bioengineering

MDPI

Article

Interpretable and Calibrated Classification of Clinical Data Using Supervised Feature Binarization

Antony Garcia [1,2], Adrián Noriega de la Colina [3,4,5,6], Gabrielle Britton [7] and Xinming Huang [1,*]

1 Department of Electrical and Computer Engineering, Worcester Polytechnic Institute, Worcester, MA 01609, USA; agarcia3@wpi.edu
2 Facultad de Ingeniería Eléctrica, Universidad Tecnológica de Panamá, Panama City 0819-07289, Panama; antony.garcia@utp.ac.pa
3 Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, USA; noriega1@mit.edu
4 Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA
5 Department of Neurology and Neurosurgery, McGill University, Montreal, QC H3A 0G4, Canada
6 Montreal Neurological Institute-Hospital, 3801 University Street, Montreal, QC H3A 2B4, Canada
7 Centro de Vacunación e Investigación (CEVAXIN), Endpoints Research Sites Network, Panama City 0830-00507, Panama; gabrielle.britton@cevaxin.com
* Correspondence: xhuang@wpi.edu

Abstract

Black-box models limit the adoption of artificial intelligence in medicine because their predictions are difficult to interpret and reproduce. We present a statistically grounded framework for interpretable, rule-based clinical classification using the Bernoulli Naïve Bayes (BNB) model. Supervised chi-square-guided binarization converts continuous variables into binary indicators by selecting thresholds that maximize association with the clinical outcome within the training folds, which allows BNB to operate on continuous medical data without sacrificing transparency. On three benchmark datasets, Pima Indians Diabetes, Wisconsin Breast Cancer, and Heart Failure Prediction, the framework reached areas under the receiver operating characteristic curve of 0.794, 0.984, and 0.918, respectively. Probabilistic reliability was assessed with a leakage-safe cross-validated calibration analysis reporting Brier score and calibration intercept and slope, and post hoc beta calibration improved probability calibration across datasets. These results indicate that an interpretable, statistically motivated framework can perform comparably to more complex models while providing explicit decision rules expressed in clinical units and risk estimates that are reliable after calibration. A complete worked example further shows that model inference can be reproduced from a printed reference table using only basic arithmetic, without software or proprietary tools, a property that may support auditable use of artificial intelligence in clinical settings.

Keywords: explainable AI; Bernoulli Naïve Bayes; feature binarization; chi-square thresholding; interpretable machine learning; medical classification; threshold optimization; probability calibration

Check for updates

Academic Editor: Andrea Cataldo

Received: 16 September 2026
Revised: 28 September 2026
Accepted: 29 September 2026
Published: 2 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Supervised classification is a fundamental task in medical data analysis. Its objective is the assignment of patients to clinically meaningful classes on the basis of their observed characteristics [1]. In practice, information such as demographics, medical history, and clin-

ical measurements is combined to determine whether a patient has a condition or is at risk of developing one [2].

These predictions are not merely technical outputs, because diagnosis, treatment planning, and patient outcomes can be directly influenced by them [3]. Several practical challenges are also presented by medical datasets: both continuous and categorical variables are often included, missing values may be present, and skewed class distributions are common [4,5]. For a classification model to be useful in clinical settings, these characteristics must be handled while maintaining sufficient interpretability for the resulting decisions to be trusted and understood by clinicians [6].

Artificial intelligence has demonstrated remarkable potential for improving diagnostic accuracy across a wide range of clinical applications [7–9]. However, many high-performing models, especially deep learning algorithms, are considered black boxes and may generalize poorly to diverse patient populations [10]. This lack of transparency can limit clinician trust and complicate regulatory adoption [11,12], making interpretable probabilistic models attractive for medical decision support.

Among supervised learning approaches, probabilistic classifiers are considered an important family because predictions are based on estimated probability distributions over class labels. This perspective is especially useful in clinical settings, where uncertainty is common. A notable example is the Bayesian network, in which conditional dependencies among features are represented and data analysis and decision-making under uncertainty are supported [13]. Within this broader family, conditional independence among features is assumed by the Naïve Bayes model, resulting in models that are easy to interpret and that often perform well even when the independence assumption is not always satisfied [14].

Several likelihood variants adapted to different types of data are included in the Naïve Bayes family: Bernoulli for binary features, Gaussian for continuous, normally distributed features, Multinomial for count-based data, and Categorical for discrete categorical features. High effectiveness has been reported for Bernoulli models in text classification tasks where the presence or absence of words is indicated by features [15–17], while Gaussian and Multinomial variants are generally preferred for continuous and frequency-based inputs, respectively, [18,19]. Categorical models handle discrete features directly without any assumption of ordering or distance between categories, making them well-adapted to tabular data with nominal features.

Despite its simplicity and interpretability, limited application in medical data analysis has been seen for the Bernoulli variant because of its reliance on binary features. Continuous variables, such as age, blood pressure, and glucose levels, are often contained in medical datasets and cannot be directly used with Bernoulli models without preprocessing. As a result, the use of Bernoulli Naïve Bayes models in clinical settings has been restricted, as preprocessing steps can introduce unnecessary complexity and introduce potential bias.

This limitation is addressed in the present work through the introduction of a statistical binarization method based on chi-square (χ^2) independence tests. Continuous features are transformed by the proposed method into binary indicators through the selection of thresholds at which the association with the outcome is least likely to occur by chance under the independence assumption.

During method development, several thresholding criteria were evaluated, including mutual information, information gain, Gini gain, ROC-based criteria, Otsu thresholding, and distribution-based cut points. Among these alternatives, the χ^2 criterion had the best mean AUC rank on the three datasets and was never significantly outperformed, while remaining closely aligned with the independence-testing rationale of the Bernoulli model. Threshold optimization was performed exclusively within the training folds, and final

performance was evaluated only on held-out data to avoid information leakage and to obtain an unbiased estimate of generalization.

By this strategy, Bernoulli models can be applied effectively to continuous medical data. The main contributions of this paper are as follows: (1) a clear formulation of the challenges involved in applying Bernoulli models to medical datasets with continuous features; (2) a leakage-safe framework that binarizes each continuous feature at a single χ^2-selected cut-point and models the result with Bernoulli or Categorical Naïve Bayes, using only empirical probabilities and no assumption about the distribution of each feature; (3) an empirical evaluation of the proposed method on three diverse and clinically relevant datasets; (4) binary features that correspond to explicit decision boundaries, thereby supporting transparent interpretation in clinical settings; and (5) a complete patient-level worked example demonstrating that the full classification pathway, from raw feature values to posterior probability, can be reproduced by hand using a printed reference table and a calculator, without requiring software access.

Alignment is also shown between this emphasis on transparent threshold-based rules and recent international frameworks, including the EU Artificial Intelligence Act [20] and the WHO guidance on the ethics and governance of artificial intelligence for health [21], in which explainability, fairness, and reproducibility are identified as key requirements for trustworthy clinical AI.

The remainder of this paper is organized as follows. Section 2 reviews prior work on Naïve Bayes models, supervised binarization, and interpretable clinical prediction. Section 3 describes the materials and methods, including the datasets, preprocessing procedures, Bernoulli and Categorical Naïve Bayes models, the proposed χ^2-based threshold-selection method, alternative thresholding strategies, and the evaluation protocol. Section 4 presents the comparative results, calibration analysis, and explainability analyses for the three medical datasets. Section 5 discusses the main implications and limitations of the study. Section 6 concludes the paper and outlines directions for future work.

2. Related Work

Naïve Bayes classifiers are one of the fundamental techniques of machine learning, often praised for their simplicity, efficiency, and interpretability [14,22]. Although Naïve Bayes assumes conditional independence between features, it has consistently achieved good performance in medical classification tasks, especially when paired with effective feature transformations.

In the medical field, Bernoulli Naïve Bayes has proven to be effective for tasks involving binary or thresholded features. BNB has demonstrated competitive performance in clinical classification tasks when preprocessing is appropriately designed [23–28]. Naïve Bayes has been successfully applied to predict lung cancer [25], detect cardiovascular diseases [26], and autism screening [27].

One of the main challenges in using BNB for clinical data lies in handling continuous biomarkers, such as glucose and cholesterol levels. Since BNB requires binary inputs, these continuous values must be converted into meaningful binarized features. In contrast, supervised methods use the target variable to identify thresholds that maximize predictive power. Unlike multi-interval discretization methods, the proposed approach focuses on identifying a single outcome-guided cutoff per feature, aligning with the binary structure required by Bernoulli Naïve Bayes [29–32]. The single-cut χ^2 criterion itself belongs to this established family of supervised discretization methods; the contribution of this work lies in combining it with Bernoulli and Categorical Naïve Bayes, which, unlike Gaussian Naïve Bayes, require no assumption about the distribution of each feature. This study demonstrates that statistically guided binarization provides a practical way to convert

continuous clinical measurements into interpretable decision rules. Prior work suggests that supervised discretization can enhance predictive discrimination while preserving model interpretability [33–36].

Recent advances have further applied feature binarization to high-dimensional and domain-specific medical datasets. A discretization guided by expert knowledge of heart disease data, for instance, has been shown to improve a Naïve Bayes classifier for heart disease diagnosis [37], while binarized gene expression data have been shown to preserve the separation between tumor types [38]. Hybrid methods combining discretization with evolutionary algorithms [39] or Bayesian scoring [40] have also reported success in biomedical domains.

From a clinical modeling perspective, threshold-based transformations yield clear, rule-like representations that can be readily interpreted and audited. Compared with black-box models, statistical binarization in BNB results in understandable, rule-based outputs that can, in principle, be justified and interpreted by clinicians themselves. This property aligns with commonly stated goals of interpretable machine learning, which emphasize simplicity, accountability, and dependability in clinical decision settings [41,42]. Work on explainability [43–45] has frequently discussed the trade-off between predictive performance and interpretability, and several authors argue that inherently interpretable models are preferable for clinical application.

3. Materials and Methods

3.1. Overview

This study presents a method for binarizing continuous features, making it possible to use the Bernoulli Naïve Bayes classifier with datasets that include continuous values. The approach uses the χ^2 test of independence as a scoring criterion: for each feature, multiple thresholds are assessed using χ^2 tests of independence to measure the association between the binarized feature and the target variable. The threshold with the lowest p-value from the χ^2 test is chosen, as it marks the point where the observed relationship is least consistent with independence.

The p-value is used only to rank candidate thresholds, not as a test of significance. Because all candidates of a feature share the same records and one degree of freedom, minimizing the p-value is equivalent to maximizing the χ^2 statistic. The minimum p-value obtained after scanning many candidates is optimistically small, as in any multiple-testing setting, and is therefore neither reported nor interpreted [46,47]. The optimism of the selection is instead handled by estimating thresholds within the training folds and evaluating performance on held-out data.

To evaluate the performance of the chi-square-based binarization method, other thresholding strategies were also implemented and compared, including mutual information, information gain, Gini gain, ROC-based criteria, Otsu thresholding, and distribution-based cut points.

The classification performance of the proposed method was compared with several reference models from the literature across three publicly available clinical datasets. Dataset characteristics and dataset-specific preprocessing protocols are detailed in Section 3.2.

3.2. Datasets and Dataset-Specific Preprocessing

Experiments were conducted on three publicly available medical datasets with binary targets: the Pima Indians Diabetes Database (PIDD), the Wisconsin Breast Cancer Dataset, and the Heart Failure Prediction Dataset. These datasets are widely used benchmarks in clinical machine learning and were selected to cover settings with continuous features, mixed feature types, and missing-value challenges.

- Pima Indians Diabetes Database (PIDD): contains 768 records with 8 clinical predictors and a binary diabetes outcome [48].
- Wisconsin Breast Cancer Dataset: contains 569 records with 30 predictors derived from digitized fine-needle aspirate images and a binary benign/malignant label [49].
- Heart Failure Prediction Dataset: contains 918 records with 11 predictors, including both numerical and categorical clinical variables, and a binary heart-disease outcome [50].

Summary statistics and class distributions for the three datasets are reported in Tables 1–3.

Table 1. Descriptive statistics for the features of the Pima Indian Diabetes Dataset. The summary includes the range (minimum and maximum values), mean, and standard deviation (Std Dev) for each risk factor across the entire dataset, as well as within two subpopulations: individuals with diabetes and individuals without diabetes.

Risk Factor	Unit	All Population (*N* = 768)			Diabetes Status				
					Diabetes (*N* = 268)			No Diabetes (*N* = 500)	
		Range	Mean	Std Dev	Range	Mean	Std Dev	Range	Mean	Std Dev
Age	Years	21.0–81.0	33.2	11.8	21.0–70.0	37.1	11.0	21.0–81.0	31.2	11.7
Pregnancies	Count	0.000–17.0	3.80	3.40	0.000–17.0	4.90	3.70	0.000–13.0	3.30	3.00
Glucose	mg/dL	0.000–199.0	120.9	32.0	0.000–199.0	141.3	31.9	0.000–197.0	110.0	26.1
Blood Pressure	mmHg	0.000–122.0	69.1	19.4	0.000–114.0	70.8	21.5	0.000–122.0	68.2	18.1
Skin Thickness	mm	0.000–99.0	20.5	16.0	0.000–99.0	22.2	17.7	0.000–60.0	19.7	14.9
Insulin	μU/mL	0.000–846.0	79.8	115.2	0.000–846.0	100.3	138.7	0.000–744.0	68.8	98.9
BMI	kg/m^2	0.000–67.1	32.0	7.90	0.000–67.1	35.1	7.30	0.000–57.3	30.3	7.70
Diabetes Pedigree Function		0.100–2.40	0.500	0.300	0.100–2.40	0.600	0.400	0.100–2.30	0.400	0.300

Table 2. Descriptive statistics for the features of the Wisconsin Breast Cancer Dataset. The summary includes the range (minimum and maximum values), mean, and standard deviation (Std Dev) for each feature across the entire dataset, as well as within two subpopulations: malignant and benign diagnoses.

Feature	Units	All Population (*N* = 569)			Diagnosis				
					Malignant (*N* = 212)			Benign (*N* = 357)	
		Range	Mean	Std Dev	Range	Mean	Std Dev	Range	Mean	Std Dev
Radius (Mean)	mm	6.98–28.1	14.1	3.52	10.9–28.1	17.5	3.20	6.98–17.9	12.1	1.78
Texture (Mean)	mm	9.71–39.3	19.3	4.30	10.4–39.3	21.6	3.78	9.71–33.8	17.9	4.00
Perimeter (Mean)	mm	43.8–188.5	92.0	24.3	71.9–188.5	115.4	21.9	43.8–114.6	78.1	11.8
Area (Mean)	mm^2	143.5–2501.0	654.9	351.9	361.6–2501.0	978.4	367.9	143.5–992.1	462.8	134.3
Smoothness (Mean)	ratio	0.050–0.160	0.100	0.010	0.070–0.140	0.100	0.010	0.050–0.160	0.090	0.010
Compactness (Mean)	ratio	0.020–0.350	0.100	0.050	0.050–0.350	0.150	0.050	0.020–0.220	0.080	0.030
Concavity (Mean)	ratio	0.000–0.430	0.090	0.080	0.020–0.430	0.160	0.080	0.000–0.410	0.050	0.040
Concave Points (Mean)	ratio	0.000–0.200	0.050	0.040	0.020–0.200	0.090	0.030	0.000–0.090	0.030	0.020
Symmetry (Mean)	ratio	0.110–0.300	0.180	0.030	0.130–0.300	0.190	0.030	0.110–0.270	0.170	0.020
Fractal Dimension (Mean)		0.050–0.100	0.060	0.010	0.050–0.100	0.060	0.010	0.050–0.100	0.060	0.010
Radius (SE)	mm	0.110–2.87	0.410	0.280	0.190–2.87	0.610	0.350	0.110–0.880	0.280	0.110
Texture (SE)	mm	0.360–4.88	1.22	0.550	0.360–3.57	1.21	0.480	0.360–4.88	1.22	0.590
Perimeter (SE)	mm	0.760–22.0	2.87	2.02	1.33–22.0	4.32	2.57	0.760–5.12	2.00	0.770
Area (SE)	mm^2	6.80–542.2	40.3	45.5	14.0–542.2	72.7	61.4	6.80–77.1	21.1	8.84
Smoothness (SE)	ratio	0.000–0.030	0.010	0.000	0.000–0.030	0.010	0.000	0.000–0.020	0.010	0.000
Compactness (SE)	ratio	0.000–0.140	0.030	0.020	0.010–0.140	0.030	0.020	0.000–0.110	0.020	0.020
Concavity (SE)	ratio	0.000–0.400	0.030	0.030	0.010–0.140	0.040	0.020	0.000–0.400	0.030	0.030
Concave Points (SE)	ratio	0.000–0.050	0.010	0.010	0.010–0.040	0.020	0.010	0.000–0.050	0.010	0.010
Symmetry (SE)	ratio	0.010–0.080	0.020	0.010	0.010–0.080	0.020	0.010	0.010–0.060	0.020	0.010
Fractal Dimension (SE)		0.000–0.030	0.000	0.000	0.000–0.010	0.000	0.000	0.000–0.030	0.000	0.000
Radius (Worst)	mm	7.93–36.0	16.3	4.83	12.8–36.0	21.1	4.28	7.93–19.8	13.4	1.98
Texture (Worst)	mm	12.0–49.5	25.7	6.15	16.7–49.5	29.3	5.43	12.0–41.8	23.5	5.49

Table 2. *Cont.*

Feature	Units	All Population (*N* = 569)			Diagnosis				
					Malignant (*N* = 212)			Benign (*N* = 357)	
		Range	Mean	Std Dev	Range	Mean	Std Dev	Range	Mean	Std Dev
Perimeter (Worst)	mm	50.4–251.2	107.3	33.6	85.1–251.2	141.4	29.5	50.4–127.1	87.0	13.5
Area (Worst)	mm^2	185.2–4254.0	880.6	569.4	508.1–4254.0	1422.3	598.0	185.2–1210.0	558.9	163.6
Smoothness (Worst)	ratio	0.070–0.220	0.130	0.020	0.090–0.220	0.140	0.020	0.070–0.200	0.120	0.020
Compactness (Worst)	ratio	0.030–1.06	0.250	0.160	0.050–1.06	0.370	0.170	0.030–0.580	0.180	0.090
Concavity (Worst)	ratio	0.000–1.25	0.270	0.210	0.020–1.17	0.450	0.180	0.000–1.25	0.170	0.140
Concave Points (Worst)	ratio	0.000–0.290	0.110	0.070	0.030–0.290	0.180	0.050	0.000–0.170	0.070	0.040
Symmetry (Worst)	ratio	0.160–0.660	0.290	0.060	0.160–0.660	0.320	0.070	0.160–0.420	0.270	0.040
Fractal Dimension (Worst)		0.060–0.210	0.080	0.020	0.060–0.210	0.090	0.020	0.060–0.150	0.080	0.010

Table 3. Descriptive statistics for the features of the Heart Failure Prediction Dataset. The summary includes the range (minimum and maximum values), mean, and standard deviation (Std Dev) for continuous features, as well as counts and percentages for categorical features, across the entire dataset and stratified by heart disease status. Cholesterol and resting blood pressure are recorded as 0 for 172 and 1 records, respectively; because neither value is physiologically possible, these entries are treated as missing and excluded from the corresponding statistics.

Risk Factor	Unit	Continuous Features								
		All Population (*N* = 918)			Heart Disease Status					
					Disease (*N* = 508)			No Disease (*N* = 410)		
		Range	Mean	Std Dev	Range	Mean	Std Dev	Range	Mean	Std Dev
Age	years	28.0–77.0	53.5	9.43	31.0–77.0	55.9	8.73	28.0–76.0	50.6	9.44
Resting Systolic Blood Pressure	mmHg	80.0–200.0	132.5	18.0	92.0–200.0	134.4	18.9	80.0–190.0	130.2	16.5
Cholesterol	mg/dL	85.0–603.0	244.6	59.2	100.0–603.0	251.1	62.5	85.0–564.0	238.8	55.4
Maximum Heart Rate	bpm	60.0–202.0	136.8	25.5	60.0–195.0	127.7	23.4	69.0–202.0	148.2	23.3
Oldpeak	mm	−2.60–6.20	0.890	1.07	−2.60–6.20	1.27	1.15	−1.10–4.20	0.410	0.700

	Unit	Categorical features		
		All Population	Disease	No Disease
Sex	count			
Male		725 (79.0%)	458 (90.2%)	267 (65.1%)
Female		193 (21.0%)	50 (9.8%)	143 (34.9%)
Chest Pain Type	count			
Atypical Angina		173 (18.8%)	24 (4.7%)	149 (36.3%)
Non-Anginal Pain		203 (22.1%)	72 (14.2%)	131 (32.0%)
Asymptomatic		496 (54.0%)	392 (77.2%)	104 (25.4%)
Typical Angina		46 (5.0%)	20 (3.9%)	26 (6.3%)
Fasting Blood Sugar	count			
No		704 (76.7%)	338 (66.5%)	366 (89.3%)
Yes		214 (23.3%)	170 (33.5%)	44 (10.7%)
Resting ECG	count			
Normal		552 (60.1%)	285 (56.1%)	267 (65.1%)
ST		178 (19.4%)	117 (23.0%)	61 (14.9%)
LVH		188 (20.5%)	106 (20.9%)	82 (20.0%)
Exercise Angina	count			
No		547 (59.6%)	192 (37.8%)	355 (86.6%)
Yes		371 (40.4%)	316 (62.2%)	55 (13.4%)
ST Slope	count			
Up		395 (43.0%)	78 (15.4%)	317 (77.3%)
Flat		460 (50.1%)	381 (75.0%)	79 (19.3%)
Down		63 (6.9%)	49 (9.6%)	14 (3.4%)

For PIDD, zeros in Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI were treated as missing values following prior studies. Two evaluation settings were considered for BNB: (i) mean-imputation preprocessing for direct comparability with published baselines, and (ii) a BNB likelihood that excludes missing features from the product term at prediction time. This second setting preserves incomplete records and was tested specifically for this dataset. The numbers of missing values were 5 (Glucose), 35 (Blood Pressure), 227 (Skin Thickness), 374 (Insulin), and 11 (BMI), affecting 376 of the 768 records; all predictors are numeric. The Wisconsin dataset has no missing values, and all 30 predictors are continuous.

For the Heart Failure dataset, the five continuous predictors (Age, Resting BP, Cholesterol, MaxHR, and Oldpeak) were binarized with the χ^2 thresholds, while the six categorical predictors (Sex, Chest Pain Type, Fasting Blood Sugar, Resting ECG, Exercise Angina, and ST Slope) were kept in their original categories (Table 4). All predictors were then modeled with Categorical Naïve Bayes (CNB), which is equivalent to BNB for the binarized features (Section 3.4). Cholesterol values recorded as 0 (172 records) were replaced by the median of the non-zero values, and records with resting BP below 80 mmHg or cholesterol outside 100–450 mg/dL were removed, leaving 908 records. In the comparisons with published studies (Section 4.1), the preprocessing of each cited study was followed instead.

Table 4. Encoding table for the Heart Failure Prediction Dataset. This table provides the categorical features, their respective categories, and the encoded values used in the dataset.

Column	Category	Encoded Value
Sex	F	0.000
	M	1.00
Chest Pain Type	ASY	0.000
	ATA	1.00
	NAP	2.00
	TA	3.00
Resting ECG	LVH	0.000
	Normal	1.00
	ST	2.00
Exercise Angina	N	0.000
	Y	1.00
ST Slope	Down	0.000
	Flat	1.00
	Up	2.00

3.3. Bernoulli Naïve Bayes Classifier

The BNB classifier is a probabilistic model specifically designed for feature vectors with binary values. It belongs to the broader family of Naïve Bayes classifiers, which rely on the assumption of conditional independence among features given the class label. Unlike Gaussian Naïve Bayes, which models continuous features using normal distributions, BNB models each binary feature using a Bernoulli distribution.

Let $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in \{0,1\}^m$ denote a binary feature vector, and let $\{C_1, \ldots, C_K\}$ represent the set of possible class labels. For each feature x_i, the class-conditional likelihood under class C_k is modeled as:

$$\mathbb{P}(x_i \mid C_k) = p_{k,i}^{x_i}(1 - p_{k,i})^{1-x_i}, \tag{1}$$

where $p_{k,i} = \mathbb{P}(x_i = 1 \mid C_k)$ denotes the probability that feature x_i is active (equals 1) given class C_k.

Assuming feature independence, the joint likelihood of the full feature vector $\mathbf{x}$ conditioned on class C_k becomes:

$$\mathbb{P}(\mathbf{x} \mid C_k) = \prod_{i=1}^{m} p_{k,i}^{x_i}(1 - p_{k,i})^{1-x_i}. \tag{2}$$

The posterior probability $\mathbb{P}(C_k \mid \mathbf{x})$ of class C_k given the observed feature vector $\mathbf{x}$ is obtained via Bayes' theorem. The predicted class label $\hat{y}$ is the one that maximizes this posterior:

$$\hat{y} = \arg\max_{k \in \{1,\ldots,K\}} \mathbb{P}(C_k) \prod_{i=1}^{m} p_{k,i}^{x_i}(1 - p_{k,i})^{1-x_i}. \tag{3}$$

The class priors and the class-conditional probabilities are estimated from the training data. If N is the number of training records, N_k the number of them in class C_k, and $n_{k,i}$ the number of class-C_k records with $x_i = 1$, the prior is $\mathbb{P}(C_k) = N_k/N$ and $p_{k,i}$ is estimated with Laplace smoothing:

$$\hat{p}_{k,i} = \frac{n_{k,i} + 1}{N_k + 2}. \tag{4}$$

Smoothing prevents a conditional probability of exactly zero when a binary state is not observed in one class. The same estimator is used for the Bernoulli models in all experiments.

Since BNB requires all features to be binary, applying it to continuous-valued datasets requires a binarization step. The method used for this transformation is detailed in the following subsection.

In this study, a modified version of the BNB model was employed in the Pima Indians Diabetes Database, which contains missing values in several features. To address this, the likelihood computation was adjusted to exclude missing values from the product term. Let $\mathcal{I}_{\text{obs}} \subseteq \{1, \ldots, m\}$ denote the indices of the observed (non-missing) features for a given instance. The modified likelihood is then computed as:

$$\mathbb{P}(\mathbf{x}_{\text{obs}} \mid C_k) = \prod_{i \in \mathcal{I}_{\text{obs}}} p_{k,i}^{x_i}(1 - p_{k,i})^{1-x_i}. \tag{5}$$

This adjustment effectively marginalizes over the missing components, preventing them from influencing the posterior estimation and thereby maintaining the integrity of the prediction. The results of this evaluation can be seen in Section 4.1.

3.4. Categorical Naïve Bayes Classifier

The Categorical Naïve Bayes classifier is another variant of the Naïve Bayes family, designed for discrete-valued features with more than two categories. While BNB is limited to binary features, CNB extends the likelihood model to handle features with multiple discrete values.

Let $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ be a feature vector where each feature $x_i \in \{v_1^{(i)}, v_2^{(i)}, \ldots, v_{r_i}^{(i)}\}$ takes values from a finite set of r_i discrete categories. Given a set of classes $\{C_1, \ldots, C_K\}$, the likelihood of a specific categorical value $x_i = v_j^{(i)}$ under class C_k is defined as:

$$\mathbb{P}(x_i = v_j^{(i)} \mid C_k) = \theta_{k,ij}, \tag{6}$$

where $\theta_{k,ij}$ denotes the class-conditional probability of observing value $v_j^{(i)}$ for feature x_i in class C_k, and satisfies the constraint:

$$\sum_{j=1}^{r_i} \theta_{k,ij} = 1 \quad \text{for all } i \text{ and } k. \tag{7}$$

Assuming conditional independence of features given the class label, the joint likelihood of the feature vector $\mathbf{x}$ is:

$$\mathbb{P}(\mathbf{x} \mid C_k) = \prod_{i=1}^{m} \theta_{k,ij} \quad \text{where } x_i = v_j^{(i)}. \tag{8}$$

The posterior probability $\mathbb{P}(C_k \mid \mathbf{x})$ is proportional to the product of the prior and the likelihood, as computed below:

$$\mathbb{P}(C_k \mid \mathbf{x}) \propto \mathbb{P}(C_k) \prod_{i=1}^{m} \theta_{k,ij}. \tag{9}$$

The predicted class label $\hat{y}$ is determined by:

$$\hat{y} = \arg \max_{k \in \{1,\ldots,K\}} \mathbb{P}(C_k) \prod_{i=1}^{m} \theta_{k,ij}. \tag{10}$$

If CNB is applied to binary features, it yields results equivalent to those of the BNB classifier, as both models reduce to the same probabilistic formulation under such conditions. This model was used to evaluate the Heart Failure Prediction Dataset, where categorical features were retained in their original form. Table 4 summarizes the encoding scheme applied to categorical features prior to CNB evaluation.

Equivalence of Categorical and Bernoulli Naïve Bayes for Binary Features

Assume that each feature is binary, $x_i \in \{0, 1\}$, with the two possible values defined as $v_1^{(i)} = 1$ ("active") and $v_2^{(i)} = 0$ ("inactive"). For Categorical Naïve Bayes the class-conditional likelihood of a feature vector $\mathbf{x}$ under class C_k is

$$\mathbb{P}(\mathbf{x} \mid C_k) = \prod_{i=1}^{m} \theta_{k,i1}^{\mathbf{I}[x_i=1]} \, \theta_{k,i2}^{\mathbf{I}[x_i=0]},$$

where the parameters satisfy $\theta_{k,i1} + \theta_{k,i2} = 1$. Setting

$$p_{k,i} = \theta_{k,i1}, \qquad 1 - p_{k,i} = \theta_{k,i2},$$

and noting that $\mathbf{I}[x_i = 1] = x_i$ and $\mathbf{I}[x_i = 0] = 1 - x_i$, the likelihood becomes

$$\mathbb{P}(\mathbf{x} \mid C_k) = \prod_{i=1}^{m} p_{k,i}^{x_i} (1 - p_{k,i})^{1-x_i},$$

which is precisely the Bernoulli Naïve Bayes formulation. Because both models share the same class priors and employ an identical maximum-a posteriori decision rule, they yield the same posterior probabilities and therefore identical predictions.

This result justifies treating Categorical NB as a seamless extension of Bernoulli NB to datasets that contain multi-level categorical variables while preserving interpretability for binary features.

3.5. Optimal Threshold Selection for Feature Binarization

Consider a dataset represented as:

$$\mathcal{D} = (x, y)$$

where x is a matrix of continuous features, and y is a binary target vector. The feature matrix is defined as:

$$x = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,m} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,m} \end{bmatrix} \in \mathbb{R}^{n \times m}$$

where n is the number of samples and m the number of features. The corresponding binary target vector is given by:

$$y = [y_1, y_2, \ldots, y_n], \quad y_i \in \{0, 1\}$$

Each feature $x_j \in \mathbb{R}^n$ is processed independently to identify its optimal binarization threshold.

To ensure meaningful comparisons, each feature vector x_j is first filtered to remove duplicate values. The resulting set of unique values defines the set of threshold candidates:

$$T_j = \{x_{i,j} \mid i \in \{1, 2, \ldots, n\}\}$$

For every threshold candidate $t \in T_j$, a binarized vector $x_{t,j} \in \{0, 1\}^n$ is computed as:

$$x_{i,j}^t = \begin{cases} 1, & x_{i,j} > t \\ 0, & x_{i,j} \leq t \end{cases} \quad \forall i \in \{1, \ldots, n\}$$

The relationship between the binarized feature vector $x_{t,j}$ and the binary target y is assessed via a chi-square test of independence. A contingency table is constructed for each threshold t:

$$C = \begin{bmatrix} n_{00} & n_{01} \\ n_{10} & n_{11} \end{bmatrix}$$

- n_{00}: number of samples with $x_{t,j} = 0$ and $y = 0$
- n_{01}: number of samples with $x_{t,j} = 0$ and $y = 1$
- n_{10}: number of samples with $x_{t,j} = 1$ and $y = 0$
- n_{11}: number of samples with $x_{t,j} = 1$ and $y = 1$

The chi-square statistic is computed as:

$$\chi^2 = \sum_{i=1}^{2} \sum_{j=1}^{2} \frac{\left(|O_{ij} - E_{ij}| - 0.5\right)^2}{E_{ij}}$$

where O_{ij} are the observed counts and E_{ij} the expected counts under the null hypothesis of independence (H_0), with Yates' continuity correction as implemented in SciPy. For a 2×2 table, the degrees of freedom are:

$$df = (2 - 1)(2 - 1) = 1$$

The p-value for threshold t, denoted p_t, is then computed as:

$$p_t = \mathbb{P}(\chi^2 \geq \chi^2_{\text{obs}} \mid H_0) = 1 - F_{\chi^2}(\chi^2_{\text{obs}}, 1)$$

where F_{χ^2} denotes the cumulative distribution function of the chi-square distribution with 1 degree of freedom, and χ^2_{obs} is the observed value of the chi-square statistic calculated from the contingency table for threshold t.

The optimal threshold is selected as the one that minimizes the p-value:

$$t_j^* = \arg\min_{t \in T_j} p_t$$

Using this optimal threshold, the final binarized feature vector is computed as:

$$x_{i,j}^{t_j^*} = \begin{cases} 1, & x_{i,j} > t_j^* \\ 0, & x_{i,j} \leq t_j^* \end{cases}$$

The process is repeated for each feature $j = 1, \ldots, m$, yielding an optimal threshold t_j^* and a corresponding binary feature vector. The final binarized feature matrix is:

$$x^* = \begin{bmatrix} x_{1,1}^{t_1^*} & x_{1,2}^{t_2^*} & \cdots & x_{1,m}^{t_m^*} \\ x_{2,1}^{t_1^*} & x_{2,2}^{t_2^*} & \cdots & x_{2,m}^{t_m^*} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1}^{t_1^*} & x_{n,2}^{t_2^*} & \cdots & x_{n,m}^{t_m^*} \end{bmatrix}$$

This matrix $x^* \in \{0,1\}^{n \times m}$ serves as input for training a Bernoulli Naïve Bayes classifier.

3.6. Other Thresholding Strategies

To evaluate the robustness of the proposed χ^2-guided thresholding, nine additional well-known binarization strategies were implemented and tested under the same cross-validation procedure. Methods were grouped into two families based on how the threshold is determined.

Score-based methods scan all unique feature values as threshold candidates and select the value that maximises a scoring function applied after binarization:

- Chi-square (χ^2): selects the threshold that minimises the p-value of the χ^2 test of independence between the binarized feature and the binary target.
- Mutual information: selects the threshold that maximises $I(X_t; Y)$, the mutual information between the thresholded feature and the target.
- Gini gain: selects the threshold that maximises the reduction in Gini impurity induced by the binary split.
- Information gain: selects the threshold that maximises the reduction in Shannon entropy (information gain) induced by the split.
- AUC split: selects the threshold that maximises $\max(\text{AUC}, 1 - \text{AUC})$, where AUC is computed using the thresholded feature as a univariate binary classifier.
- Youden's J: selects the threshold that maximises $J = \text{sensitivity} + \text{specificity} - 1$.

For a binary split of a binary outcome, the χ^2 statistic is proportional to the Gini gain ΔG:

$$\Delta G = \frac{2\pi(1-\pi)}{n} \chi^2, \tag{11}$$

where π is the outcome prevalence and n the number of records, so both criteria select the same threshold up to Yates' correction. The χ^2 criterion was preferred because the χ^2 test of independence is widely used and familiar in medical research [51,52], and it is computed directly from the 2×2 table of counts on which the model is based.

Direct estimators compute the threshold analytically from the feature distribution without scanning candidates:

- Otsu threshold: maximises the between-class variance of the feature histogram, computed via Otsu's method.
- Median threshold: sets the threshold at the feature median.
- Mean threshold: sets the threshold at the feature mean.
- k-means midpoint: fits a two-cluster k-means model ($k = 2$) and sets the threshold at the midpoint between the two cluster centroids.

For all methods, thresholding is applied as $x_{i,j}^t = \mathbf{I}[x_{i,j} > t]$, and thresholds are estimated exclusively on the training fold within each cross-validation iteration to prevent data leakage.

3.7. Evaluation Metrics

Classifier performance is evaluated using standard metrics, including accuracy, weighted precision, weighted recall, weighted F1-score, and ROC-AUC [53]. Special attention is given to the area under the receiver operating characteristic curve, which serves as an overall measure of binary classifier performance. ROC-AUC indicates the model's

ability to separate positive and negative classes by quantifying the total area under the ROC curve. As discussed in the literature [54], ROC-AUC is considered the most informative single metric for model evaluation in this study, as it does not depend on the classification threshold or class imbalance.

For each model, predictions from all held-out folds are concatenated to form pooled out-of-fold (OOF) vectors $(y_{\text{true}}, \hat{y}, \hat{p})$. Point estimates are then computed on these pooled OOF predictions. Alongside each point estimate, a 95% confidence interval (CI) is reported in the result tables using non-parametric bootstrap resampling with replacement ($B = 1000$ replicates). The CI bounds are the percentile limits:

$$\text{CI}_{95\%} = \left[q_{0.025}\left(\hat{\theta}^{*(1:B)}\right),\ q_{0.975}\left(\hat{\theta}^{*(1:B)}\right) \right],$$

where $\hat{\theta}$ denotes the metric of interest (accuracy, weighted precision, weighted recall, weighted F1-score, or ROC-AUC), and $\hat{\theta}^{*(b)}$ is the value from bootstrap replicate b.

To compare models statistically, paired tests are performed on matched OOF predictions obtained from the same samples. Each comparison is made against the proposed Bernoulli Naïve Bayes method. All hypothesis tests are two-sided with significance level $\alpha = 0.05$.

Differences in discrimination between BNB and each comparator were assessed using DeLong's test for correlated ROC-AUC values [55].

$$H_0 : \text{AUC}_A - \text{AUC}_B = 0$$

$$H_1 : \text{AUC}_A - \text{AUC}_B \neq 0.$$

Under H_0, both models have the same discrimination performance. The test statistic is computed as

$$z = \frac{\widehat{\text{AUC}}_A - \widehat{\text{AUC}}_B}{\text{SE}\left(\widehat{\text{AUC}}_A - \widehat{\text{AUC}}_B\right)},$$

with p-value $p = 2\Phi(-|z|)$, where Φ is the standard normal cumulative distribution function.

Differences in paired classification outcomes were evaluated using McNemar's test [56], based on discordant counts b and c computed from matched predictions of BNB and each comparator on the same cases.

$$H_0 : b = c$$

$$H_1 : b \neq c.$$

The null hypothesis states that both models have equal error rates on discordant pairs. When $b + c \geq 25$, the continuity-corrected chi-square form is used:

$$\chi^2 = \frac{(|b - c| - 1)^2}{b + c}.$$

When $b + c < 25$, the exact binomial version of McNemar's test is used.

Because many pairwise comparisons are made, all DeLong and McNemar p-values were adjusted for multiple comparisons with the Benjamini–Hochberg false discovery rate procedure [57], separately for each dataset and each test. An adjusted p-value below 0.05 was considered statistically significant, and only adjusted p-values are reported in Section 4.

Because relatively few published comparison studies report confidence intervals and the fold-level outputs required for paired statistical testing, additional self-implementations were developed for each comparator model. These implementations followed the model

and preprocessing descriptions in each paper, but all were evaluated with the protocol of this study (stratified 10-fold cross-validation with pooled out-of-fold predictions) rather than with the original validation scheme, which in several studies was a single train–test split. Each comparator was re-implemented to match or exceed its published performance, so that the comparison with BNB is fair. Many re-implemented values are close to the published ones; the remaining differences mainly reflect the change in validation scheme. All statistical tests compare BNB with the re-implementations on identical folds, and the published values are listed for reference only. This can be seen in Section 4.

3.8. Calibration Metrics

In clinical practice, a predicted probability is not merely a model output; it is a number that a clinician may use to decide whether to refer a patient, initiate treatment, or counsel a family. A model that assigns a 70% risk to a patient should, in practice, identify individuals who experience the outcome roughly seven times out of ten; if it does not, the probability estimate is misleading regardless of how well the model ranks patients relative to one another. In addition to ROC-AUC, calibration was evaluated to assess the reliability of predicted probabilities, while ROC-AUC measures discrimination, that is, the ability to rank cases correctly, it does not assess whether predicted probabilities correspond to observed event frequencies. For clinical decision support, this distinction is important because risk estimates may be used directly for interpretation and threshold-based decisions [58].

For this study, calibration analysis was incorporated into the same 10-fold cross-validation framework used for model evaluation. For each fold, χ^2-guided thresholds were estimated on the training split only, the corresponding Naïve Bayes model was trained on the transformed training data, and predicted probabilities were generated for the held-out fold. Pooled out-of-fold predictions were then used for calibration assessment, preserving separation between training and testing data and avoiding information leakage.

Let $y_i \in \{0, 1\}$ denote the observed outcome for sample i, and let $\hat{p}_i \in (0, 1)$ denote the predicted probability of the positive class. Using pooled OOF predictions, the following calibration metrics were computed.

3.8.1. Calibration-in-the-Large

Calibration-in-the-large evaluates whether predicted risks are too high or too low on average [59]. A simple descriptive summary compares the observed event rate

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

with the mean predicted risk

$$\bar{p} = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_i.$$

Good calibration-in-the-large is indicated when $\bar{p} \approx \bar{y}$.

A model-based estimate of calibration-in-the-large is the calibration intercept, obtained from logistic recalibration. An intercept of 0 is ideal.

3.8.2. Brier Score

The Brier score measures the mean squared error of probabilistic predictions [60]:

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

Lower values indicate better probabilistic accuracy, reflecting both discrimination and calibration.

3.8.3. Calibration Intercept and Slope

Calibration intercept and slope were estimated by fitting a logistic recalibration model to pooled OOF predictions [60]:

$$\text{logit}(\Pr(y_i = 1)) = \alpha + \beta \, \text{logit}(\hat{p}_i),$$

where

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right).$$

Here, α is the calibration intercept and β is the calibration slope. Ideal calibration corresponds to

$$\alpha = 0, \qquad \beta = 1.$$

An intercept $\alpha < 0$ suggests overall overprediction of risk, whereas $\alpha > 0$ suggests underprediction. A slope $\beta < 1$ is commonly associated with overconfident predictions, while $\beta > 1$ indicates underconfident predictions.

3.8.4. Post Hoc Beta Calibration

Because Naïve Bayes models can produce overconfident probabilities, especially when the conditional independence assumption is violated, an additional post hoc beta calibration step was applied [60]. Beta calibration is a parametric probability transformation that generalizes Platt scaling and is well suited to correcting asymmetric distortions in predicted probabilities [61].

For each cross-validation fold, beta calibration was fitted using training-fold predictions only and then applied to the corresponding held-out fold predictions. In one common parameterization, the calibrated probability $\tilde{p}$ is obtained from the raw probability $\hat{p}$ via

$$\text{logit}(\tilde{p}) = a \, \log(\hat{p}) + b \, \log(1 - \hat{p}) + c,$$

followed by

$$\tilde{p} = \frac{1}{1 + \exp[-(a \log(\hat{p}) + b \log(1 - \hat{p}) + c)]}.$$

This transformation was learned on the training fold and applied only to the held-out fold, producing calibrated OOF probabilities that were then evaluated using the same calibration metrics: calibration-in-the-large, Brier score, intercept, and slope.

3.9. Cross-Validation Strategy

To ensure a reliable evaluation of the Bernoulli Naïve Bayes classifier, stratified 10-fold cross-validation ($k = 10$) is used. In stratified cross-validation, each fold preserves the overall class distribution of the dataset, which is especially important when dealing with imbalanced classes. The dataset is split into ten equally sized folds, maintaining the same proportion of class labels in each. In every iteration, nine folds form the training partition and the remaining fold serves as a held-out validation fold, so that each fold is used once for performance evaluation.

To avoid data leakage, threshold optimization and model fitting are performed exclusively within the training partition in each iteration. Binarization thresholds are calculated only on the training folds and are then applied to both the training data and the held-out validation fold, ensuring that the validation data does not influence preprocessing or parameter estimation.

After binarization, the classifier is trained on the binarized training data and used to predict the corresponding held-out validation fold. Rather than computing each metric separately within every fold and averaging the ten resulting values, the predictions from

all held-out folds are concatenated into the pooled out-of-fold vectors defined in Section 3.7, and every reported metric is computed once on those pooled vectors. Pooling is required by the evaluation protocol: the paired DeLong and McNemar comparisons operate on matched per-sample predictions rather than on per-fold summaries, and the calibration analysis in Section 3.8 is performed on the same pooled predictions. Because each sample is held out exactly once, the pooled vectors contain one prediction per sample and therefore provide a single leakage-free estimate over the full dataset. For each metric, a 95% confidence interval is computed via non-parametric bootstrap resampling with replacement ($B = 1000$ replicates) on these pooled out-of-fold predictions.

To assess threshold stability, the procedure was repeated ten times with different random fold assignments, and the thresholds selected in each of the resulting 100 training partitions were recorded.

All reported results use this 10-fold procedure. The only exception is the illustrative worked example in Section 4.3, where the model was fitted on the remaining $N - 1 = 767$ records (leave-one-out). This keeps the patient out of model fitting, uses nearly all available data as in deployment, and avoids dependence on a random fold assignment.

3.10. Comparison Study Selection Criteria

Studies included for comparison in Section 4.1 were identified by searching for research that used the same datasets and reported ROC-AUC values, preferably estimated with k-fold cross-validation; because few studies met this criterion, some studies that used a single train–test split were also included. Preference was given to studies that described their preprocessing steps, allowing for fair and direct comparison. The selection was based on the relevance of the studies and the availability of results that could be matched to the evaluation criteria of the proposed method.

4. Results

This section reports the comparative performance of the proposed Bernoulli Naïve Bayes framework on the three benchmark clinical datasets defined in Section 3.2. Methodological details on preprocessing, thresholding, cross-validation, and statistical testing are provided in Section 3.

4.1. Proposed Model Compared to Reference Methods

The proposed χ^2-based binarization was compared with alternative thresholding strategies within the same Bernoulli Naïve Bayes framework: mutual information, Gini gain, information gain, AUC-split, Otsu, median, mean, Youden's J, and k-means midpoint.

Table 5 reports AUC (95% CI) and pairwise DeLong/McNemar comparisons versus χ^2. No alternative showed statistically significant improvement over χ^2 in any dataset.

For the Pima Indians Diabetes Database, no statistically significant difference from χ^2 was detected for any variant under either test. For the Wisconsin Breast Cancer dataset, χ^2 and Gini gain achieved the highest AUC (0.984), while mutual information, information gain, median and Youden's J underperformed in DeLong, and no McNemar difference was significant after correction. For the Heart Failure Prediction dataset, where only the five continuous predictors are thresholded, no variant differed significantly from χ^2 (AUC 0.916–0.918).

Although χ^2 was not always the single highest-performing method, it had the best mean AUC rank on the three datasets (2.67 among the ten criteria) and was never significantly outperformed. Some alternatives were equal or numerically better in individual settings, but those same methods showed significant underperformance in others (for example, mutual information and information gain in Breast Cancer). Gini gain ranked

second and never differed significantly from χ^2, as expected: it is proportional to the χ^2 statistic (Equation (11)), so both criteria selected the same threshold in almost all cases and differed only when Yates' correction broke a near-tie.

Table 5. Binarization-variant sensitivity analysis for Bernoulli Naïve Bayes across the three datasets (10-fold cross-validation). For each dataset, pairwise DeLong and McNemar *p*-values are computed against the χ^2 baseline; ** denotes significant underperformance. *p*-values are adjusted with the Benjamini–Hochberg procedure within each dataset and test.

Thresholding Method	Pima Indians Diabetes			Wisconsin Breast Cancer			Heart Failure		
	AUC [95% CI]	DeLong *p*	McNemar *p*	AUC [95% CI]	DeLong *p*	McNemar *p*	AUC [95% CI]	DeLong *p*	McNemar *p*
χ^2	0.794 [0.760, 0.826]	–	–	0.984 [0.974, 0.992]	–	–	0.918 [0.900, 0.935]	–	–
Mutual information	0.799 [0.767, 0.830]	0.5187	0.8131	0.981 [0.970, 0.990]	**0.0299** **	0.4985	0.918 [0.900, 0.935]	0.9062	1.0000
Gini gain	0.792 [0.759, 0.824]	0.5187	1.0000	0.984 [0.974, 0.992]	0.3698	1.0000	0.918 [0.900, 0.935]	1.0000	1.0000
Information gain	0.799 [0.767, 0.830]	0.5187	0.8131	0.981 [0.970, 0.990]	**0.0299** **	0.4985	0.918 [0.900, 0.935]	0.9062	1.0000
AUC split	0.790 [0.759, 0.823]	0.5187	0.8458	0.979 [0.968, 0.989]	0.0843	0.4985	0.917 [0.899, 0.935]	0.9062	1.0000
Otsu	0.780 [0.745, 0.810]	0.5187	0.8458	0.983 [0.973, 0.991]	0.6904	0.4985	0.918 [0.899, 0.934]	0.9062	1.0000
Median	0.787 [0.755, 0.821]	0.5187	0.8131	0.972 [0.958, 0.983]	**0.0299** **	0.2517	0.917 [0.898, 0.933]	0.9062	1.0000
Mean	0.781 [0.748, 0.812]	0.5187	0.8131	0.982 [0.971, 0.991]	0.4183	0.4985	0.916 [0.898, 0.933]	0.9062	1.0000
Youden's *J*	0.790 [0.759, 0.823]	0.5187	0.8458	0.978 [0.966, 0.989]	**0.0299** **	0.6089	0.918 [0.899, 0.935]	0.9062	1.0000
k-means midpoint	0.775 [0.740, 0.808]	0.5187	0.8131	0.972 [0.958, 0.985]	0.1035	0.4985	0.917 [0.899, 0.935]	0.9062	1.0000

4.1.1. Pima Indians Diabetes Database

Table 6 reports the Pima Indians Diabetes Database comparisons for the preprocessing settings described in Section 3.2. In these blocks, the self-implementations are generally aligned with the reported studies and frequently reproduce equal or higher values under the corresponding preprocessing protocols.

The statistical pattern in Table 6 is not one of uniform superiority over BNB. For most classical baselines no statistically significant difference from BNB was detected, with non-significant DeLong/McNemar outcomes in many rows. After correction, significant DeLong advantages over BNB were limited to the neural networks, random forest and Gaussian Naïve Bayes of one block, Gaussian Naïve Bayes and *k*-nearest neighbors of another, and three SVM formulations, while significant underperformance appeared in tree-based baselines and two RBF-based SVM variants. No McNemar comparison remained significant.

Table 6. Comparison of reported and re-implemented algorithms for the Pima Indians Diabetes dataset under the preprocessing strategies used in the cited studies. Reported study values are shown alongside the corresponding implementations and their computed metrics. DeLong and McNemar *p*-values are calculated against the BNB benchmark row within each preprocessing block; *p*-values are adjusted with the Benjamini–Hochberg procedure within the table and test; bold *p*-values indicate statistical significance ($p < 0.05$), with * denoting significant outperformance over BNB and ** denoting significant underperformance. Only AUC is shown here; the full comparison including accuracy, recall, precision, and F1-score is given in Table S1 of the Supplementary Materials.

Reference	Preprocessing	Algorithm	AUC		Statistical Tests	
			Reported	Re-Implemented [95% CI]	DeLong	McNemar
[62]	Mean imputation	Gaussian Naïve Bayes	0.846	0.820 [0.790, 0.849]	**0.022** *	0.599
		k-Nearest Neighbors	0.838	0.779 [0.744, 0.811]	0.400	0.600
		J48 Decision Tree	0.785	0.707 [0.670, 0.740]	**<0.001** **	0.505
		Random Forest	0.836	0.825 [0.795, 0.854]	**0.010** *	0.773
		Feedforward Neural Network	0.851	0.835 [0.806, 0.862]	**<0.001** *	0.505
		Convolutional Neural Network	0.837	0.835 [0.806, 0.863]	**<0.001** *	0.196
		Bernoulli Naïve Bayes	–	0.784 [0.751, 0.815]	0.116	0.245
	–	Bernoulli Naïve Bayes (NaN-aware)	–	0.793 [0.759, 0.823]	–	–

Table 6. *Cont.*

Reference	Preprocessing	Algorithm	AUC		Statistical Tests	
			Reported	Re-Implemented [95% CI]	DeLong	McNemar
[63]	Removal of missing values	Gaussian Naïve Bayes	0.810	0.831 [0.790, 0.872]	0.792	0.600
		J48 Decision Tree	0.750	0.801 [0.750, 0.849]	0.302	1.000
		Random Forest	0.810	0.855 [0.813, 0.896]	0.127	0.924
		Logistic Regression	0.830	0.849 [0.808, 0.890]	0.246	0.924
		Bernoulli Naïve Bayes	–	0.827 [0.783, 0.869]	–	–
[64]	Removal of missing values	Classification Tree	0.648	0.699 [0.649, 0.748]	**<0.001** **	0.176
		Support Vector Machine	0.707	0.836 [0.790, 0.879]	0.652	0.505
		k-Nearest Neighbors	0.737	0.797 [0.750, 0.845]	0.237	0.599
		Gaussian Naïve Bayes	0.818	0.831 [0.790, 0.872]	0.792	0.600
		Random Forest	0.808	0.851 [0.810, 0.892]	0.158	0.924
		Multilayer Perceptron	0.824	0.804 [0.752, 0.851]	0.314	0.505
		AdaBoost	0.681	0.843 [0.798, 0.886]	0.302	0.924
		Logistic Regression	0.825	0.849 [0.808, 0.889]	0.240	0.924
		Bernoulli Naïve Bayes	–	0.827 [0.783, 0.869]	–	–
[65]	Mean imputation, feature selection	Gaussian Naïve Bayes	0.800	0.829 [0.799, 0.857]	**0.002** *	0.196
		Decision Tree	0.710	0.759 [0.725, 0.790]	**0.047** **	0.599
		Random Forest	0.830	0.812 [0.779, 0.842]	0.208	0.388
		k-Nearest Neighbors	0.800	0.836 [0.806, 0.863]	**0.002** *	0.100
		Bernoulli Naïve Bayes	–	0.788 [0.755, 0.818]	0.252	0.505
	–	Bernoulli Naïve Bayes (NaN-aware)	–	0.793 [0.758, 0.823]	–	–
[66]	SMOTE, Mean imputation and Outlier Removal	SVM (Linear)	0.747	0.824 [0.796, 0.853]	**0.019** *	0.741
		SVM (Polynomial)	0.726	0.774 [0.736, 0.810]	0.357	0.599
		SVM (RBF Minkowski)	0.836	0.735 [0.695, 0.770]	**0.004** **	0.388
		SVM (RBF City Block)	0.841	0.733 [0.695, 0.767]	**0.004** **	0.294
		SVM (RBF Mahalanobis)	0.817	0.788 [0.752, 0.821]	0.761	0.924
		SVM (RBF Bray-Curtis)	0.744	0.814 [0.784, 0.842]	0.158	0.505
		SVM (RBF Canberra)	0.789	0.834 [0.805, 0.861]	**<0.001** *	0.505
		SVM (RBF Euclidean Distance)	0.809	0.826 [0.796, 0.856]	**0.019** *	0.864
		Bernoulli Naïve Bayes	–	0.780 [0.748, 0.811]	0.252	0.176
	–	Bernoulli Naïve Bayes (NaN-aware)	–	0.794 [0.760, 0.826]	–	–

For this dataset, the NaN-aware BNB implementation handles missing values internally by omitting likelihood contributions for missing entries rather than applying imputation. In most comparisons with standard BNB using χ^2-based binarization, no statistically significant difference was detected, although NaN-aware BNB showed slightly higher values in most settings.

4.1.2. Wisconsin Breast Cancer Dataset

Table 7 reports the Wisconsin Breast Cancer dataset comparisons without missing-value preprocessing, including BNB and the reference models from [67–70]. The implemented results in this table also show broad reproducibility, with many rows matching or surpassing the corresponding paper-level values.

Compared with the Pima Indians Diabetes Database and Heart Failure Prediction tables, Table 7 contains a larger number of statistically significant outperformance cases versus BNB. The models that outperform BNB are mainly ensemble methods and neural/deep learning models, plus selected SVM configurations with non-linear kernels and logistic regression in one block, whereas several classical comparators remain non-significant and some tree-based models are significantly below BNB. These alternatives generally improve predictive performance at the cost of interpretability, because their complex and non-linear decision structures are less transparent for direct clinical reasoning.

Table 7. Comparison of reported and re-implemented algorithms for the Wisconsin Breast Cancer dataset. Reported study values are shown alongside the corresponding implementations and their computed metrics. DeLong and McNemar *p*-values are calculated against the Bernoulli Naïve Bayes (BNB) benchmark within each study block; *p*-values are adjusted with the Benjamini–Hochberg procedure within the table and test; bold *p*-values indicate statistical significance ($p < 0.05$), with * denoting significant outperformance over BNB and ** denoting significant underperformance. Only AUC is shown here; the full comparison including accuracy, recall, precision, and F1-score is given in Table S2 of the Supplementary Materials.

Reference	Algorithm	AUC		Statistical Tests	
		Reported	Re-Implemented [95% CI]	DeLong	McNemar
[71]	Multi-layer Perceptron (MLP)	0.993	0.990 [0.978, 0.997]	0.277	0.063
	Radial Basis Function Network (RBF)	0.990	0.989 [0.980, 0.995]	0.159	1.000
	Probabilistic Neural Network (PNN)	0.994	0.993 [0.986, 0.998]	**0.029** *	0.214
	Bernoulli Naïve Bayes	–	0.984 [0.974, 0.992]	–	–
[68]	SELF Stacking Ensemble	0.991	0.992 [0.984, 0.998]	**0.042** *	**0.003** *
	Extra Trees	0.994	0.993 [0.986, 0.998]	**0.016** *	**0.032** *
	Random Forest	0.989	0.988 [0.978, 0.996]	0.104	0.634
	AdaBoost	0.991	0.994 [0.989, 0.998]	**0.016** *	**0.003** *
	Gradient Boosting	0.995	0.991 [0.983, 0.996]	**0.042** *	0.167
	k-Nearest Neighbor (k = 9)	0.979	0.988 [0.978, 0.997]	0.301	**0.035** *
	Support Vector Machine (RBF)	0.963	0.994 [0.987, 0.999]	**0.029** *	**0.002** *
	Multilayer Perceptron (MLP)	0.967	0.991 [0.982, 0.998]	0.177	**0.026** *
	Classification and Regression Tree (CART)	0.963	0.907 [0.882, 0.932]	**<0.001** **	**0.026** **
	Stochastic Gradient Descent (Log-loss)	0.951	0.995 [0.989, 0.999]	**0.016** *	**0.012** *
	Bernoulli Naïve Bayes	–	0.984 [0.974, 0.992]	–	–
[67]	Support Vector Machine (SVM)	0.966	0.994 [0.987, 0.999]	**0.029** *	**0.002** *
	Random Forest	0.960	0.989 [0.978, 0.997]	0.104	0.167
	Logistic Regression	0.947	0.992 [0.985, 0.998]	**0.042** *	**0.039** *
	Decision Tree (C4.5 approximation)	0.945	0.932 [0.909, 0.954]	**<0.001** **	0.544
	k-Nearest Neighbors (KNN)	0.952	0.986 [0.975, 0.996]	0.612	**0.035** *
	Bernoulli Naïve Bayes	–	0.984 [0.974, 0.992]	–	–
[70]	Artificial Neural Network (ANN)	0.990	0.990 [0.982, 0.996]	0.131	**0.035** *
	Support Vector Machine (SVM, RBF)	0.970	0.995 [0.990, 0.999]	**0.016** *	**0.002** *
	Random Forest	0.960	0.990 [0.980, 0.997]	0.051	0.252
	Bernoulli Naïve Bayes	–	0.984 [0.974, 0.992]	–	–
[69]	Decision Tree (Entropy)	0.990	0.970 [0.955, 0.984]	0.057	0.540
	Gaussian Naïve Bayes	0.890	0.984 [0.975, 0.992]	0.878	0.214
	Random Forest	0.990	0.989 [0.978, 0.997]	0.081	0.123
	Bernoulli Naïve Bayes	–	0.984 [0.974, 0.992]	–	–

4.1.3. Heart Failure Prediction Dataset

Table 8 reports Heart Failure Prediction dataset comparisons for BNB and the reference classifiers, including Gaussian Naïve Bayes [72]. As in the other datasets, the independent implementations are generally consistent with the reported studies and in many rows achieve equal or higher computed metrics.

The statistical outcomes in Table 8 are predominantly non-significant for many classical baselines, so no difference from BNB was detected for these comparators. After correction, significant outperformance was limited to the random forest in two blocks (McNemar test only), whereas the decision-tree baselines were significantly worse than BNB. This table reinforces the role of BNB as a competitive reference model and suggests that apparent SVM gains depending on non-linear kernel choices should be interpreted carefully.

Table 8. Comparison of reported and re-implemented algorithms for the Heart Failure Prediction dataset. Reported study values are shown alongside the corresponding implementations and their computed metrics. DeLong and McNemar *p*-values are calculated against the Bernoulli Naïve Bayes (BNB) benchmark within each study block; *p*-values are adjusted with the Benjamini–Hochberg procedure within the table and test; bold *p*-values indicate statistical significance ($p < 0.05$), with * denoting significant outperformance over BNB and ** denoting significant underperformance. Only AUC is shown here; the full comparison including accuracy, recall, precision, and F1-score is given in Table S3 of the Supplementary Materials. In the BNB rows, the categorical predictors are modeled with Categorical Naïve Bayes (Section 3.2).

Reference	Algorithm	AUC		Statistical Tests	
		Reported	Re-Implemented [95% CI]	DeLong	McNemar
[73]	SVM	0.826	0.921 [0.902, 0.940]	0.868	0.144
	Random Forest	0.838	0.929 [0.911, 0.947]	0.157	**0.009** *
	Naïve Bayes	0.853	0.911 [0.891, 0.930]	0.170	0.685
	Logistic Regression	0.832	0.911 [0.889, 0.930]	0.170	0.744
	KNN	0.827	0.910 [0.887, 0.929]	0.225	0.335
	Decision Tree	0.805	0.798 [0.771, 0.823]	**<0.001** **	**0.009** **
	Bernoulli Naïve Bayes	–	0.920 [0.900, 0.938]	–	–
[74]	Logistic Regression (LR)	0.940	0.911 [0.891, 0.930]	0.157	0.535
	Random Forest (RF)	0.940	0.928 [0.910, 0.945]	0.170	0.062
	Extra Trees (ET)	0.940	0.926 [0.908, 0.944]	0.316	0.203
	GBDT	0.940	0.922 [0.903, 0.940]	0.868	0.744
	Multilayer Perceptron (MLP)	0.920	0.923 [0.905, 0.940]	0.738	0.956
	Stacking Ensemble	0.950	0.929 [0.911, 0.946]	0.157	0.061
	Bernoulli Naïve Bayes	–	0.921 [0.902, 0.938]	–	–
[72]	Logistic Regression	0.869	0.912 [0.891, 0.930]	0.291	1.000
	Decision Tree	0.825	0.799 [0.773, 0.824]	**<0.001** **	**0.015** **
	Random Forest	0.866	0.927 [0.908, 0.944]	0.157	0.202
	Support Vector Machine (RBF)	0.879	0.916 [0.894, 0.935]	0.752	0.203
	k-Nearest Neighbors	0.903	0.909 [0.887, 0.927]	0.171	1.000
	Gaussian Naïve Bayes	0.863	0.909 [0.889, 0.928]	0.170	1.000
	Bernoulli Naïve Bayes	–	0.918 [0.900, 0.935]	–	–
[75]	Random Forest	0.990	0.930 [0.911, 0.947]	0.157	**0.009** *
	Logistic Regression	0.910	0.910 [0.889, 0.929]	0.157	0.744
	k-Nearest Neighbours (K = 13)	0.840	0.920 [0.901, 0.939]	0.908	0.640
	Bernoulli Naïve Bayes	–	0.920 [0.900, 0.938]	–	–

4.2. Calibration Analysis

In addition to ROC-AUC, calibration analysis was performed following the protocol described in Section 3. Calibration results are summarized in Table 9 and the corresponding calibration curves are shown in Figure 1. Four metrics are reported for both raw and beta-calibrated probabilities for each dataset: calibration-in-the-large, Brier score, calibration intercept, and calibration slope.

Raw BNB probabilities were miscalibrated, with calibration slopes of 0.530, 0.162, and 0.505 for the Pima, Wisconsin, and Heart Failure datasets, indicating overconfident estimates. Discrimination and calibration therefore diverged: the Wisconsin model had the highest AUC (0.984) but the most overconfident raw probabilities. Beta calibration improved the slope substantially in all three datasets, with the Heart Failure dataset reaching a slope of 0.982 and an intercept of 0.006, close to ideal calibration. Brier scores also decreased after calibration in all cases. The Wisconsin Breast Cancer dataset had the lowest Brier score (0.044 post-calibration), consistent with its high discrimination (AUC 0.984).

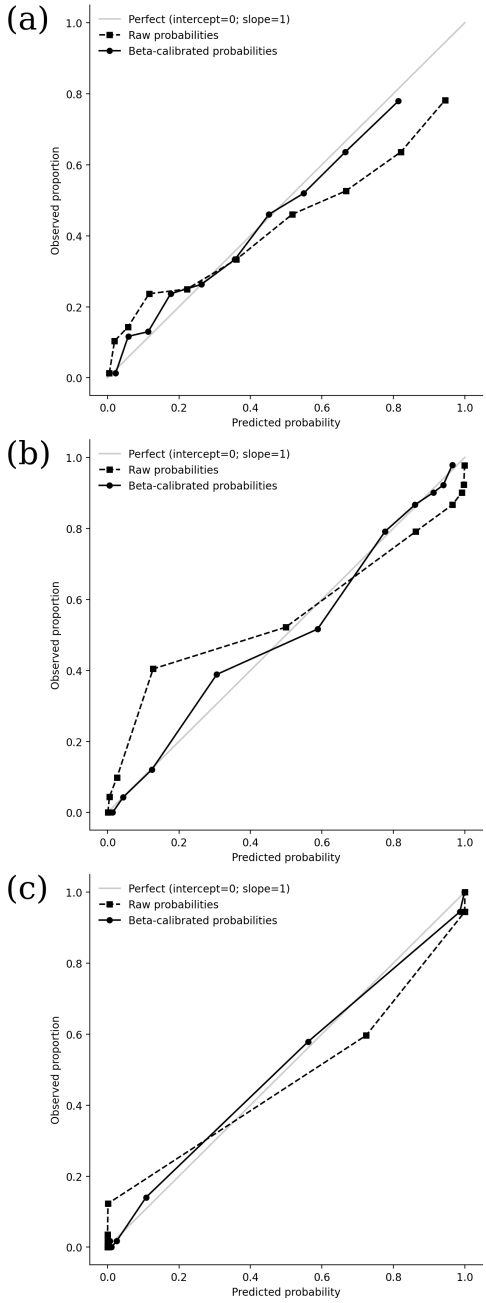


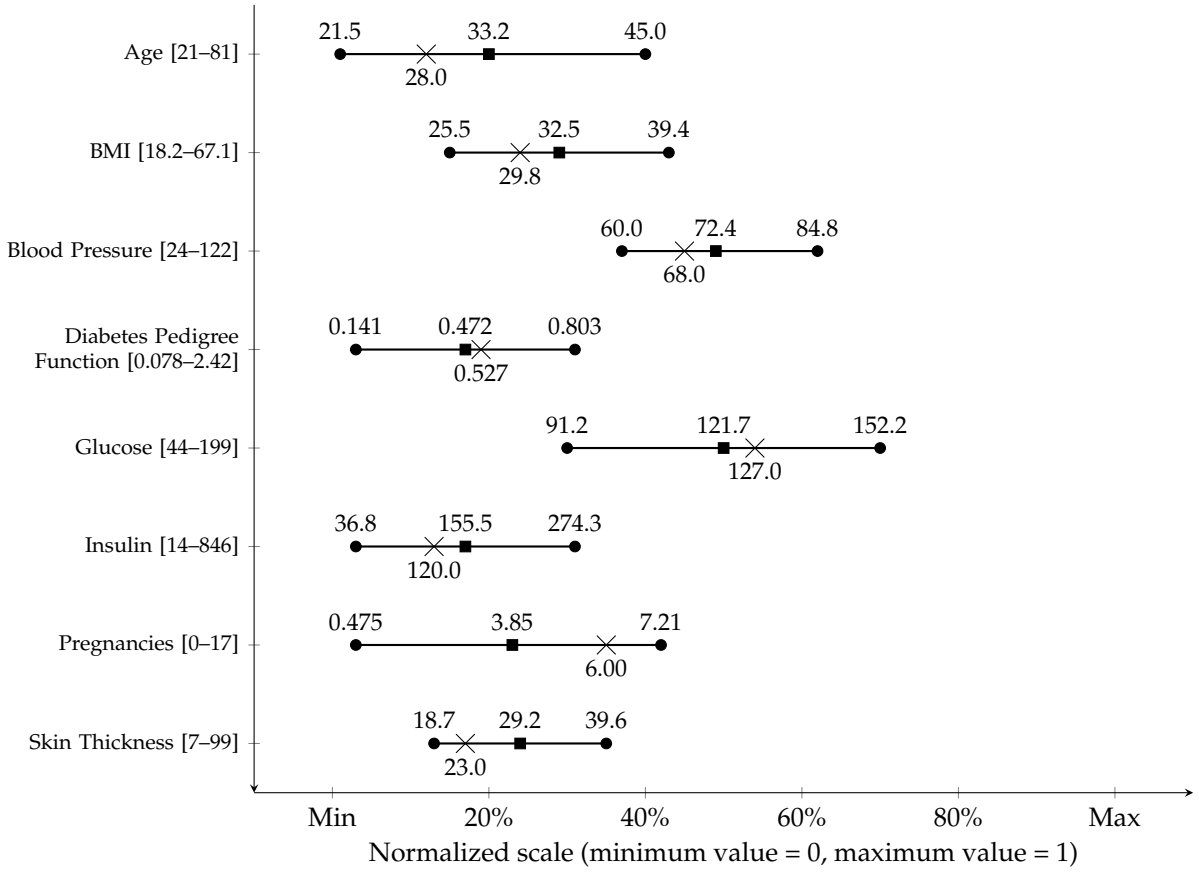
Figure 1. Calibration curves across three clinical datasets: (**a**) Pima Indians Diabetes Database, (**b**) Heart Failure Prediction dataset, and (**c**) Wisconsin Breast Cancer dataset. The diagonal line indicates perfect calibration. Dashed lines show raw probabilities and solid lines show three-parameter beta-calibrated probabilities.

Table 9. Calibration performance based on pooled out-of-fold predictions from 10-fold cross-validation. For each fold, beta calibration was fit on training-fold predictions and applied to the corresponding held-out fold. Ideal calibration intercept and slope are 0 and 1, respectively.

Dataset	Model	Brier	Intercept	Slope	Event Rate	Mean Risk
Pima Indians Diabetes	Raw	0.184	−0.284	0.530	0.349	0.374
	Beta-calibrated	0.174	−0.056	0.842	0.349	0.347
Wisconsin Breast Cancer	Raw	0.052	−0.205	0.162	0.373	0.373
	Beta-calibrated	0.044	−0.093	0.863	0.373	0.371
Heart Failure	Raw	0.121	0.098	0.505	0.553	0.548
	Beta-calibrated	0.111	0.006	0.982	0.553	0.553

4.3. Explainability

The interpretability of the proposed BNB framework is supported by complementary graphical and tabular analyses. Figures 2–4 present feature-wise thresholds on a normalized $[0, 1]$ scale together with the mean and mean $\pm$ standard deviation. This visualization makes the threshold location explicit relative to each feature distribution and clarifies how each predictor is partitioned into the Bernoulli states used for classification ($x \leq t$ and $x > t$).



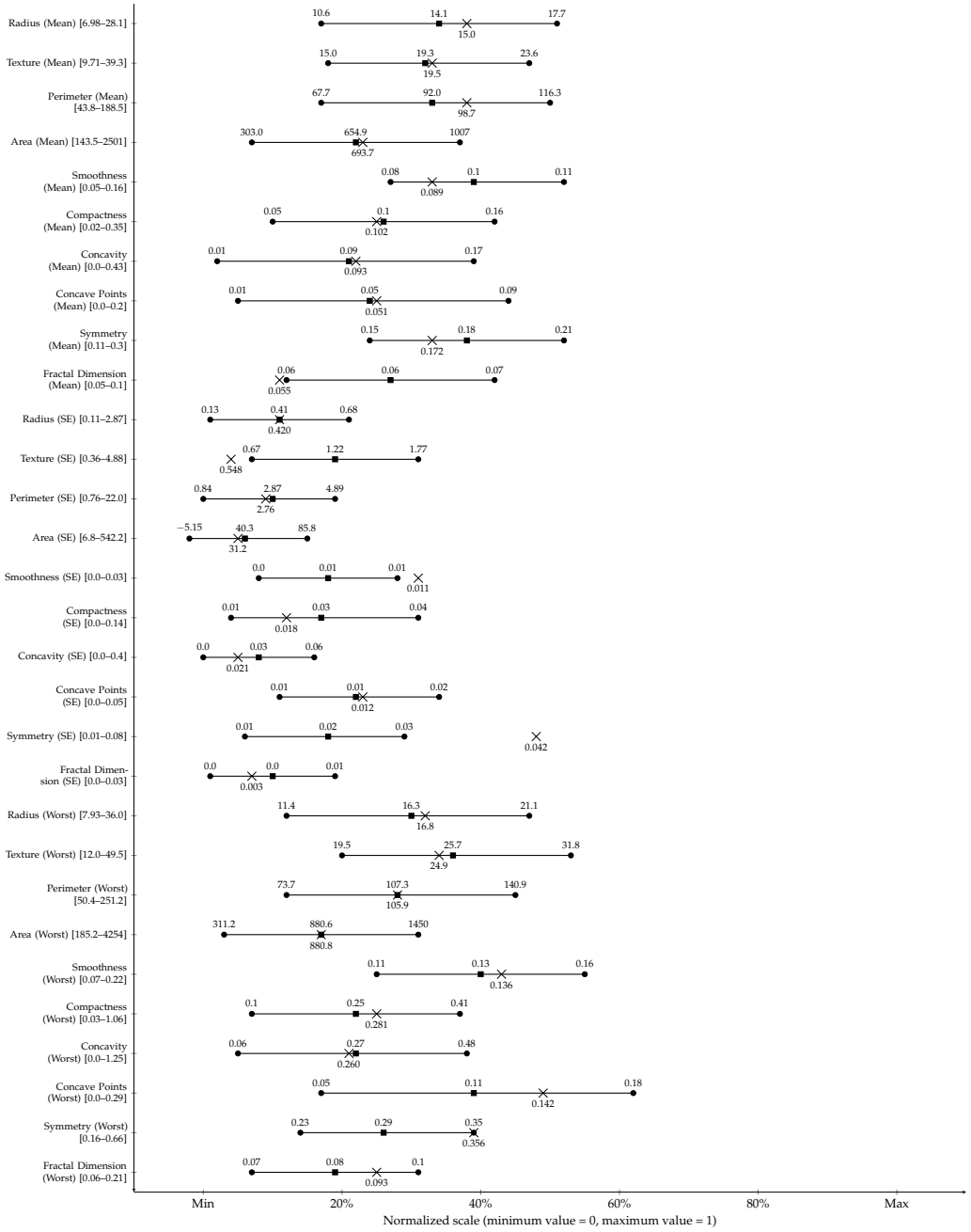
Legend: ■ Mean, ● ± Standard Deviation, × Binarization Threshold.

Figure 2. Forest Plot of Pima Indian Diabetes Database. The plot presents the distribution of key features on a normalized scale (0 to 1). The y-axis lists the features along with their respective value ranges, while the x-axis represents their normalized values. Horizontal lines indicate the spread of data, with markers representing key statistical values.

Tables 10–12 report outcome probabilities stratified by feature region, with 95% confidence intervals obtained by bootstrap resampling ($n = 1000$). For thresholded variables, these tables compare outcome probability at or below and above the threshold and therefore quantify both the direction and magnitude of feature effects on class membership. The absolute difference, ΔP, is included as a practical proxy for class-separation strength, with larger values indicating stronger discrimination. For categorical rows in the Heart Failure CNB table, threshold-based ΔP is not applicable.

For the Heart Failure dataset (Figure 4), thresholds are shown only for the five continuous predictors; the categorical predictors are not thresholded, and their category-specific probabilities are reported in Table 12.

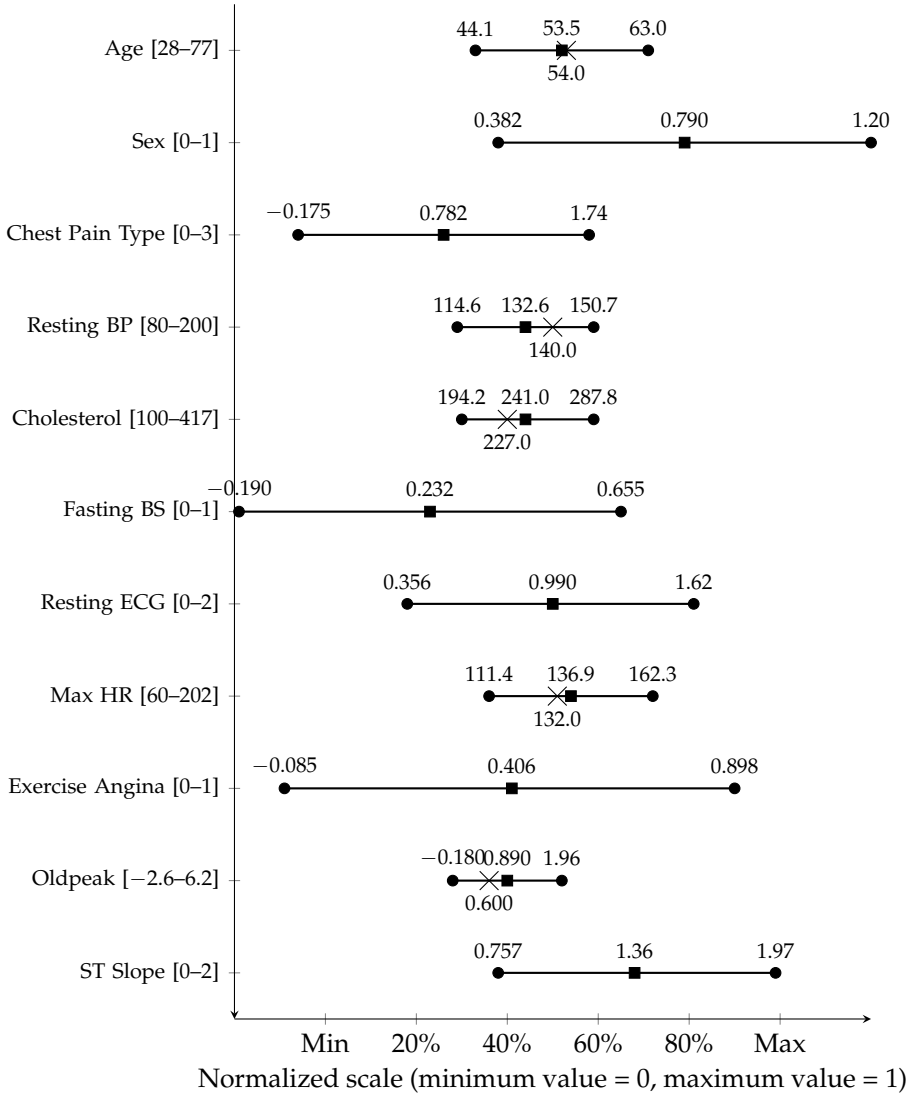
In the Heart Failure dataset, categorical features were encoded for Categorical Naïve Bayes (CNB), as summarized in Table 4.



Legend: ■ Mean, ● ± Standard Deviation, × Binarization Threshold.

Figure 3. Forest Plot of Wisconsin Breast Cancer Diagnosis dataset. This plot displays the range, mean, standard deviation, and optimal binarization threshold for each feature on a normalized scale (0 to 1). The binarization threshold is used to transform continuous medical data into a Bernoulli distribution, enabling the application of the Bernoulli Naïve Bayes classifier for supervised classification.

Table 13 reports the thresholds selected in the 100 training partitions of the repeated cross-validation. Most thresholds were reproducible; for example, the Pregnancies threshold was 6 in all 100 partitions. Diabetes Pedigree Function and Blood Pressure varied more, but on average at least 92% of patients kept the same binary state as under the full-data threshold in all three datasets, and the pooled ROC-AUC varied little between repetitions (SD $\leq$ 0.005). The bootstrap intervals in Tables 10–12 are wider because each replicate resamples a random 80% subset of the records, a stronger perturbation than the difference between cross-validation training sets.



Legend: ■ Mean, ● ± Standard Deviation, × Binarization Threshold.

Figure 4. Forest Plot of Heart Failure Prediction Dataset. This plot presents the distribution of key features on a normalized scale (0 to 1). The y-axis lists the features along with their respective value ranges, while the x-axis represents their normalized values. Horizontal lines indicate the spread of data, with markers representing key statistical values.

Table 10. Summary statistics for PIDD features. Here, t denotes the chi-square-selected threshold for each feature, $P(y = 1 \mid x \leq t) \equiv P(\text{diabetes} \mid \text{feature} \leq \text{threshold})$ and $P(y = 1 \mid x > t) \equiv P(\text{diabetes} \mid \text{feature} > \text{threshold})$ are the conditional probabilities of diabetes at or below and above the threshold, respectively, and ΔP is their absolute difference. Values in brackets denote 95% confidence intervals estimated via bootstrap resampling ($n = 1000$); each replicate resamples with replacement from a random 80% subset of the records.

Feature	Min	Max	t [95% CI]	$P(y = 1 \mid x \leq t)$ [95% CI]	$P(y = 1 \mid x > t)$ [95% CI]	ΔP [95% CI]
Pregnancies	0.000	17.0	6.00 [2.00, 7.00]	0.289 [0.209, 0.328]	0.562 [0.424, 0.667]	0.273 [0.167, 0.383]
Glucose	44.0	199.0	127.0 [112.0, 158.0]	0.192 [0.114, 0.275]	0.615 [0.531, 0.882]	0.423 [0.377, 0.632]
Blood Pressure	24.0	122.0	68.0 [60.0, 88.0]	0.247 [0.115, 0.334]	0.404 [0.348, 0.643]	0.157 [0.127, 0.328]
Skin Thickness	7.00	99.0	23.0 [19.0, 34.0]	0.157 [0.075, 0.270]	0.415 [0.366, 0.557]	0.258 [0.194, 0.360]
Insulin	14.0	846.0	120.0 [87.0, 148.1]	0.146 [0.044, 0.200]	0.505 [0.413, 0.603]	0.359 [0.296, 0.483]
BMI	18.2	67.1	29.8 [26.2, 31.2]	0.161 [0.041, 0.201]	0.463 [0.395, 0.533]	0.303 [0.265, 0.422]
Diabetes Pedigree Function	0.078	2.42	0.527 [0.200, 1.09]	0.291 [0.143, 0.343]	0.463 [0.371, 0.745]	0.173 [0.147, 0.432]
Age	21.0	81.0	28.0 [24.0, 31.0]	0.193 [0.116, 0.245]	0.491 [0.424, 0.566]	0.298 [0.243, 0.389]

Table 11. Summary statistics for Wisconsin Breast Cancer Diagnosis Dataset features. Here, t denotes the chi-square-selected threshold for each feature, $P(y = 1 \mid x \leq t) \equiv P(\text{cancer} \mid \text{feature} \leq \text{threshold})$ and $P(y = 1 \mid x > t) \equiv P(\text{cancer} \mid \text{feature} > \text{threshold})$ are the conditional probabilities of malignancy at or below and above the threshold, respectively, and ΔP is their absolute difference. Values in brackets denote 95% confidence intervals estimated via bootstrap resampling ($n = 1000$); each replicate resamples with replacement from a random 80% subset of the records.

Feature	Min	Max	t [95% CI]	$P(y = 1 \mid x \leq t)$ [95% CI]	$P(y = 1 \mid x > t)$ [95% CI]	ΔP [95% CI]
Radius (Mean)	6.98	28.1	15.0 [14.3, 15.3]	0.128 [0.083, 0.168]	0.936 [0.852, 0.978]	0.808 [0.743, 0.865]
Texture (Mean)	9.71	39.3	19.5 [18.1, 20.2]	0.174 [0.086, 0.218]	0.621 [0.536, 0.704]	0.447 [0.386, 0.550]
Perimeter (Mean)	43.8	188.5	98.7 [89.8, 99.6]	0.132 [0.053, 0.159]	0.946 [0.813, 0.985]	0.814 [0.738, 0.875]
Area (Mean)	143.5	2501.0	693.7 [641.2, 711.8]	0.126 [0.082, 0.163]	0.942 [0.870, 0.980]	0.816 [0.746, 0.873]
Smoothness (Mean)	0.053	0.163	0.089 [0.088, 0.108]	0.135 [0.070, 0.279]	0.493 [0.444, 0.695]	0.358 [0.299, 0.470]
Compactness (Mean)	0.019	0.345	0.102 [0.095, 0.124]	0.117 [0.070, 0.209]	0.692 [0.627, 0.849]	0.575 [0.515, 0.687]
Concavity (Mean)	0.000	0.427	0.093 [0.071, 0.111]	0.090 [0.034, 0.135]	0.845 [0.763, 0.936]	0.755 [0.703, 0.839]
Concave Points (Mean)	0.000	0.201	0.051 [0.045, 0.058]	0.057 [0.026, 0.098]	0.873 [0.823, 0.951]	0.815 [0.768, 0.885]
Symmetry (Mean)	0.106	0.304	0.172 [0.171, 0.211]	0.190 [0.124, 0.336]	0.493 [0.454, 0.827]	0.302 [0.267, 0.523]
Fractal Dimension (Mean)	0.050	0.097	0.055 [0.052, 0.073]	0.621 [0.310, 1.000]	0.344 [0.286, 0.700]	0.276 [0.181, 0.656]
Radius (SE)	0.112	2.87	0.420 [0.372, 0.546]	0.164 [0.107, 0.230]	0.805 [0.718, 0.960]	0.641 [0.576, 0.755]
Texture (SE)	0.360	4.88	0.548 [0.538, 1.89]	0.097 [0.000, 0.445]	0.388 [0.119, 0.468]	0.292 [0.123, 0.414]
Perimeter (SE)	0.757	22.0	2.76 [2.39, 3.43]	0.148 [0.095, 0.213]	0.775 [0.693, 0.900]	0.627 [0.564, 0.735]
Area (SE)	6.80	542.2	31.2 [30.5, 42.8]	0.101 [0.067, 0.168]	0.834 [0.805, 0.958]	0.734 [0.695, 0.836]
Smoothness (SE)	0.002	0.031	0.011 [0.004, 0.014]	0.392 [0.086, 0.503]	0.167 [0.000, 0.447]	0.225 [0.122, 0.485]
Compactness (SE)	0.002	0.135	0.018 [0.016, 0.025]	0.169 [0.107, 0.248]	0.538 [0.487, 0.663]	0.370 [0.314, 0.473]
Concavity (SE)	0.000	0.396	0.021 [0.020, 0.026]	0.095 [0.046, 0.158]	0.576 [0.506, 0.658]	0.481 [0.410, 0.568]
Concave Points (SE)	0.000	0.053	0.012 [0.009, 0.013]	0.178 [0.046, 0.216]	0.622 [0.505, 0.715]	0.444 [0.400, 0.551]
Symmetry (SE)	0.008	0.079	0.042 [0.011, 0.042]	0.360 [0.315, 0.800]	0.923 [0.224, 1.000]	0.563 [0.142, 0.675]
Fractal Dimension (SE)	0.001	0.030	0.003 [0.002, 0.004]	0.265 [0.137, 0.314]	0.476 [0.398, 0.578]	0.211 [0.170, 0.339]
Radius (Worst)	7.93	36.0	16.8 [15.6, 17.5]	0.087 [0.032, 0.125]	0.942 [0.843, 0.992]	0.855 [0.795, 0.906]
Texture (Worst)	12.0	49.5	24.9 [23.0, 29.3]	0.137 [0.063, 0.231]	0.573 [0.502, 0.733]	0.436 [0.381, 0.550]
Perimeter (Worst)	50.4	251.2	105.9 [101.9, 117.2]	0.049 [0.019, 0.120]	0.871 [0.840, 1.000]	0.821 [0.798, 0.915]
Area (Worst)	185.2	4254.0	880.8 [739.1, 932.7]	0.096 [0.026, 0.123]	0.956 [0.827, 0.987]	0.860 [0.789, 0.911]
Smoothness (Worst)	0.071	0.223	0.136 [0.130, 0.148]	0.202 [0.132, 0.278]	0.603 [0.537, 0.763]	0.401 [0.340, 0.532]
Compactness (Worst)	0.027	1.06	0.281 [0.201, 0.345]	0.175 [0.079, 0.240]	0.760 [0.616, 0.908]	0.585 [0.514, 0.704]
Concavity (Worst)	0.000	1.25	0.260 [0.208, 0.349]	0.069 [0.017, 0.144]	0.760 [0.680, 0.893]	0.691 [0.639, 0.782]
Concave Points (Worst)	0.000	0.291	0.142 [0.111, 0.149]	0.090 [0.034, 0.118]	0.937 [0.836, 0.985]	0.847 [0.787, 0.906]
Symmetry (Worst)	0.157	0.664	0.356 [0.281, 0.360]	0.306 [0.172, 0.335]	0.892 [0.553, 0.982]	0.587 [0.354, 0.694]
Fractal Dimension (Worst)	0.055	0.207	0.093 [0.078, 0.102]	0.286 [0.181, 0.333]	0.659 [0.487, 0.803]	0.373 [0.276, 0.518]

Perhaps the most direct demonstration of the framework's interpretability is that the entire classification procedure can be carried out by hand. Table 14 presents a complete leave-one-out worked example that integrates all explainability components. It shows how thresholds, conditional probabilities, and class priors are combined and normalized to classify a single sample. For each feature, the table reports the training-set count from which the conditional probability is obtained, so that every quantity can be checked independently, and the final rows give the product of the conditional probabilities, the effect of multiplying by the class prior, and the normalized posterior probabilities.

As shown in Table 14, the sample is classified as Diabetes, with posterior probability 0.984, compared with 0.0165 for No Diabetes, which agrees with the patient's recorded outcome. This value is the uncalibrated model output; as shown in Table 9, raw probabilities are overconfident and should be calibrated before being communicated as risk. This calculation does not include the usual logarithmic transformation of probabilities, which is commonly applied in Naïve Bayes implementations to avoid numerical underflow. This, however, does not affect the interpretability of the example, as the same terms are combined in the same way regardless of whether log transformation is applied.

Table 12. Summary statistics for Heart Failure Prediction Dataset. For continuous features, the table includes minimum, maximum, and threshold values, along with conditional probabilities $P(\text{heart disease} \mid \text{feature} \leq \text{threshold})$ and $P(\text{heart disease} \mid \text{feature} > \text{threshold})$, and absolute probability difference ΔP. Categorical features are maintained as categorical with probability of heart disease shown for each category. The modeled outcome is the dataset's `HeartDisease` indicator; the dataset is distributed under the name "Heart Failure Prediction Dataset", but the label it carries denotes presence of heart disease. Fasting blood sugar is recorded in this dataset as a binary indicator of fasting blood sugar above 120 mg/dL, and is therefore treated as categorical alongside Sex and Exercise Angina rather than being thresholded. Values in brackets denote 95% confidence intervals estimated via bootstrap resampling ($n = 1000$); each replicate resamples with replacement from a random 80% subset of the records.

Continuous Features

Feature	Min.	Max.	t [95% CI]	$P(y = 1 \mid x \leq t)$ [95% CI]	$P(y = 1 \mid x > t)$ [95% CI]	ΔP [95% CI]
Oldpeak	−2.60	6.20	0.600 [0.400, 1.00]	0.343 [0.284, 0.401]	0.776 [0.727, 0.837]	0.433 [0.367, 0.511]
Cholesterol	100.0	417.0	227.0 [204.0, 236.0]	0.415 [0.324, 0.487]	0.627 [0.578, 0.684]	0.212 [0.151, 0.311]
MaxHR	60.0	202.0	132.0 [128.0, 157.0]	0.762 [0.650, 0.825]	0.383 [0.216, 0.426]	0.378 [0.331, 0.475]
Age	28.0	77.0	54.0 [45.0, 57.0]	0.413 [0.238, 0.466]	0.700 [0.611, 0.770]	0.287 [0.238, 0.403]
RestingBP	80.0	200.0	140.0 [120.0, 156.0]	0.512 [0.429, 0.555]	0.680 [0.581, 0.824]	0.168 [0.106, 0.307]

Categorical Features

Feature	Category Code	Category Name	Probability [95% CI]
Exercise Angina	1	Yes	0.851 [0.807, 0.892]
	0	No	0.349 [0.299, 0.401]
Chest Pain Type	0	ASY	0.792 [0.743, 0.833]
	3	TA	0.435 [0.250, 0.621]
	2	NAP	0.350 [0.272, 0.433]
	1	ATA	0.140 [0.079, 0.205]
Fasting Blood Sugar	1	Yes	0.796 [0.732, 0.856]
	0	No	0.479 [0.435, 0.524]
Resting ECG	2	ST	0.657 [0.578, 0.745]
	0	LVH	0.567 [0.477, 0.651]
	1	Normal	0.514 [0.464, 0.570]
Sex	1	Male	0.630 [0.587, 0.672]
	0	Female	0.262 [0.184, 0.344]
ST Slope	1	Flat	0.828 [0.783, 0.872]
	0	Down	0.790 [0.660, 0.912]
	2	Up	0.198 [0.150, 0.250]

Table 13. Stability of the χ^2-selected thresholds over the 100 training partitions of ten repetitions of stratified 10-fold cross-validation. Threshold: value selected on the full dataset. Same threshold: percentage of partitions selecting that value. Same side: mean percentage of patients whose binary state is unchanged when the partition threshold is used. Pooled ROC-AUC over the ten repetitions (mean ± SD): 0.795 ± 0.005 (Pima), 0.984 ± 0.001 (Wisconsin) and 0.917 ± 0.001 (Heart Failure).

Feature	Threshold	Partition Median [2.5%, 97.5%]	Same Threshold (%)	Same Side (%)
PIDD				
Pregnancies	6.00	6.00 [6.00, 6.00]	100	100.0
Glucose	127.0	127.0 [123.0, 154.0]	69	95.9
Blood Pressure	68.0	68.0 [61.0, 75.0]	58	93.3
Skin Thickness	23.0	23.0 [23.0, 29.0]	93	98.6
Insulin	120.0	112.0 [108.0, 126.0]	31	96.7
BMI	29.8	28.8 [27.3, 29.9]	43	96.0
Diabetes Pedigree Function	0.527	0.527 [0.407, 0.718]	43	94.0
Age	28.0	28.0 [28.0, 30.0]	90	99.4

Table 13. *Cont.*

Feature	Threshold	Partition Median [2.5%, 97.5%]	Same Threshold (%)	Same Side (%)
Wisconsin Breast Cancer				
Radius (Mean)	15.04	15.04 [15.00, 15.04]	90	100.0
Texture (Mean)	19.46	19.31 [18.45, 19.60]	31	96.3
Perimeter (Mean)	98.73	98.73 [96.39, 99.58]	51	98.9
Area (Mean)	693.7	693.7 [690.2, 693.7]	90	100.0
Smoothness (Mean)	0.089	0.089 [0.089, 0.090]	49	99.2
Compactness (Mean)	0.102	0.102 [0.102, 0.114]	92	99.5
Concavity (Mean)	0.093	0.093 [0.086, 0.103]	41	98.8
Concave Points (Mean)	0.051	0.051 [0.049, 0.056]	64	99.4
Symmetry (Mean)	0.172	0.172 [0.171, 0.206]	43	92.3
Fractal Dimension (Mean)	0.055	0.055 [0.054, 0.057]	49	97.6
Radius (SE)	0.420	0.420 [0.386, 0.420]	75	99.5
Texture (SE)	0.548	0.548 [0.542, 0.771]	64	97.7
Perimeter (SE)	2.76	2.76 [2.75, 2.89]	82	99.8
Area (SE)	31.24	31.24 [30.90, 35.13]	51	97.9
Smoothness (SE)	0.011	0.011 [0.009, 0.012]	73	98.9
Compactness (SE)	0.018	0.022 [0.018, 0.024]	37	94.6
Concavity (SE)	0.021	0.021 [0.021, 0.021]	73	99.9
Concave Points (SE)	0.012	0.012 [0.009, 0.012]	32	95.4
Symmetry (SE)	0.042	0.042 [0.018, 0.042]	86	97.7
Fractal Dimension (SE)	0.003	0.003 [0.002, 0.004]	31	92.1
Radius (Worst)	16.77	16.77 [16.76, 16.77]	90	100.0
Texture (Worst)	24.85	24.85 [23.31, 25.91]	42	96.8
Perimeter (Worst)	105.9	109.8 [105.0, 115.0]	41	96.0
Area (Worst)	880.8	880.8 [867.1, 880.8]	82	99.9
Smoothness (Worst)	0.136	0.136 [0.136, 0.139]	73	99.5
Compactness (Worst)	0.281	0.281 [0.266, 0.285]	51	99.0
Concavity (Worst)	0.260	0.260 [0.260, 0.267]	86	99.8
Concave Points (Worst)	0.142	0.142 [0.141, 0.146]	62	99.6
Symmetry (Worst)	0.356	0.356 [0.340, 0.358]	56	98.9
Fractal Dimension (Worst)	0.093	0.093 [0.090, 0.096]	58	98.9
Heart Failure Prediction				
Age	54.0	54.0 [54.0, 55.0]	70	98.7
Resting BP	140.0	140.0 [140.0, 140.0]	98	99.6
Cholesterol	227.0	227.0 [227.0, 236.0]	57	98.2
MaxHR	132.0	132.0 [130.0, 134.0]	64	99.5
Oldpeak	0.600	0.600 [0.600, 0.800]	52	98.8

Table 14. Leave-one-out worked example of the Bernoulli Naïve Bayes procedure for one Pima Indians Diabetes sample. For each feature, the patient's measurement is compared with the binarization threshold, and the count column n gives the number of training patients of that class falling on the same side of the threshold as the patient. The conditional probability is the Laplace-smoothed proportion (see Equation (4)). The leave-one-out training set contains 267 diabetes patients (class prior 0.348) and 500 non-diabetes patients (class prior 0.652). Thresholds can be estimated with any statistical software using the algorithm described in Section 3; every remaining quantity in the table can be obtained with a calculator.

Feature	Binarization Threshold	Patient Value	Binarized Value	Diabetes		No Diabetes	
				n	$P(x \mid D)$	n	$P(x \mid \bar{D})$
Pregnancies	6.00	9.00	1	94	0.353	74	0.149
Glucose	127.0	171.0	1	173	0.652	109	0.220
Blood Pressure	68.0	110.0	1	181	0.719	268	0.557

Table 14. *Cont.*

Feature	Binarization Threshold	Patient Value	Binarized Value	Diabetes		No Diabetes	
				n	$P(x \mid D)$	n	$P(x \mid \bar{D})$
Skin Thickness	23.0	24.0	1	152	0.845	216	0.598
Insulin	112.0	240.0	1	107	0.824	113	0.429
BMI	29.8	45.4	1	220	0.828	256	0.521
Diabetes Pedigree Function	0.527	0.721	1	119	0.446	139	0.279
Age	28.0	54.0	1	196	0.732	204	0.408
Product of the eight conditional probabilities				0.0312		0.000279	
Multiplied by the class prior (0.348 and 0.652)				0.0109		0.000182	
Normalized posterior probability				0.9835		0.0165	

5. Discussion

The benchmark datasets used in this study, Pima Indians Diabetes, Wisconsin Breast Cancer, and Heart Failure Prediction, were selected for their broad adoption as evaluation standards in clinical machine learning, enabling direct comparison with published methods. However, the practical value of an interpretable, threshold-based framework extends well beyond these specific conditions. Consider the memory clinic setting: a nurse practitioner conducting a cognitive triage assessment needs not only a risk estimate but an accountable explanation, one that can be communicated to a worried patient or family member in real time, without reference to a black-box model or a proprietary algorithm. Similar requirements arise in community screening for type 2 diabetes in low-resource settings, in primary care cardiovascular risk communication, and in any clinical context where the decision-maker is not the data scientist who built the model. The Bernoulli Naïve Bayes framework proposed here is intended for these conditions: settings where interpretability is not a desirable feature but an operational requirement, where the clinician must be able to trace, explain, and if necessary defend every classification to a patient, a colleague, or a regulatory body. Future work should evaluate the framework prospectively in such settings, including through clinician usability studies that assess whether the decision rules produced are not only formally interpretable but practically useful at the point of care.

In this context, the present study evaluated whether Bernoulli Naïve Bayes combined with supervised χ^2-guided binarization can provide an interpretable option for clinical prediction without a substantial loss in discriminatory ability. The comparative analyses place the method in a competitive position relative to many conventional tabular classifiers on the three benchmark datasets. Although BNB was not uniformly superior, no statistically significant difference from a broad set of baselines was detected in a large share of the comparisons. This pattern supports the view that a threshold-based Bernoulli representation can retain useful predictive information while preserving a verifiable decision structure.

A strength of the proposed method is that the full classification pathway can be reconstructed directly from the training data. Once the thresholds have been estimated, each continuous predictor is converted into a binary rule, the relevant class-conditional probabilities can be tabulated, and the posterior probability for a new patient can be obtained with straightforward arithmetic. Table 14 demonstrates that the model can be presented as a complete worked example rather than as an opaque scoring mechanism.

The comparative results also clarify the nature of the trade-off between discrimination and interpretability. When BNB was surpassed, the gains generally came from ensemble methods, neural network models, or selected non-linear SVM configurations. Such models can achieve stronger predictive performance, but their internal decision paths are less accessible for direct inspection at the level of an individual patient. By contrast,

the proposed BNB approach expresses each predictor through an explicit threshold together with a compact set of conditional probabilities. In applications where explicit reasoning and reproducible calculation are necessary, that simplicity may justify accepting a modest reduction in raw performance.

The threshold sensitivity analysis provides further support for the proposed binarization rule. As shown in Table 5, alternative thresholding strategies occasionally matched the χ^2 criterion numerically, yet none achieved a statistically significant improvement in any dataset. Several alternatives were significantly worse in at least one setting. This pattern suggests that the usefulness of the method does not depend on a narrowly tuned threshold rule. Rather, the χ^2 procedure appears to offer a stable and defensible way to convert continuous variables into Bernoulli features while preserving model clarity.

The calibration analysis adds a complementary perspective to the discrimination results. In clinical prediction, the quality of the probability estimates is often as important as rank-based separation, especially when probabilities may inform downstream decision thresholds. The raw probabilities were overconfident (calibration slopes 0.16–0.53), consistent with the tendency of Naïve Bayes to count correlated evidence more than once when the conditional independence assumption is violated; the effect was largest in the Wisconsin dataset, whose radius, perimeter, and area features are strongly correlated. Raw BNB probabilities should therefore not be used as clinical risk estimates. The discrimination of the model and its explicit rules do not depend on calibration, but when probabilities are communicated or used for decisions, a calibration step such as beta calibration, fitted on training data, is required.

The explainability analyses support the interpretability claims of the method. Figures 2–4 place the selected thresholds in relation to the empirical feature distributions, while Tables 10–12 report outcome probabilities stratified by the selected feature thresholds or categorical levels, together with uncertainty intervals. Taken together with the patient-level posterior calculation in Table 14, these components permit inspection of the model at the level of thresholds, feature-specific probabilities, and final posterior values. This level of detail is difficult to obtain from more complex classifiers without reliance on post hoc explanation methods.

The results also point to settings in which caution is warranted. Thresholding categorical predictors in the same manner as continuous variables can merge clinically meaningful categories and reduce the informational content of the original feature. A more appropriate strategy is to binarize only continuous variables while preserving categorical predictors in encoded form, as was done in the Categorical Naïve Bayes analysis for the Heart Failure dataset. As shown in the methods section, Bernoulli Naïve Bayes is the special case of Categorical Naïve Bayes in which each *predictor* takes only two levels; this concerns the number of levels per feature and is independent of the number of outcome classes. Using Categorical Naïve Bayes preserves the original category structure while maintaining the simplicity of the overall approach.

Similar caution applies to the interpretation of ΔP. The reported values provide an intuitive proxy for class-separation strength by describing how outcome probability changes on the two sides of a threshold, but they should not be treated as definitive measures of feature importance. Clinical predictors are often correlated, and the Naïve Bayes independence assumption does not fully represent those dependencies or their possible interactions. Consequently, the apparent contribution of one variable may change when other predictors are added or removed. Feature selection is therefore better regarded as a subset optimization problem than as a univariate ranking exercise [76–79]. Even so, the ordering of variables such as Glucose, Insulin, and Age in the Pima dataset remains clinically plausible and aligns with earlier reports [80,81].

These considerations should be interpreted together with the standard limitations of Naïve Bayes. The independence assumption is unlikely to hold exactly in clinical data, the thresholds of some features, such as Diabetes Pedigree Function and Blood Pressure, vary between resamples and should therefore be read as data-driven cut-points to be re-estimated in the target population rather than as fixed clinical cutoffs, and the present evaluation is restricted to benchmark datasets rather than external cohorts. These limitations constrain the scope of generalization. It should also be noted that interpretability in this study was assessed structurally, through explicit thresholds, conditional probability tables, and a worked classification example, rather than through formal evaluation with clinical users. Whether clinicians find these representations intuitive and actionable in practice remains an open question, and usability studies in real clinical settings are an important direction for future work.

6. Conclusions

This study evaluated a Bernoulli Naïve Bayes method based on supervised χ^2-guided binarization for tabular clinical prediction. The results support its use as an interpretable reference model that combines transparent decision rules with competitive predictive performance on the evaluated datasets. The framework achieved ROC-AUC values of 0.794 for the NaN-aware Pima Indians Diabetes analysis, 0.984 for the Wisconsin Breast Cancer dataset, and 0.918 for the Heart Failure Prediction dataset. The method is also computationally light at both training and inference, because each predictor is reduced to a single thresholded Bernoulli term and classification requires only simple probability calculations. Rather than serving as a replacement for more flexible classifiers, the method is best viewed as a simple and auditable alternative when transparency and computational efficiency are primary design goals.

The method nevertheless retains the usual limitations of Naïve Bayes. The conditional independence assumption is unlikely to hold exactly in clinical data, threshold estimates are influenced by training data variability, and mixed continuous and categorical predictors require careful handling. Exhaustive threshold evaluation over all candidate cut points may also become burdensome for larger datasets, and the absence of built-in feature weighting may limit performance when predictors differ substantially in informativeness. Strong ROC-AUC performance should therefore not be interpreted as guaranteed superiority on threshold-dependent metrics such as accuracy, precision, recall, or F1-score. The present evaluation was also limited to benchmark datasets, which constrains the scope of immediate clinical generalization. Future work should prioritize external validation in independent cohorts, assessment of threshold stability in other populations, and study of feature-selection or weighting strategies that preserve interpretability. Cross-dataset studies, in which the model is developed on one cohort and tested on another, are also needed to assess how well the method generalizes. Additionally, other classifiers could benefit from similar binarization strategies, so more research is needed to explore how broadly this method can be applied.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/bioengineering13101152/s1. Table S1: Complete comparison for the Pima Indians Diabetes Database; Table S2: Complete comparison for the Wisconsin Breast Cancer dataset; Table S3: Complete comparison for the Heart Failure Prediction dataset. Tables S1–S3 report accuracy, recall, precision and F1-score in addition to the AUC and statistical tests shown in the main text.

Author Contributions: Conceptualization, A.G.; methodology, A.G.; software, A.G.; formal analysis, A.G., A.N.d.l.C., G.B., and X.H.; data curation, A.G.; visualization, A.G.; writing—original

draft preparation, A.G.; writing—review and editing, A.G., A.N.d.l.C., G.B., and X.H.; supervision, A.N.d.l.C. and X.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study used only publicly available, de-identified benchmark datasets and did not involve new data collection, direct interaction with human participants, or access to identifiable private information.

Informed Consent Statement: Not applicable.

Data Availability Statement: All datasets analyzed in this study are publicly available benchmark datasets: the Pima Indians Diabetes Database [48], the Breast Cancer Wisconsin (Diagnostic) Database [49], and the Heart Failure Prediction Dataset [50]. No new data were generated. The code used for all analyses is publicly available at https://github.com/AntonyGarcia/chi2-binarized-naive-bayes (accessed on 28 September 2026).

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
AUC	Area under the curve
BMI	Body mass index
BNB	Bernoulli Naïve Bayes
CI	Confidence interval
CNB	Categorical Naïve Bayes
ECG	Electrocardiogram
NB	Naïve Bayes
OOF	Out-of-fold
PIDD	Pima Indians Diabetes Database
RBF	Radial basis function
ROC	Receiver operating characteristic
SVM	Support vector machine
WHO	World Health Organization

References

1. Moreno-Ibarra, M.A.; Villuendas-Rey, Y.; Lytras, M.D.; Yáñez-Márquez, C.; Salgado-Ramírez, J.C. Classification of Diseases Using Machine Learning Algorithms: A Comparative Study. *Mathematics* **2021**, *9*, 1817. [CrossRef]
2. An, Q.; Rahman, S.; Zhou, J.; Kang, J.J. A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges. *Sensors* **2023**, *23*, 4178. [CrossRef] [PubMed]
3. Ennab, M.; Mcheick, H. Enhancing interpretability and accuracy of AI models in healthcare: A comprehensive review on challenges and future directions. *Front. Robot. AI* **2024**, *11*, 1444763. [CrossRef] [PubMed]
4. Chen, X.; Chen, H.; Nan, S.; Kong, X.; Duan, H.; Zhu, H. Dealing with Missing, Imbalanced, and Sparse Features During the Development of a Prediction Model for Sudden Death Using Emergency Medicine Data: Machine Learning Approach. *JMIR Med. Inform.* **2023**, *11*, e38590. [CrossRef] [PubMed]
5. Razzaghi, T.; Roderick, O.; Safro, I.; Marko, N. Multilevel Weighted Support Vector Machine for Classification on Healthcare Data with Missing Values. *PLoS ONE* **2016**, *11*, e0155119. [CrossRef] [PubMed]
6. Kumar, A.; Veeraiah, V.; Gongada, T.N.; Ahamad, S.; Khan, H.; Gupta, A. Explainable Machine Learning Models for Clinical Decision Support Systems. In Proceedings of the 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), Kamand, India, 24–28 June 2024; pp. 1–6. [CrossRef]
7. Richens, J.G.; Lee, C.M.; Johri, S. Improving the accuracy of medical diagnosis with causal machine learning. *Nat. Commun.* **2020**, *11*, 3923. [CrossRef] [PubMed]

8. Fuse, Y.; Murphy, S.N.; Ikari, H.; Takahashi, A.; Fuse, K.; Kawakami, E. Artificial intelligence in clinical data analysis: A review of large language models, foundation models, digital twins, and allergy applications. *Allergol. Int.* **2025**, *74*, 499–513. [CrossRef] [PubMed]
9. Ahsan, M.M.; Luna, S.A.; Siddique, Z. Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. *Healthcare* **2022**, *10*, 541. [CrossRef] [PubMed]
10. Tjoa, E.; Guan, C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *32*, 4793–4813. [CrossRef] [PubMed]
11. Topol, E.J. High-performance medicine: The convergence of human and artificial intelligence. *Nat. Med.* **2019**, *25*, 44–56. [CrossRef] [PubMed]
12. Kelly, C.J.; Karthikesalingam, A.; Suleyman, M.; Corrado, G.; King, D. Key challenges for delivering clinical impact with artificial intelligence. *BMC Med.* **2019**, *17*, 195. [CrossRef] [PubMed]
13. Pérez, A.; Larrañaga, P.; Inza, I. Supervised classification with conditional gaussian networks: Increasing the structure complexity from naïve bayes. *Int. J. Approx. Reason.* **2006**, *43*, 1–25. [CrossRef]
14. Domingos, P.; Pazzani, M. On the optimality of the simple bayesian classifier under zero–one loss. *Mach. Learn.* **1997**, *29*, 103–130. [CrossRef]
15. McCallum, A.; Nigam, K. A comparison of event models for naïve bayes text classification. In Proceedings of the AAAI Workshop on Learning for Text Categorization, Madison, WI, USA, 26 July 1998; pp. 41–48.
16. Singh, G.; Kumar, B.; Gaur, L.; Tyagi, A. Comparison between multinomial and bernoulli naïve bayes for text classification. In Proceedings of the 2019 International Conference on Automation, Computational and Technology Management (ICACTM), London, UK, 24–26 April 2019; pp. 593–596. [CrossRef]
17. Kharisma, R.S.; Muttafi'ah; Dahlan, A. Comparison of naïve bayes algorithm model combinations with term weighting techniques in sentiment analysis. In Proceedings of the 2021 4th International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 30–31 August 2021; pp. 160–163. [CrossRef]
18. Xu, S. Bayesian naïve bayes classifiers to text classification. *J. Inf. Sci.* **2018**, *44*, 48–59. [CrossRef]
19. Kibriya, A.M.; Frank, E.; Pfahringer, B.; Holmes, G. Multinomial naive bayes for text categorization revisited. In *AI 2004: Advances in Artificial Intelligence* ; Springer: Berlin/Heidelberg, Germany, 2004; pp. 488–499.
20. Aboy, M.; Minssen, T.; Vayena, E. Navigating the EU AI Act: Implications for regulated digital medical products. *npj Digit. Med.* **2024**, *7*, 237. [CrossRef] [PubMed]
21. World Health Organization. *Ethics and Governance of Artificial Intelligence for Health: WHO Guidance*, 1st ed.; World Health Organization: Geneva, Switzerland, 2021.
22. Zhang, H. The optimality of naive bayes. In *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference (FLAIRS 2004)*; AAAI: Menlo Park, CA, USA, 2004; pp. 562–567.
23. Weegar, R.; Sundström, K. Using machine learning for predicting cervical cancer from Swedish electronic health records by mining hierarchical representations. *PLoS ONE* **2020**, *15*, e0237911. [CrossRef] [PubMed]
24. Taghizadeh, E.; Heydarheydari, S.; Saberi, A.; JafarpoorNesheli, S.; Rezaeijo, S.M. Breast cancer prediction with transcriptome profiling using feature selection and machine learning methods. *BMC Bioinform.* **2022**, *23*, 410. [CrossRef] [PubMed]
25. Patra, R. Prediction of Lung Cancer Using Machine Learning Classifier. In *Computing Science, Communication and Security*; Chaubey, N., Parikh, S., Amin, K., Eds.; Communications in Computer and Information Science, Volume 1235; Springer: Singapore, 2020; pp. 132–142. [CrossRef] [PubMed]
26. Miranda, E.; Irwansyah, E.; Amelga, A.Y.; Maribondang, M.M.; Salim, M. Detection of cardiovascular disease risk's level for adults using naive Bayes classifier. *Healthc. Inform. Res.* **2016**, *22*, 196–205. [CrossRef] [PubMed]
27. Gupta, K.; Paul, C.; Jain, N. Early Autism Spectrum Disorder Prediction Using Fine-Tuned Bernoulli's Naive Bayes Algorithm. In *Computation of Artificial Intelligence and Machine Learning*; Bairwa, A.K., Tiwari, V., Vishwakarma, S.K., Tuba, M., Ganokratanaa, T., Eds.; Springer Nature: Cham, Switzerland, 2025; Volume 2184, pp. 278–289. [CrossRef]
28. Singh, R.; Kaur, S.; Singh, G.; Kaur, M.; Kaur, P. Optimizing early diagnosis by integrating multiple classifiers for predicting brain stroke and critical diseases. *Sci. Rep.* **2024**, *14*, 28429. [CrossRef] [PubMed]
29. Fayyad, U.M.; Irani, K.B. Multi-interval discretization of continuous-valued attributes for classification learning. In Proceedings of the 13th International Joint Conference on Artificial Intelligence (IJCAI 1993), Chambéry, France, 28 August–3 September 1993; pp. 1022–1029.
30. Kerber, R. ChiMerge: Discretization of numeric attributes. In Proceedings of the 10th National Conference on Artificial Intelligence (AAAI-92), San Jose, CA, USA, 12–16 July 1992; pp. 123–128.
31. Tay, F.E.H.; Shen, L. A modified chi2 algorithm for discretization. *IEEE Trans. Knowl. Data Eng.* **2002**, *14*, 666–670. [CrossRef]
32. Su, C.; Hsu, J. An extended chi2 algorithm for discretization of real-value attributes. *IEEE Trans. Knowl. Data Eng.* **2005**, *17*, 437–441. [CrossRef]
33. Liu, H.; Hussain, F.; Tan, C.L.; Dash, M. Discretization: An enabling technique. *Data Min. Knowl. Discov.* **2002**, *6*, 393–423. [CrossRef]

34. Ramírez-Gallego, S.; García, S.; Mouriño-Talín, H.; Martínez-Rego, D.; Bolón-Canedo, V.; Alonso-Betanzos, A.; Benítez, J.M.; Herrera, F. Data discretization: Taxonomy and big data challenge. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2016**, *6*, 5–21. [CrossRef]
35. Kotsiantis, S.; Kanellopoulos, D. Discretization techniques: A recent survey. *GESTS Int. Trans. Comput. Sci. Eng.* **2006**, *32*, 47–58.
36. García, S.; Luengo, J.; Herrera, F. A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. *IEEE Trans. Knowl. Data Eng.* **2013**, *25*, 734–750. [CrossRef]
37. Bohacik, J.; Zabovsky, M. Discretization for naive Bayes taking the specifics of heart data into account. *J. Inf. Organ. Sci.* **2019**, *43*, 1–14. [CrossRef]
38. Shmulevich, I.; Zhang, W. Binary analysis and optimization-based normalization of gene expression data. *Bioinformatics* **2002**, *18*, 555–565. [CrossRef] [PubMed]
39. Tran, B.; Xue, B.; Zhang, M. A new representation in PSO for discretization-based feature selection. *IEEE Trans. Cybern.* **2018**, *48*, 1733–1746. [CrossRef] [PubMed]
40. Lustgarten, J.L.; Visweswaran, S.; Gopalakrishnan, V.; Cooper, G.F. Application of an efficient Bayesian discretization method to biomedical data. *BMC Bioinform.* **2011**, *12*, 309. [CrossRef] [PubMed]
41. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **2019**, *1*, 206–215. [CrossRef] [PubMed]
42. Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; Elhadad, N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015; pp. 1721–1730.
43. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. Causability and explainability of artificial intelligence in medicine. *WIREs Data Min. Knowl. Discov.* **2019**, *9*, e1312. [CrossRef] [PubMed]
44. Ghassemi, M.; Naumann, T.; Schulam, P.; Beam, A.L.; Chen, I.Y.; Ranganath, R. A Review of Challenges and Opportunities in Machine Learning for Health. *AMIA Summits Transl. Sci. Proc.* **2020**, *2020*, 191–200. [PubMed]
45. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* **2019**, *51*, 93. [CrossRef]
46. Miller, R.; Siegmund, D. Maximally Selected Chi Square Statistics. *Biometrics* **1982**, *38*, 1011–1016. [CrossRef] [PubMed]
47. Altman, D.G.; Lausen, B.; Sauerbrei, W.; Schumacher, M. Dangers of Using "Optimal" Cutpoints in the Evaluation of Prognostic Factors. *J. Natl. Cancer Inst.* **1994**, *86*, 829–835. [CrossRef] [PubMed]
48. Smith, J.W.; Everhart, J.; Dickson, W.; Knowler, W.; Johannes, R. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. In Proceedings of the Annual Symposium on Computer Application in Medical Care, Washington, DC, USA, November 1988; pp. 261–265.
49. William Wolberg, O.M. Breast Cancer Wisconsin (Diagnostic). 1993. Available online: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic (accessed on 28 June 2026).
50. Janosi, A.; Steinbrunn, W.; Pfisterer, M.; Detrano, R. Heart Failure Prediction Dataset. Kaggle. Compiled and Distributed on Kaggle by Fedesoriano; Combines the Cleveland, Hungarian, Switzerland, Long Beach VA and Statlog Heart Datasets into 918 Records with 11 Shared Features. 2021. Available online: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction (accessed on 23 February 2025).
51. Emerson, J.D.; Colditz, G.A. Use of Statistical Analysis in The New England Journal of Medicine. *N. Engl. J. Med.* **1983**, *309*, 709–713. [CrossRef] [PubMed]
52. Horton, N.J.; Switzer, S.S. Statistical Methods in the Journal. *N. Engl. J. Med.* **2005**, *353*, 1977–1979. [CrossRef] [PubMed]
53. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. [CrossRef]
54. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [CrossRef]
55. DeLong, E.R.; DeLong, D.M.; Clarke-Pearson, D.L. Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics* **1988**, *44*, 837–845. [CrossRef] [PubMed]
56. McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **1947**, *12*, 153–157. [CrossRef] [PubMed]
57. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B (Methodol.)* **1995**, *57*, 289–300. [CrossRef]
58. Van Calster, B.; McLernon, D.J.; van Smeden, M.; Wynants, L.; Steyerberg, E.W. Calibration: The Achilles heel of predictive analytics. *BMC Med.* **2019**, *17*, 230. [CrossRef] [PubMed]
59. Van Calster, B.; Nieboer, D.; Vergouwe, Y.; De Cock, B.; Pencina, M.J.; Steyerberg, E.W. A calibration hierarchy for risk models was defined: From utopia to empirical data. *J. Clin. Epidemiol.* **2016**, *74*, 167–176. [CrossRef] [PubMed]
60. Huang, Y.; Li, W.; Macheret, F.; Gabriel, R.A.; Ohno–Machado, L. A tutorial on calibration measurements and calibration models for clinical prediction models. *J. Am. Med. Inform. Assoc.* **2020**, *27*, 621–633. [CrossRef] [PubMed]

61. Kull, M.; Filho, T.S.; Flach, P. Beta calibration: A well-founded and easily implemented improvement on logistic calibration for binary classifiers. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, Fort Lauderdale, FL, USA, 20–22 April 2017; Volume 54, pp. 623–631.
62. Ashour, A.F.; Fouda, M.M.; Fadlullah, Z.M.; Ibrahem, M.I. Optimized Neural Networks for Diabetes Classification Using Pima Indians Diabetes Database. In Proceedings of the 2024 IEEE 3rd International Conference on Computing and Machine Intelligence (ICMI), Mt Pleasant, MI, USA, 13–14 April 2024; pp. 1–7. [CrossRef]
63. Battineni, G.; Sagaro, G.G.; Nalini, C.; Amenta, F.; Tayebati, S.K. Comparative Machine-Learning Approach: A Follow-Up Study on Type 2 Diabetes Predictions by Cross-Validation Methods. *Machines* **2019**, *7*, 74. [CrossRef]
64. Bhoi, S.K.; Panda, S.K.; Jena, K.K.; Abhisekh, P.A.; Sahoo, K.S.; Sama, N.U.; Pradhan, S.S.; Sahoo, R.R. Prediction of diabetes in females of pima Indian heritage: A complete supervised learning approach. *Turk. J. Comput. Math. Educ.* **2021**, *12*, 3074–3084.
65. Pranto, B.; Mehnaz, S.M.; Mahid, E.B.; Sadman, I.M.; Rahman, A.; Momen, S. Evaluating Machine Learning Methods for Predicting Diabetes among Female Patients in Bangladesh. *Information* **2020**, *11*, 374. [CrossRef]
66. Reza, M.; Hafsha, U.; Amin, R.; Yasmin, R.; Ruhi, S. Improving SVM performance for type II diabetes prediction with an improved non-linear kernel: Insights from the PIMA dataset. *Comput. Methods Programs Biomed. Update* **2023**, *4*, 100118. [CrossRef]
67. Naji, M.A.; Filali, S.E.; Aarika, K.; Benlahmar, E.H.; Abdelouhahid, R.A.; Debauche, O. Machine learning algorithms for breast cancer prediction and diagnosis. *Procedia Comput. Sci.* **2021**, *191*, 487–492. [CrossRef]
68. Jakhar, A.K.; Gupta, A.; Singh, M. SELF: A stacked-based ensemble learning framework for breast cancer classification. *Evol. Intell.* **2023**, *17*, 1341–1356. [CrossRef]
69. Shah, A.A.; Malik, H.A.M.; Mohammad, A.; Khan, Y.D.; Alourani, A. Machine learning techniques for identification of carcinogenic mutations, which cause breast adenocarcinoma. *Sci. Rep.* **2022**, *12*, 11738. [CrossRef] [PubMed]
70. Rovshenov, A.; Peker, S. Performance Comparison of Different Machine Learning Techniques for Early Prediction of Breast Cancer using Wisconsin Breast Cancer Dataset. In Proceedings of the 2022 3rd International Informatics and Software Engineering Conference (IISEC), Ankara, Turkey, 15–16 December 2022; pp. 1–6. [CrossRef]
71. Azar, A.T.; El-Said, S.A. Probabilistic neural network for breast cancer classification. *Neural Comput. Appl.* **2013**, *23*, 1737–1751. [CrossRef]
72. Muhammad, G.; Naveed, S.; Nadeem, L.; Mahmood, T.; Khan, A.R.; Amin, Y.; Bahaj, S.A.O. Enhancing Prognosis Accuracy for Ischemic Cardiovascular Disease Using K Nearest Neighbor Algorithm: A Robust Approach. *IEEE Access* **2023**, *11*, 97879–97895. [CrossRef]
73. Kumar Mygapula, S.P.; Lal Saini, M.; Raj Dheeraj, C.S. Performance Evaluation of Machine Learning Algorithms for Prediction of Cardiac Failure. In Proceedings of the 2023 International Conference on Sustainable Communication Networks and Application (ICSCNA), Theni, India, 15–17 November 2023; pp. 1599–1604. [CrossRef]
74. Liu, J.; Dong, X.; Zhao, H.; Tian, Y. Predictive Classifier for Cardiovascular Disease Based on Stacking Model Fusion. *Processes* **2022**, *10*, 749. [CrossRef]
75. Patidar, S.; Jain, A.; Gupta, A. Comparative Analysis of Machine Learning Algorithms for Heart Disease Predictions. In Proceedings of the 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 25–27 May 2022; pp. 1340–1344. [CrossRef]
76. Habib, M.; Aljarah, I.; Faris, H.; Mirjalili, S. Multi-objective particle swarm optimization: Theory, literature review, and application in feature selection for medical diagnosis. In *Evolutionary Machine Learning Techniques: Algorithms and Applications*; Mirjalili, S., Faris, H., Aljarah, I., Eds.; Springer: Singapore, 2020; pp. 175–201. [CrossRef]
77. Remeseiro, B.; Bolon-Canedo, V. A review of feature selection methods in medical applications. *Comput. Biol. Med.* **2019**, *112*, 103375. [CrossRef] [PubMed]
78. Oh, I.S.; Lee, J.S.; Moon, B.R. Hybrid genetic algorithms for feature selection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2004**, *26*, 1424–1437. [CrossRef] [PubMed]
79. Pudjihartono, N.; Fadason, T.; Kempa-Liehr, A.W.; O'Sullivan, J.M. A review of feature selection methods for machine learning-based disease risk prediction. *Front. Bioinform.* **2022**, *2*, 927312. [CrossRef] [PubMed]
80. Tasin, I.; Nabil, T.U.; Islam, S.; Khan, R. Diabetes prediction using machine learning and explainable AI techniques. *Healthc. Technol. Lett.* **2023**, *10*, 1–10. [CrossRef] [PubMed]
81. Chang, V.; Bailey, J.; Xu, Q.A.; Sun, Z. Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms. *Neural Comput. Appl.* **2022**, *35*, 16157–16173. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.