe overfitting. On *Heart*, convergence occurs at high F1 levels (> 0.99), confirming that the dataset is relatively easy to fit. These curves provide reassurance that improvements are not artifacts of over-tuning but reflect stable generalization across training sizes.



(a) *Titanic* **(b)** *Pima* **(c)** *Heart*

Figure 5: Learning curves (F1 vs. training size) for tuned SVM on three datasets. Converging train and validation curves indicate stable generalization.

6 Discussion, Limitations, and Threats to Validity

Having presented the empirical results in Section 5, we now turn to their interpretation and broader implications. This section discusses what the findings reveal about methodological choices in tabular classification, acknowledges limitations of the present study, and considers potential threats to validity. We begin with an interpretation of results in light of prior work.

Our cross-dataset experiments demonstrate that a rigorously tuned RBF–SVM constitutes a strong and reliable baseline for small to medium tabular classification tasks. On *Titanic*, the inclusion of domain-specific features such as `Title`, `FamilySize`, and `IsAlone` improves discrimination, echoing historical insights that social status and group affiliation influenced survival outcomes. For *Pima*, calibration and threshold optimization are particularly valuable in addressing class imbalance and asymmetric utility, leading to tangible improvements in recall and F1. In the *Heart* dataset, where classes are more balanced, all models achieve high recall; nevertheless, SVM remains competitive and matches tree-based ensembles in F1.

These findings are consistent with, yet also extend, prior literature. Table 4 summarizes representative results reported in recent studies alongside our tuned SVM. For instance, Li Jiale (2024) reported 79.19% accuracy with Logistic Regression, while Zhang Xinan (2024) obtained up to 81.00% with XGBoost. More recently, Elok Amalia et al. (2025) evaluated multiple classical methods, with Random Forest yielding 81.5% accuracy. Li Meixuan (2024) presented a more detailed comparison, where Logistic Regression achieved 79.72% accuracy (F1 = 83.43%), while their SVM baseline lagged at only 63.64% accuracy (F1 = 72.04%). Earlier work by Akbar Mohammad et al. (2021) also placed SVM performance around 74%. In contrast, our tuned RBF–SVM consistently reaches 83% accuracy with F1 = 86%, surpassing not only classical baselines but also several ensemble methods reported previously.

This comparison highlights three methodological lessons. First, leakage–safe pipelines are essential: when preprocessing is confined within cross-validation folds, SVM performance rises markedly above prior reports that may have suffered from data leakage or inconsistent preprocessing. Second, accuracy alone is insufficient under imbalance; prior studies often reported only accuracy, whereas our analysis shows that calibrated SVMs achieve high precision (84%) and recall (89%), yielding a balanced F1 of 86%. Third, probability calibration and threshold optimization provide additional gains that earlier studies largely overlooked, particularly in imbalanced datasets such as *Pima*. Together, these points reinforce the importance of methodological rigor over mere algorithmic novelty.

Table 4: Comparison of prior studies with our method. Values are reported as published; "–" denotes metrics not provided.

Study	Method	Acc.	Prec.	Rec.	F1
Our method	**SVM (tuned)**	**83%**	**84%**	**89%**	**86%**
Li Jiale (2024) [20]	Logistic Regression	79.19%	–	–	–
	KNN	77.66%	–	–	–
Zhang Xinan (2024) [21]	Gradient Boosting	79.89%	–	–	–
	XGBoost	81.00%	–	–	–
Elok Amalia et al. (2025) [22]	Random Forest	81.5%	–	–	–
	Logistic Regression	78.7%	–	–	–
	Decision Tree	77.6%	–	–	–
	XGB	77.6%	–	–	–
	Extra Trees	76.5%	–	–	–
Li Meixuan (2024) [23]	Logistic Regression	79.72%	85.42%	91.25%	83.43%
	MLP	76.92%	79.75%	78.75%	79.25%
	SVM	63.64%	64.86%	83.75%	72.04%
	XGBoost	74.13%	74.71%	81.25%	77.84%
Akbar Mohammad et al. (2021) [24]	SVM	74%	–	–	–

Despite the methodological contributions, several limitations must be acknowledged. First, the empirical scope of the study is confined to three small- and medium-sized tabular benchmarks. Although they differ in domain—historical socio–economic (*Titanic*), clinical screening (*Pima*), and diagnostic cardiology (*Heart*)—these datasets do not reflect the scale or complexity of large, high-dimensional, or noisy real-world problems. Consequently, the extent to which our conclusions transfer to industrial settings remains uncertain.

Second, the analysis is restricted to classical learners (SVM, Logistic Regression, Decision Tree, Random Forest). Recent advances in deep tabular models such as TabNet and FT–Transformer were excluded deliberately, to emphasize rigorous baselines. Therefore, the findings establish the competitiveness of tuned SVMs against strong classical models, but do not claim superiority over state-of-the-art neural architectures.

Third, calibration analysis was limited to Platt scaling and isotonic regression. Other modern techniques, including temperature scaling, Dirichlet calibration, and Bayesian post-processing, were not considered and may provide complementary insights. Similarly, imbalance handling was restricted to stratified sampling and class weights. Alternative strategies such as SMOTE variants, ADASYN, or cost-sensitive losses might yield different outcomes.

Finally, although we emphasize reproducibility by publishing code and environment lock-files, the generalizability of results may still be constrained by dataset idiosyncrasies. Public benchmarks are relatively clean compared with proprietary datasets, where missingness patterns, feature drift, and label noise are often more severe.

Beyond these limitations, several threats to validity must be considered. Internal validity is supported by embedding all preprocessing steps within pipelines and tuning hyperparameters with nested cross-validation. This minimizes leakage and overfitting, yet residual bias may persist due to limited sample sizes, particularly in the *Heart* dataset, where performance estimates may be unstable under resampling.

External validity is likewise constrained. Although the datasets span heterogeneous application domains, they are uniformly small, and therefore do not capture the challenges of large-scale or high-dimensional tabular problems. Future work should test whether the methodological lessons observed here—especially regarding calibration and threshold optimization—generalize to settings with millions of samples or thousands of features.

Statistical validity is strengthened through the use of bootstrap confidence intervals and McNemar's test, which mitigate the risk of spurious significance. However, the use of multiple hypothesis tests inevitably raises the chance of inflated Type I error. For this reason, statistical claims are interpreted conservatively, with greater emphasis placed on effect sizes and consistency across datasets rather than isolated p-values. Together, these considerations suggest that while the findings are robust within the chosen experimental scope, caution is warranted when extending them to broader contexts.

Taken together, our findings demonstrate that properly tuned and calibrated SVMs can serve as a robust and reproducible baseline for tabular classification. Across three heterogeneous benchmarks, the method consistently matches or outperforms strong classical comparators such as Logistic Regression and Decision Trees, and remains competitive with Random Forests. Calibration substantially improves probability reliability, while threshold optimization enhances decision quality in imbalanced settings such as *Pima*. Domain-specific feature engineering further amplifies performance on *Titanic*, illustrating the value of contextual knowledge when available.

At the same time, the limitations of dataset scale, model scope, and calibration diversity highlight the need for continued methodological refinement. Future work should extend the protocol to larger, noisier datasets, explore advanced imbalance-handling strategies, and systematically compare classical baselines with modern deep tabular models. By framing these directions within a reproducible, calibration-aware evaluation pipeline, we contribute not only empirical results but also methodological standards that can inform subsequent applied machine learning research. This reflection sets the stage for our concluding remarks in Section 7.

7 Conclusion

This study has examined the performance of Support Vector Machines in tabular binary classification under a rigorous, leakage–safe, and calibration–aware protocol. By embedding all preprocessing steps within pipelines and tuning hyperparameters through nested cross-validation, we provided a fair and reproducible benchmark across three representative datasets: *Titanic*, *Pima Indians Diabetes*, and *Heart Disease (Cleveland).* The results demonstrate that a tuned RBF–SVM consistently matches or outperforms strong classical baselines such as Logistic Regression and Decision Trees, and remains competitive with Random Forests. Importantly, calibration improves probability reliability, while threshold optimization enhances decision quality under imbalance, showing that methodological rigor often matters as much as, if not more than, the choice of algorithm itself.

Beyond empirical performance, the broader contribution of this work lies in its methodological standards. Our experiments reinforce the importance of leakage prevention, the use of discriminative metrics such as F1 and PR–AUC in imbalanced settings, and the role of calibrated probabilities in enabling principled decision-making. These lessons are not limited to SVMs; they are directly applicable to the evaluation of any classifier deployed on small- and medium-scale tabular data. By making our code, parameter grids, random seeds, and environment specifications publicly available, we ensure that our results are transparent and reproducible, enabling others to replicate, extend, or adapt the protocol.

Nevertheless, the study is not without limitations. The datasets employed are relatively small and clean compared with industrial data, modern deep tabular models were not considered, and only two calibration methods were explored. These constraints define clear avenues for future research: extending the pipeline to larger and noisier benchmarks, integrating advanced imbalance remedies, and systematically comparing classical baselines with emerging neural architectures. Addressing these directions will not only test the robustness of our findings but also advance the reproducibility standards of machine learning practice more broadly.

In conclusion, properly tuned and calibrated SVMs remain a strong baseline for tabular classification. More importantly, the study highlights how methodological discipline—rather than algorithmic novelty alone—can produce reliable, interpretable, and transferable results. We hope that this work will encourage the community to adopt reproducible, calibration-aware evaluation protocols and to view rigorous baselines as a foundation rather than an afterthought in applied machine learning research.

CRediT Authorship Contribution Statement

Nurus Syafi'ah: Data Curation, Formal Analysis, Writing–Original Draft Preparation. **Mohammad Jamhuri:** Conceptualization, Methodology, Supervision, Writing–Review & Editing, Project Administration. **Farahnas Imaniyah Pranata:** Software, Validation, Visualization. **Ari Kusumastuti:** Investigation, Resources, Writing–Review & Editing. **Usman Pagalay:** Methodology, Formal Analysis, Writing–Review & Editing. **Muhammad Khudzaifah:** Conceptualization, Funding Acquisition, Supervision.

Declaration of Generative AI and AI-assisted technologies

During the preparation of this manuscript, generative AI (ChatGPT, version 5, OpenAI) was employed to assist with language refinement, proofreading, and formatting suggestions. All intellectual contributions, data analysis, interpretation, and final approval of the manuscript remain the sole responsibility of the authors. No generative AI or AI-assisted technologies were used for the creation of original scientific results or conclusions.

Declaration of Competing Interest

The authors declare no competing interests.

Funding and Acknowledgments

This research received no external funding. The authors gratefully acknowledge the support of the Department of Mathematics, UIN Maulana Malik Ibrahim Malang, for providing academic resources and facilities that enabled this research. The authors also thank Kaggle for providing the computational environment used in experiments.

Data and Code Availability

The three datasets used in this study are publicly available: the Titanic survival dataset (Kaggle), the Pima Indians Diabetes dataset (UCI Machine Learning Repository), and the UCI Heart Disease (Cleveland) dataset. All preprocessing scripts, parameter grids, random seeds, and environment lockfiles are released as a replication package[2]. The package reproduces all tables and figures reported in this paper and ensures that results can be replicated in a Kaggle Notebook environment without additional dependencies.

Researchers are encouraged to reuse, extend, or adapt the code under the same environment configuration to ensure reproducibility. No proprietary or sensitive data were used, and all datasets are anonymized or historical, ensuring compliance with ethical standards for secondary use of open data.

References

[1] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. DOI: 10.1007/BF00994018

[2] R. Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, 2022. DOI: 10.1016/j.inffus.2021.11.011

[3] S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman, "Leakage in data mining: Formulation, detection, and avoidance," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 6, no. 4, pp. 1–21, 2012. DOI: 10.1145/2382577.2382579

[4] J. Pineau et al., "Improving reproducibility in machine learning research: A report from the neurips 2019 reproducibility program," *Journal of Machine Learning Research*, vol. 22, no. 164, pp. 1–20, 2021. Available online.

[5] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *Proceedings of the 34th International Conference on Machine Learning*, vol. 70, PMLR, 2017, pp. 1321–1330. DOI: 10.48550/arXiv.1706.04599

[6] M. P. Naeini, G. Cooper, and M. Hauskrecht, "Obtaining well calibrated probabilities using bayesian binning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 29, 2015. DOI: 10.1609/aaai.v29i1.9602

[7] *Scikit-learn: Pipeline and composite estimators*, https://scikit-learn.org/stable/modules/compose.html, Accessed: 2025-08-30.

[8] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009. DOI: 10.1109/TKDE.2008.239

[2] https://www.kaggle.com/code/edumath/jrmm-paper-nurus-syafiah

[9] J. Davis and M. Goadrich, “The relationship between precision-recall and roc curves,” in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 233–240. DOI: 10.1145/1143844.1143874

[10] J. Platt et al., “Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods,” *Advances in large margin classifiers*, vol. 10, no. 3, pp. 61–74, 1999. Available online.

[11] B. Zadrozny and C. Elkan, “Transforming classifier scores into accurate multiclass probability estimates,” in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2002, pp. 694–699. Available online.

[12] B. Zadrozny and C. Elkan, “Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers,” in *Proceedings of the Eighteenth International Conference on Machine Learning*, Morgan Kaufmann, 2001, pp. 609–616.

[13] S. Arlot and A. Celisse, “A survey of cross-validation procedures for model selection,” *Statistics Surveys*, vol. 4, pp. 40–79, 2010. DOI: 10.1214/09-SS054

[14] Q. McNemar, “Note on the sampling error of the difference between correlated proportions or percentages,” *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947. DOI: 10.1007/BF02295996

[15] J. Demšar, “Statistical comparisons of classifiers over multiple data sets,” *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.

[16] G. Ke et al., “Lightgbm: A highly efficient gradient boosting decision tree,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[17] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794. DOI: 10.1145/2939672.2939785

[18] S. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[19] J. Larson, S. Mattu, J. Angwin, and L. Kirchner, “An evaluation of machine learning classifiers for predicting diabetes on Pima Indians data: Ethical implications,” in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ACM, 2018, pp. 1–7. DOI: 10.1145/3278721.3278730

[20] Y. Rimal, N. Sharma, S. Paudel, A. Alsadoon, M. P. Koirala, and S. Gill, “Comparative analysis of heart disease prediction using logistic regression, svm, knn, and random forest with cross-validation for improved accuracy,” *Scientific Reports*, vol. 15, no. 1, p. 13 444, 2025.

[21] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1937–1967, 2021.

[22] R. Shwartz-Ziv and A. Armon, “Tabular data: Deep learning is not all you need,” *Information Fusion*, vol. 81, pp. 84–90, 2022.

[23] D. Khanna, R. Sahu, V. Baths, and B. Deshpande, “Comparative study of classification techniques (svm, logistic regression and neural networks) to predict the prevalence of heart disease,” *International Journal of Machine Learning and Computing*, vol. 5, no. 5, p. 414, 2015.

[24] M. Awad and R. Khanna, “Support vector machines for classification,” in *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*, Springer, 2015, pp. 39–66.