

Article

The Point-of-Balanced Performance in Binary Classification: How to Simplify the Performance Metrics Ecosystem

Ioannis Markoulidakis ^{1,*} and Georgios Markoulidakis ²¹ Department of Research and Innovation, Public Power Company, 10432 Athens, Greece² Department of Electrical and Computer Engineering, National Technical University of Athens, 15773 Athens, Greece; el21050@mail.ntua.gr

* Correspondence: i.markoulidakis@ppcgroup.com; Tel.: +30-6944-600646

Abstract

Binary classification is one of the most common supervised machine-learning problems. Several metrics have been defined in the literature to assess the performance of binary classification machine-learning models. However, using different metrics to compare two or more models may yield different results, often prompting comparative studies on the best metric for performance analysis. The current paper addresses this topic by developing a theoretical framework, which is validated through examples of real-world binary classification problems. As a first step, the paper defines the concept of equivalent metrics and identifies all pairs of State-of-the-Art metrics that yield the same conclusion when two classifiers are compared. The paper then identifies a specific classification threshold, called the “Point-of-Balanced Performance” (PoBP), for which the entire set of State-of-the-Art performance metrics yields consistent results when comparing classifiers. The paper also identifies the geometrical representation of the PoBP in the Receiver Operating Characteristic curve. Although identifying the PoBP during the training phase is trivial, this is not the case for inference. The paper defines and compares various approximation methods for identifying the PoBP during inference. The results of the analysis are then applied to real-world examples, indicating that the PoBP can become the preferred approach without excluding the option of selecting a State-of-the-Art approach depending on the specific problem characteristics. Overall, the paper provides useful theoretical insights and new tools for approaching binary classification analysis.

Keywords: binary classification; machine learning; confusion matrix; performance metrics; class probabilities



Academic Editors: Chun Yin, Jiusi Zhang, Quan Qian and Tenglong Huang

Received: 26 January 2026

Revised: 19 February 2026

Accepted: 21 February 2026

Published: 25 February 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Binary classification is a fundamental task in supervised machine learning [1]. It involves categorizing data points into one of two distinct classes based on a set of input features. Typically, the two classes are named “Positive” and “Negative” classes with appropriate meaning in each specific problem. Binary classification is prevalent across various applications, ranging from healthcare and finance to image recognition. Due to its basic nature, binary classification has attracted research interest in recent decades, leading to significant progress in the field of Machine Learning (ML).

One direction of research refers to the efficiency of binary classifiers, leading to the definition of a wide variety of binary classification algorithms indicatively including: Logistic Regression (LogReg), Naïve Bayes (NB), Support Vector Machines (SVM), Decision Trees

(DT), Random Forest (RF), k-Nearest Neighbors (k-NN), Convolutional Neural Networks (CNN), Artificial Neural Networks (ANN), Extreme Gradient Boosting (XGB), etc. [2,3]. In another line of research, the performance of binary classifiers has been analyzed using appropriate metrics. The performance metrics of a binary classification model are typically calculated based on the so-called Confusion Matrix (CM) [4,5]. A quite high number of performance metrics have been proposed in the literature, each providing certain performance insights. Such metrics include Accuracy, Precision, Recall, the F1-Score, or the Matthews Correlation Coefficient (MCC) [6].

A typical study of a binary classification problem involves comparing different binary classifiers to identify the most efficient one. Such a process is based on selecting one or more performance metrics. The wide variety of performance metrics available for binary classification can often lead to conflicting conclusions about the best-performing model. To address this issue, data scientists must select the metric or set of metrics to consider for the scope of their study, taking into account the pros and cons of each. For example, it is well known that Accuracy (the ratio of correct predictions) is not representative of a binary classifier's performance on datasets with highly imbalanced class distributions [7–9].

The paper addresses the problem of the wide variety of binary classification performance metrics. The literature in principle follows an approach of identifying the best performance metric [7–29]. A different approach is followed in this paper, starting with the definition of equivalent performance metrics that correspond to the case where two metrics lead to the same conclusion when the performance of a pair of classifiers is considered. The paper then defines the PoBP as the point at which the number of predicted positives equals the number of actual positives. It is shown that at this specific point, all performance metrics become equivalent, eliminating the issue of which metric is most appropriate for the problem analysis. The paper then considers the geometrical representation of the PoBP on the Receiver Operating Characteristic (ROC) curve, which is another common State-of-the-Art method for performance analysis. The problem of PoBP identification during inference is also addressed through a number of approximation methods. Finally, the method for exploiting the PoBP during the analysis of real binary classification problems is illustrated. Overall, the paper aims to provide key insights into what drives the observed differences in performance metrics and to propose a new method for analyzing the performance of binary classifiers.

The paper is organized as follows: Section 2 provides the modeling framework of binary classification. Section 3 provides State-of-the-Art performance metrics for binary classification and demonstrates that different metrics may lead to different conclusions about the best-performing classifier. Section 4 defines the concept of Equivalent Performance Metrics and shows that there are several pairs of metrics with this property. Section 5 introduces the concept of the PoBP and shows that, at the PoBP, all performance metrics become equivalent. Section 6 provides the geometrical representation of the PoBP in the ROC curve space. Section 7 provides alternative methods for identifying the PoBP during inference. Section 8 provides specific examples that compare PoBP performance vs. alternative points of performance. Section 9 provides the paper's conclusions and the next research steps. The entire analysis is supported by example results from real-world binary classification problems and State-of-the-Art classifiers listed in Appendix A.

2. Modeling of the Binary Classification Problem

For a specific input dataset X with the corresponding set of class labels Y , we define the following parameters: S as the total number of input samples, P as the number of samples with a “positive” class label, and N as the number of samples with a “negative” class label. A trained binary classifier predicts the class label of each sample, leading to the set of

predicted labels, $\hat{Y}$. We define PP as the number of samples with a positive predicted label and PN as the number of samples with a negative predicted label. The Confusion Matrix (CM) of a binary classification problem is provided in Table 1 [5]. The elements of the CM correspond to the number of the dataset samples for which a specific combination of actual and predicted class occurs: TP (True Positives): both the predicted and the actual class is positive, TN (True Negatives): both the predicted and the actual class is negative, FP (False Positives): predicted class positive and actual class negative and FN (False Negatives): predicted class negative and actual class positive. Table 1 also depicts the relation between the CM elements and P, N, PP, PN, and S.

Table 1. The Confusion Matrix of a binary classification problem.

	Actual Positive	Actual Negative	Total
Predicted Positive	TP	FP	TP + FP = PP
Predicted Negative	FN	TN	FN + TN = PN
Total	TP + FN = P	FP + TN = N	P + N = PP + PN = S

For a wide variety of binary classifiers, it is possible to identify the class probabilities corresponding to the conditional probability of the class label having a specific value (positive or negative) given a certain input sample. In this paper, we consider the following process for predicting class $\hat{y}_i$ of the input sample i ($i = 1, 2, \dots, S$) based on a trained model [23]: (a) estimation of the class probabilities ($\hat{p}_p(i)$ positive and $\hat{p}_n(i) = 1 - \hat{p}_p(i)$ negative class probabilities) and (b) application of a classification threshold θ to predict the class label of each input sample based on the following condition:

$$\hat{y}_i = \begin{cases} \text{Positive} & \text{if } \hat{p}_p(i) \geq \theta \\ \text{Negative} & \text{if } \hat{p}_p(i) < \theta \end{cases} \quad (1)$$

The most common classification threshold is $\theta = 0.5$, with the predicted class corresponding to the maximum class probability for each input sample. However, depending on the specific classification problem and the performance metric in focus, a different classification threshold may be adopted.

3. Performance Metrics for Binary Classification

The performance analysis of binary classifiers has, over time, led to the definition of a wide range of metrics, as shown in Table 2 [7–27]. All listed metrics in Table 2 are defined under the assumption that their denominators are nonzero. According to the literature, each metric captures certain insights into the performance of a binary classifier. Moreover, each metric is characterized by certain pros and cons leading to the need of a proper metric selection in the study of a specific problem. In particular, refs. [7–12] considers the advantages of MCC over a number of different metrics, while [13] highlights the fact that MCC is not a balanced measurement. In [14,15], the issue of Accuracy is highlighted, especially in the presence of prevalence variance. In [16], the issues related to Cohen’s kappa metric are presented. In [17], it is recommended to move from analysis based on Precision, Recall, and F-measure metrics to ROC, Informedness, Markedness, and correlation analysis. In [18], the Efficiency Index is introduced as a new proposed metric. The fact that appropriate instance selection improves geometric mean Accuracy for imbalanced input datasets is addressed in [19]. The benefits of the G4 metric for the medical device validation problem are illustrated in [20]. The classification threshold for maximizing F1-score is provided in [21]. In [22], an overview of binary classification performance metrics is provided, while [23]

provides a unified view of performance metrics, highlighting the relationship between the classification threshold and the expected classification loss. In [24–26], the issues related to F-score (F1-score, $F\beta$ -score) are highlighted, proposing alternative metrics such as F^* (Jaccard Index) or $F1^*$ (arithmetic mean of F1-score and Accuracy). A comparison of 17 metrics and their validity across different binary classification problem datasets is provided in [27]. Finally, refs. [28,29] provide alternative methods (meta-metrics) to measure the efficiency of binary classification performance metrics.

Table 2. An indicative list of Binary Classification Performance Metrics.

A/A	Metric Name	Symbol and Definition
1	Accuracy	$Acc = (TP + TN) / S$
2	Fraction of Inefficiency	$FiC = (FP + FN) / S$
3	True Positive Rate (Recall, Sensitivity)	$TPR = TP / P$
4	True Negative Rate (Specificity)	$TNR = TN / N$
5	Balanced Accuracy	$BA = (TPR + TNR) / 2$
6	False Positive Rate (Fall-Out Rate)	$FPR = FP / N$
7	False Negative Rate (Miss Rate)	$FNR = FN / P$
8	Positive Predictive Value (Precision)	$PPV = TP / PP$
9	Negative Predictive Value	$NPV = TN / PN$
10	False Discovery Rate	$FDR = FP / PP$
11	False Omission Rate	$FOR = FN / PN$
12	F1-Score	$F_1 = 2 \cdot TPR \cdot PPV / (TPR + PPV)$
13	$F\beta$ -Score	$F_\beta = (1 + \beta^2) \cdot TPR \cdot PPV / (\beta^2 \cdot TPR + PPV)$
14	Cohen's kappa (see Appendix B)	$k = 2 \cdot (TPR + TNR - 1) / (PP/P + PN/N)$
15	Youden's J Statistic (Informedness)	$BM = TPR + TNR - 1$
16	Markedness	$MK = PPV + NPV - 1$
17	Matthews Correlation Coefficient (see Appendix B)	$MCC = (TPR + TNR - 1) \sqrt{(P \cdot N) / (PP \cdot PN)}$
18	Fowlkes–Mallows Index	$FM = \sqrt{TPR \cdot PPV}$
19	Jaccard Index (Threat Score)	$TS = TP / (TP + FP + FN)$
20	Prevalence Threshold	$PT = (\sqrt{TPR \cdot FPR} - FPR) / (TPR - FPR)$
21	P4 metric	$P4 = 1 / (1/TPR + 1/PPV + 1/TNR + 1/NPV)$
22	Positive Likelihood Ratio	$LR+ = TPR / FPR$
23	Negative Likelihood Ratio	$LR- = FNR / TNR$
24	Diagnostics Odd Ratio	$DOR = LR+ / LR-$
25	Efficiency Index	$EI = (TP + TN) / (FP + FN)$
26	Inefficiency Index	$InI = (FP + FN) / (TP + TN)$
27	G-Mean Score	$GM = \sqrt{TPR \cdot TNR}$
28	G4 Score	$G4 = \sqrt[4]{TPR \cdot PPV \cdot TNR \cdot NPV}$

Table 3 provides the performance metrics for a set of binary classifiers (see Appendix A: ANN, LogReg, RF, DT, NB, CNN, and XGB) applied to Problem 1 (see Appendix A) on a test dataset of size $S = 3295$ samples and a classification threshold of $\theta = 0.5$. The table

illustrates that comparing two or more models may yield different conclusions, depending on the performance metric used.

Table 3. Metric values vs. the applied classifier for Problem 1 (best-performing classifier per metric marked with bold).

	ANN	LogReg	RF	DT	NB	CNN	XGB
Accuracy	0.916	0.906	0.904	0.905	0.905	0.913	0.918
Fraction of Inefficiency	0.084	0.094	0.096	0.095	0.095	0.087	0.082
True Positive Rate (Recall, Sensitivity)	0.569	0.377	0.276	0.536	0.344	0.429	0.561
True Negative Rate (Specificity)	0.959	0.978	0.982	0.953	0.976	0.974	0.964
Balanced Accuracy	0.764	0.678	0.629	0.745	0.660	0.701	0.762
False Positive Rate (Fall-Out Rate)	0.041	0.022	0.018	0.047	0.024	0.026	0.036
False Negative Rate (Miss Rate)	0.431	0.623	0.724	0.464	0.656	0.571	0.439
Positive Predictive Value (Precision)	0.638	0.700	0.654	0.602	0.645	0.678	0.669
Negative Predictive Value	0.946	0.920	0.917	0.940	0.922	0.931	0.944
False Discovery Rate	0.362	0.300	0.346	0.398	0.355	0.322	0.331
False Omission Rate	0.054	0.080	0.083	0.060	0.078	0.069	0.056
F1-Score	0.602	0.490	0.388	0.567	0.449	0.525	0.610
F β -Score ($\beta = 2$)	0.623	0.597	0.513	0.588	0.549	0.607	0.644
Cohen's kappa	0.555	0.443	0.345	0.514	0.402	0.480	0.564
Informedness	0.528	0.355	0.258	0.490	0.320	0.403	0.525
Markedness	0.585	0.620	0.570	0.542	0.567	0.608	0.613
Mathews' Correlation Coefficient	0.556	0.469	0.384	0.515	0.426	0.495	0.567
Fowlkes–Mallows Index	0.603	0.514	0.425	0.568	0.471	0.539	0.612
Jaccard Index (Threat Score)	0.430	0.325	0.241	0.396	0.289	0.356	0.439
Prevalence Threshold	0.211	0.195	0.204	0.228	0.209	0.197	0.202
P4 metric	0.184	0.162	0.138	0.177	0.152	0.169	0.186
Positive Likelihood Ratio	13.998	17.098	15.292	11.487	14.391	16.559	15.586
Negative Likelihood Ratio	0.449	0.637	0.737	0.486	0.672	0.586	0.456
Diagnostics Odd Ratio	31.165	26.849	20.748	23.623	21.419	28.235	34.214
Efficiency Index	10.856	9.632	9.463	9.497	9.564	10.450	11.162
Inefficiency Index	0.092	0.104	0.106	0.105	0.105	0.096	0.090
G-Mean Score	0.739	0.607	0.521	0.715	0.580	0.646	0.735
G4 Score	0.758	0.698	0.635	0.734	0.668	0.716	0.764

4. The Concept of Equivalent Performance Metrics

In this section, we introduce the concept of equivalent performance metrics as a first step towards understanding how the relation between a pair of metrics can influence the performance analysis of different classifiers. From the definition of the performance metrics (see Table 2), it can be observed that there are pairs of closely related metrics expressing effectively the same performance aspects of a binary classifier, e.g., TPR can be considered to be complementary to the FNR metric since $TPR = 1 - FNR$. Let us assume that TPR is used for the comparison of the performance of two classifiers, f_1 and f_2 , applied to the same dataset and for a common classification threshold θ . If f_1 outperforms f_2 based on TPR (i.e., $TPR(f_1, \theta) > TPR(f_2, \theta)$) then we already know the result of the comparison for the FNR metric since it will definitely be $FNR(f_1, \theta) < FNR(f_2, \theta)$. To formalize this property, we adopt the following definition:

Definition of Equivalent Performance Metrics: Let us consider a pair of trained binary classification models f_1, f_2 , which are applied over the same binary classification problem for the same input dataset X and class labels Y , with classification thresholds θ_1 and θ_2 respectively. Let us also consider two performance metrics PM_1 and PM_2 which are used to compare the performance of the two models. The metrics PM_1, PM_2 are called “Equivalent” if they provide a common model performance comparison conclusion. If for both metrics a better model performance occurs as their values either increase or decrease, then their equivalence corresponds to the following property:

$$PM_1(f_1, \theta_1) \geq PM_1(f_2, \theta_2) \Leftrightarrow PM_2(f_1, \theta_1) \geq PM_2(f_2, \theta_2) \quad (2)$$

In the case that for one of the metrics a better model performance occurs as its value increases, while for the other metric a better model performance occurs as its value decreases, then their equivalence corresponds to the following property:

$$PM_1(f_1, \theta_1) \geq PM_1(f_2, \theta_2) \Leftrightarrow PM_2(f_1, \theta_1) \leq PM_2(f_2, \theta_2) \quad (3)$$

It should be noted that the above definition of equivalence may also apply to classifiers like k-NN, even if the definition of the classification threshold is not applicable. The equivalence property apparently has a transitive property; if PM_1 is equivalent to PM_2 and PM_2 is equivalent to PM_3 , it is trivial to show that PM_1 is equivalent to PM_3 . Moreover, if two metrics are equal, they are also equivalent.

Table 4 provides the resulting pairs of equivalent metrics from Table 2, based on the equation that describes their relationship. Appendix C provides the proof for the non-trivial cases. Note that, as expected, the metrics grouped as equivalent in Table 4 yield the same classifier performance comparison results as in the example in Table 3. The groups of equivalent metrics presented in Table 4 allow narrowing the list of Table 2 metrics to consider for comparing different classifiers (e.g., comparing the performance of a pair of classifiers based on Informedness eliminates the need to compare their performance based on their equivalent metric, Balanced Accuracy).

Table 4. The list of pairs of equivalent metrics.

Groups of Equivalent Metrics	Equations
Accuracy (Acc)	
Fraction of Inefficiency (FiC)	$Acc = 1 - FiC$
Efficiency Index (EI)	$EI = 1/FiC - 1$
Inefficiency Index (InI)	$EI = 1/InI$

Table 4. Cont.

Groups of Equivalent Metrics	Equations
Balanced Accuracy (BA) Informedness (BM)	$BA = (BM + 1)/2$
Recall (TPR) Miss Rate (FNR)	$TPR = 1 - FNR$
Specificity (TNR) Fall Out Rate (FPR)	$TNR = 1 - FPR$
Precision (PPV) False Discovery Rate (FDR)	$PPV = 1 - FDR$
Negative Predictive Value (NPV) False Omission Rate (FOR)	$NPV = 1 - FOR$
F1-Score (F1) Jaccard Index (TS)	$TS = F_1 / (2 - F_1)$
Prevalence Threshold (PT) Positive Likelihood Ratio (LR+)	$PT = 1 / (1 + \sqrt{LR+})$

5. The Point-of-Balanced Performance (PoBP)

Selecting a classification threshold affects a classifier's performance metrics. Depending on the metric, we may define characteristic performance points, for example, the point at which Accuracy, F1-score, or MCC score is maximized. In this paper, we define a new characteristic point, the PoBP, based on a property that characterizes an efficient classifier. As will be shown later in this paper, PoBP exhibits properties that allow for further simplification of the performance analysis of different classifiers.

PoBP Definition: Let us assume a binary classification problem and a trained binary classification model f . The PoBP is defined as the selection of a proper classification threshold θ_b for which the trained model produces a number of predicted positives equal to the number of actual positives for a given input dataset X and the corresponding set of actual class labels Y :

$$PP(f, \theta_b) = P \quad (4)$$

This property leads to a number of predicted negatives equal to the actual negatives:

$$PN(f, \theta_b) = S - PP(f, \theta_b) = S - P = N \quad (5)$$

At the PoBP, the CM is symmetric as false positives become equal to false negatives:

$$FP(f, \theta_b) = PP(f, \theta_b) - TP(f, \theta_b) = P - TP(f, \theta_b) = FN(f, \theta_b) \quad (6)$$

5.1. The Classification Threshold of the Point-of-Balanced Performance

In this section, we focus on identifying and classifying the PoBP threshold to indicate that the PoBP can always be identified as long as the number of positives is known. Let us assume a trained binary classifier f applied over an input dataset X with a set of class labels Y . The model produces an indexed list $\hat{L}$ of positive class probabilities $\hat{p}_p(i)$, $i = 1, 2, \dots, S$. Let us consider the revised list $\hat{L}_d$ which occurs by sorting the list $\hat{L}$ at a decreasing order of the positive class probabilities. In this list, a classification threshold θ leads to the selection of the first $PP(\theta)$ set of samples, which, according to their ranking, justify the property: $\hat{p}_p(i) \geq \theta$, $i = 1, 2, \dots, PP(\theta)$.

Based on the definition of $\hat{L}_d$, we can prove that PoBP can always be identified as long as the number of actual positives is known: the PoBP classification threshold θ_b should be equal to the positive class probability at position P in the ordered list $\hat{L}_d$:

$$\theta_b = \hat{p}_p(i = P) \quad (7)$$

There is, however, the possibility of multiple input samples with equal positive class probabilities. This is quite common when class probabilities have been calibrated. In this case, Equation (7) does not lead to a unique classification threshold value. This issue does not limit the capability to identify the PoBP. This is possible by simply assigning a positive predicted class to the first P samples in the list $\hat{L}_d$. Meanwhile, a negative one is assigned for the rest. This process can be expressed by a revised definition of the PoBP classification threshold:

$$\hat{y}_i = \begin{cases} \text{Positive} & \text{if } \hat{p}_p(i) \geq \theta_b \text{ and } i \leq P \\ \text{Negative} & \text{if } \hat{p}_p(i) < \theta_b \text{ or } i > P \end{cases}, i = 1, 2, \dots, S \quad (8)$$

As discussed later, this condition can be applied when the number of positives is known (during model training) or approximated (during inference).

5.2. Performance Metrics That Become Equal at the Point-of-Balanced Performance

As a first step in the analysis of PoBP properties, we consider the set of metrics that become equal (and consequently equivalent) at the PoBP. In particular, the following equalities of metrics apply at the PoBP (see overview in Table 5):

Table 5. Equalities of Binary Classification Metrics at the Point-of-Balanced Performance.

Groups of Metrics	Equality at the Point-of-Balanced Performance
True Positive Rate (TPR) Precision (PPV) F1-score (F1) Fβ-score (Fβ) Fowlkes–Mallows Index (FM)	$TPR(f, \theta_b) = PPV(f, \theta_b) = F_1(f, \theta_b) = F_\beta(f, \theta_b) = FM(f, \theta_b)$
Informedness (BM) Markedness (MK) Matthews Correlation Coefficient (MCC) Cohen's kappa (k)	$BM(f, \theta_b) = MK(f, \theta_b) = MCC(f, \theta_b) = k(f, \theta_b)$
True Negative Rate (TNR) Negative Predictive Value (NPV)	$TNR(f, \theta_b) = NPV(f, \theta_b)$
G-Mean Score G4 Score	$GM(f, \theta_b) = G4(f, \theta_b)$

Precision becomes equal to Recall:

$$TPR(f, \theta_b) = \frac{TP(f, \theta_b)}{P} = \frac{TP(f, \theta_b)}{PP(f, \theta_b)} = PPV(f, \theta_b) \quad (9)$$

The Miss Rate becomes equal to the False Discovery Rate:

$$FNR(f, \theta_b) = \frac{FN(f, \theta_b)}{P} = \frac{FP(f, \theta_b)}{PP(f, \theta_b)} = FDR(f, \theta_b) \quad (10)$$

The Negative Predictive Value becomes equal to the True Negative Value:

$$NPV(f, \theta_b) = \frac{TN(f, \theta_b)}{PN(f, \theta_b)} = \frac{TN(f, \theta_b)}{N} = TNR(f, \theta_b) \quad (11)$$

The False Positive Rate becomes equal to the False Omission Rate:

$$FPR(f, \theta_b) = \frac{FP(f, \theta_b)}{N} = \frac{FN(f, \theta_b)}{PN(f, \theta_b)} = FOR(f, \theta_b) \quad (12)$$

The F_β -score, based on the fact that Precision equals Recall at the PoBP, becomes equal to Recall:

$$F_\beta(f, \theta_b) = (1 + \beta^2) \frac{TPR^2(f, \theta_b)}{(1 + \beta^2)TPR(f, \theta_b)} = TPR(f, \theta_b) \quad (13)$$

Therefore, at the PoBP, F1-score will be equal to F_β -score for any factor β , and both will be equal to Precision and Recall:

$$F_\beta(f, \theta_b) = F_1(f, \theta_b) = PPV(f, \theta_b) = TPR(f, \theta_b) \quad (14)$$

Fowlkes–Mallows Index becomes equal to Precision and the metrics of Equation (14):

$$FM(f, \theta_b) = \sqrt{TPR(f, \theta_b) \cdot PPV(f, \theta_b)} = TPR(f, \theta_b) \quad (15)$$

Informedness (BM) becomes equal to Markedness (MK) based on Equations (9) and (11):

$$BM(f, \theta_b) = PPV(f, \theta_b) + NPV(f, \theta_b) - 1 = MK(f, \theta_b) \quad (16)$$

Matthews Correlation Coefficient as well as Cohen's kappa k (see Appendix B) both become equal to Informedness at the PoBP based on the following equations:

$$MCC(f, \theta_b) = BM(f, \theta_b) \cdot \sqrt{\frac{P \cdot N}{PP(f, \theta_b) \cdot PN(f, \theta_b)}} = BM(f, \theta_b) \quad (17)$$

$$k(f, \theta_b) = \frac{2 \cdot BM(f, \theta_b)}{PP(f, \theta_b)/P + PN(f, \theta_b)/N} = BM(f, \theta_b) \quad (18)$$

G4 score becomes equal to the G-mean score based on Equations (9) and (11):

$$G4(f, \theta_b) = \sqrt[4]{TPR(f, \theta_b) \cdot PPV(f, \theta_b) \cdot TNR(f, \theta_b) \cdot NPV(f, \theta_b)} \\ = \sqrt{TPR(f, \theta_b) \cdot TNR(f, \theta_b)} = GM(f, \theta_b) \quad (19)$$

To verify these properties, we consider the application of an ANN to a test dataset of size $S = 3295$ for Problem 1 (see Appendix A), with $P = 369$ actual positives. For this example, Figure 1 clearly shows that TPR, PPV, F1-score, F2-score (F_β -score for $\beta = 2$), and FM are equal at the PoBP (i.e., the point at which PP equals P). Moreover, Figure 2 illustrates that MCC, BM, MK, and Cohen's kappa become equal at the PoBP. Both figures present the metrics vs. the number of PP (ranging from $PP = 0$ to $PP = S$) in order to clearly present the conditions that apply at the PoBP (i.e., $PP = P$). This is an equivalent way of presenting the metrics vs. the classification threshold θ varying from 1 to 0 (since for $\theta = 1$, $PP = 0$ and for $\theta = 0$, $PP = S$).

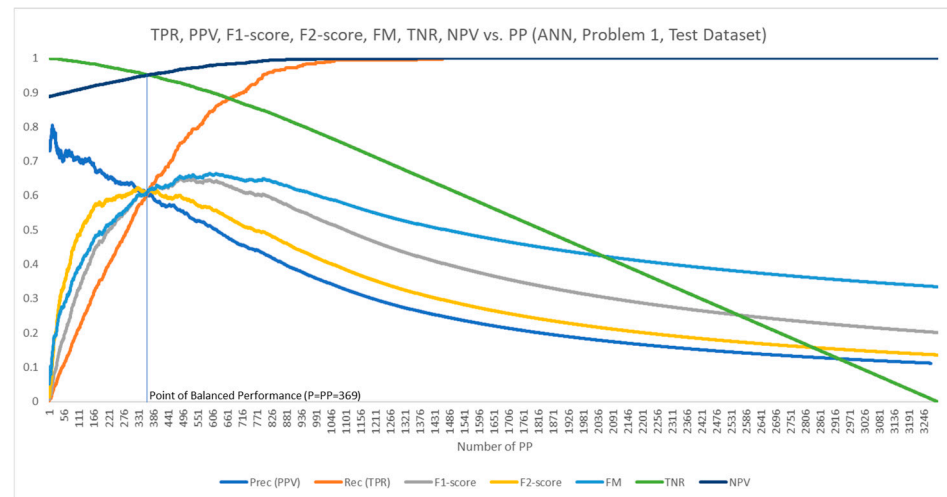


Figure 1. Metric equalities at the PoBP ($P = PP = 369$): TPR, PPV, F1-score, F2-score, FM, TNR, NPV vs. the number of PP for the ANN classifier, Problem 1, Test dataset ($S = 3295$).

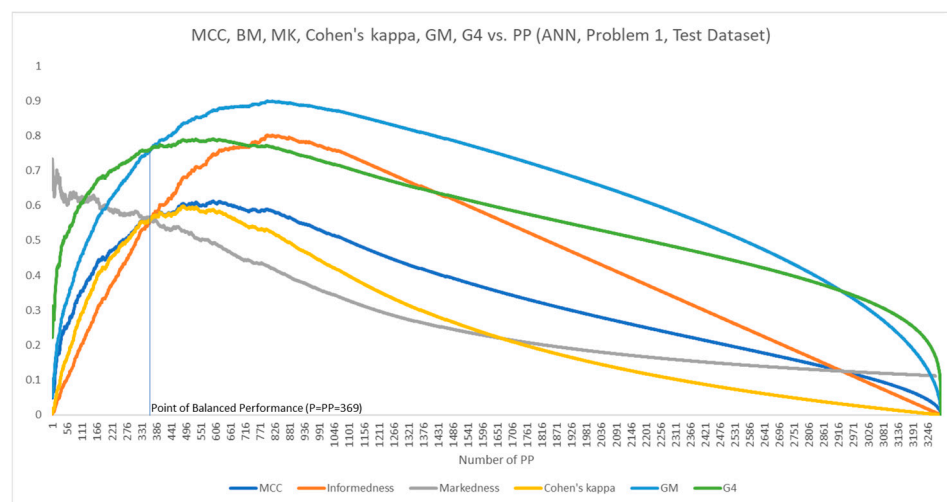


Figure 2. Metric equalities at the PoBP ($P = PP = 369$): MCC, BM, MK, Cohen's kappa, GM, G4 vs. the number of PP for ANN Classifier, Problem 1, Test dataset ($S = 3295$).

Figure 2 illustrates that MCC, BM, MK, GM, and G4 are equal at a second point, corresponding to the case where PP equals the number of actual negatives (i.e., $PP = N = 2926$ in the example dataset). Unlike PoBP, this point has no specific interest from the classifier performance viewpoint.

Both figures indicate that none of the presented metrics is maximized at the PoBP. However, PoBP provides a tradeoff between pairs of metrics with opposite trends as follows: (a) TPR vs. PPV, (b) TNR vs. NPV, and (c) MCC or BM (Informedness) vs. MK (Markedness).

5.3. Performance Metrics That Become Equivalent at the Point-of-Balanced Performance

In this section, in addition to the metrics that are equal at the PoBP (Table 5), we further show that all binary classification performance metrics are equivalent. This unique property of the PoBP makes it an interesting method for comparing the performance of different classifiers. Let us prove the following theorem for any pair of performance metrics:

Theorem 1. *If two performance metrics PM_1 , PM_2 indicate an improved binary classification model performance as the number of True Predictions increases (TP and/or TN) and/or the number of False Predictions decreases (FP and/or FN), then the metrics will be equivalent at the PoBP.*

Proof of Theorem 1. Let us consider two classifiers, f_1 and f_2 , applied to the same dataset for a binary classification problem. As already shown, for each classifier, the PoBP exists and, in general, corresponds to a different PoBP threshold (i.e., θ_{b1} for f_1 and θ_{b2} for f_2). Let us also consider a performance metric PM_1 which is defined based on any set of parameters of the Confusion Matrix. The metric PM_1 is defined such that it corresponds to an improved model performance as the number of true predictions (i.e., TP and/or TN) increases and/or the number of false predictions (i.e., FP and/or FN) decreases. This property applies to all metrics presented in Table 2 (e.g., TPR increases with the number of TP, indicating better model performance, while FPR decreases as the number of FP decreases, indicating better model performance).

Apart from the Confusion Matrix elements (TP , FP , FN , TN), the metric PM_1 (Table 2) may also depend on parameters like P , N , PP , PN , S . At the PoBP these parameters become constant for the same input dataset as $PP = P$ and $PN = N$.

Without loss of generality, let us assume that the higher the PM_1 value, the better the model performance, and at the PoBP, the following inequality holds:

$$PM_1(f_1, \theta_{b1}) \geq PM_1(f_2, \theta_{b2}) \quad (20)$$

Under this assumption and based on the definition of PM_1 , at least one of the parameters, TP, FP, FN, and TN, will be better for the model f_1 than the corresponding ones for the model f_2 at the PoBP. The term “better” in this case means a higher number of true predictions (for TP, TN) and a lower number of false predictions (for FP, FN).

Without loss of generality, let us assume that Equation (20) applies due to the following property: $TP(f_1, \theta_{b1}) \geq TP(f_2, \theta_{b2})$. It is easy to show, based on $TP + FP = PP$, $TN + FN = PN$ as well as Equations (4)–(6), that in this case, the following properties apply at the PoBP:

$$\begin{aligned} TP(f_1, \theta_{b1}) \geq TP(f_2, \theta_{b2}) &\Leftrightarrow FP(f_1, \theta_{b1}) \leq FP(f_2, \theta_{b2}) \Leftrightarrow \\ &\Leftrightarrow FN(f_1, \theta_{b1}) \leq FN(f_2, \theta_{b2}) \Leftrightarrow TN(f_1, \theta_{b1}) \geq TN(f_2, \theta_{b2}) \end{aligned} \quad (21)$$

According to these properties, if classifier f_1 has a better performance than classifier f_2 when compared at their PoBP based on any of the parameters: TP, FP, FN, or TN, then f_1 will have a better performance than f_2 for the rest of the parameters. Therefore, if Equation (20) holds, then the properties of Equation (21) will also hold at the PoBP.

Now, let us consider another metric PM_2 which, as with PM_1 , is defined in such a way that it corresponds to a better model performance as the number of true predictions (i.e., TP and/or TN) increases and/or the number of false predictions (i.e., FP and/or FN) decreases. At the PoBP, assuming that Equation (20) and thus Equation (21) apply, it is evident that PM_2 will also indicate a better performance for the model f_1 .

In case that the higher the PM_2 value, the better the model performance, the previous analysis leads to the following condition at the PoBP:

$$PM_2(f_1, \theta_{b1}) \geq PM_2(f_2, \theta_{b2}) \quad (22)$$

Therefore, PM_1 and PM_2 will be equivalent (see Equation (2)). Note that if the lower PM_2 the value, the better the model performance, then PM_1 and PM_2 would be again equivalent based on the property of Equation (3).

The above analysis indicates that at the PoBP, any pair of performance metrics that correspond to improved model performance as the number of true predictions (TP and/or TN) increases and/or the number of false predictions (FP and/or FN) decreases will be equivalent. $\square$

Based on the above theorem, all State-of-the-Art performance metrics presented in Table 2 are expected to be either equal or equivalent at the PoBP. The analysis provided in Appendix D on the equivalence of parameters at the PoBP provides the necessary evidence for this property. Therefore, if the performance of two binary classifiers is compared at their PoBP, the comparison conclusion will be the same for any metric of Table 2. In this context, at the PoBP, we need to consider only one metric from Table 2 to compare the performance of two classifiers. Therefore, the problem identified in Table 3 is eliminated, as any pair of metrics at the PoBP yields the same conclusions when comparing the performance of binary classifiers.

Table 6 provides an overview of the resulting metrics at the PoBP for Problem 1, using the same dataset as in Table 3. Table 6 confirms the theoretical predictions regarding the metrics that become equal at the PoBP. Table 6 also confirms that all metrics are equivalent at the PoBP; as in the example presented, XGB becomes the best-performing classifier at the PoBP for all metrics.

Table 6. Performance Metric comparison at the PoBP (Problem 1). The best-performing classifier is marked in bold.

	ANN	LogReg	RF	DT	NB	CNN	XGB
Accuracy	0.9135	0.8932	0.8829	0.9035	0.8914	0.9108	0.9144
Fraction of Inefficiency	0.0865	0.11	0.12	0.10	0.11	0.089	0.0856
True Positive Rate	0.61	0.55	0.47	0.59	0.51	0.60	0.63
True Negative Rate	0.9515	0.94	0.93	0.95	0.94	0.950	0.9517
Balanced Accuracy	0.78	0.75	0.70	0.77	0.73	0.78	0.79
False Positive Rate	0.049	0.06	0.07	0.055	0.06	0.050	0.048
False Negative Rate	0.39	0.45	0.53	0.41	0.49	0.40	0.37
Positive Predictive Value	0.61	0.55	0.47	0.59	0.51	0.60	0.63
Negative Predictive Value	0.951	0.94	0.93	0.945	0.94	0.950	0.952
False Discovery Rate	0.39	0.45	0.53	0.41	0.49	0.40	0.37
False Omission Rate	0.049	0.061	0.066	0.055	0.061	0.050	0.048
F1-Score	0.61	0.55	0.47	0.59	0.51	0.60	0.63
F β -Score ($\beta = 2$)	0.61	0.55	0.47	0.59	0.51	0.60	0.63
Cohen's kappa	0.56	0.49	0.40	0.53	0.45	0.55	0.58
Informedness	0.56	0.49	0.40	0.53	0.45	0.55	0.58
Markedness	0.57	0.49	0.40	0.53	0.45	0.55	0.58
Matthews Correlation Coefficient	0.56	0.49	0.40	0.53	0.45	0.55	0.58
Fowlkes–Mallows Index	0.61	0.55	0.47	0.59	0.51	0.60	0.63
Jaccard Index	0.44	0.38	0.30	0.41	0.35	0.43	0.46
Prevalence Threshold	0.220	0.249	0.273	0.234	0.256	0.224	0.217

Table 6. Cont.

	ANN	LogReg	RF	DT	NB	CNN	XGB
P4 metric	0.186	0.174	0.156	0.181	0.166	0.185	0.189
Positive Likelihood Ratio	12.62	9.14	7.10	10.73	8.42	12.03	12.98
Negative Likelihood Ratio	0.41	0.47	0.57	0.44	0.52	0.42	0.39
Diagnostics Odd Ratio	31.00	19.27	12.44	24.50	16.30	28.86	33.10
Efficiency Index	10.56	8.36	7.54	9.36	8.21	10.21	10.69
Inefficiency Index	0.095	0.120	0.133	0.107	0.122	0.098	0.094
G-Mean Score	0.76	0.72	0.66	0.74	0.70	0.76	0.77
G4 Score	0.76	0.72	0.66	0.74	0.70	0.76	0.77

It should be noted that Tables 3 and 6 indicate that ANN performs similarly to XGB across all metrics. As the scope of the paper is to validate the properties of PoBP rather than identify the best-performing classifier for the specific problem, no further hyperparameter optimization has been performed for ANN and XGB.

6. PoBP at the Receiver Operating Characteristic Curve

In this section, we consider the PoBP in relation to the ROC curve [30]. The ROC curve is a method for analyzing the performance of binary classifiers. ROC is the curve in a two-dimensional space representing the True Positive Rate vs. False Positive Rate as the classification threshold θ ranges from 0 to 1. Beyond the ROC curve, the Area Under the Curve (AUC) is a metric used to compare different models, taking into account the entire range of classification thresholds [31].

Since the definition of the PoBP of a classification model f is based on a certain classification threshold θ_b , it is evident that PoBP corresponds to a single point on the ROC curve corresponding to the pair of coordinates $(TPR(f, \theta_b), FPR(f, \theta_b))$.

Comparing two classifiers at their PoBP and taking into account the equivalence of all metrics at the PoBP (and hence the equivalence between TPR and FPR), it is expected that the better a classifier's performance, the higher its TPR and the lower its FPR at the PoBP. Therefore, the PoBP point on the ROC curve of a more efficient classifier will be located to the left and higher than the PoBP point on the ROC curve of a less efficient classifier.

Moreover, assuming a certain binary classification problem, the following condition applies at the PoBP of a classifier:

$$TPR(f, \theta_b) = \frac{TP(f, \theta_b)}{P} = \frac{PP(f, \theta_b) - FP(f, \theta_b)}{P} = 1 - \frac{N}{P} \cdot FPR(f, \theta_b) \quad (23)$$

Geometrically, Equation (23) corresponds to a straight line in the ROC space. Therefore, the PoBP point on the ROC curve of any classifier will lie on a line we call the PoBP ROC space line. The endpoints of the PoBP ROC space line can be determined by two points: the ideal classifier with $FPR = 0$ and $TPR = 1$; the random-guess classifier for which $TPR = FPR$. Equation (23) provides: $TPR(f, \theta_b) = FPR(f, \theta_b) = P/S$ (i.e., the prevalence of the input dataset).

Figure 3 presents the ROC curves for all applied classifiers for a test set of Problem 3 (see Appendix A). The same figure provides the PoBP ROC space line and the PoBP point for each applied classifier, indicating that XGB is the best-performing classifier at the PoBP.

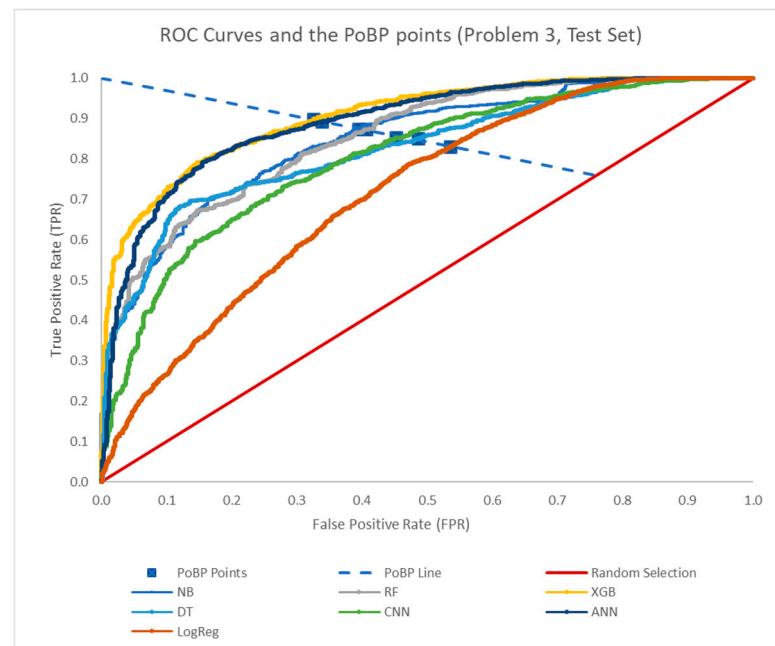


Figure 3. The ROC curves and PoBP points for the classifiers applied to Problem 3 (Test set size 2606).

7. Methods for the Identification of the Point-of-Balanced Performance

Following the analysis of the PoBP properties, we focus in this section on identifying the PoBP. Apparently, the identification of PoBP for a classifier applied to an input dataset X and a set of actual class labels Y requires determining the classification threshold for which the number of predicted positives equals the number of actual positives. During the training phase of a classifier, the available dataset is typically split into a training set and, depending on the applied method, a number of test or validation sets [32]. Since the number of actual positives is known for all of these datasets, it is a straightforward process to identify the PoBP when the trained model is applied to any of them. This section further considers the PoBP identification during inference, where the number of actual positives is unknown.

7.1. Methods for the Identification of the Point-of-Balanced Performance During Inference

During inference, the actual class labels are unknown; hence, the PoBP should be approximated. The following three alternative methods are considered:

Method 1: PoBP identification based on Training Dataset Prevalence. Assuming that the Prevalence in the inference dataset X_{inf} is identical to that of the training dataset X_{train} , we can estimate the number of positives for the inference dataset $\hat{P}(X_{inf})$ as follows:

$$\frac{P(X_{inf})}{S(X_{inf})} \approx \frac{P(X_{train})}{S(X_{train})} \Rightarrow \hat{P}(X_{inf}) \approx \frac{P(X_{train})}{S(X_{train})} \cdot S(X_{inf}) \quad (24)$$

Method 2: PoBP identification based on the PoBP Classification threshold of the Training dataset. Assuming that both the training X_{train} and the inference datasets X_{inf} follow the same distribution, we may consider that the PoBP classification threshold applied in the training dataset $\theta_b(X_{train})$ is identical to the one of the inference dataset $\theta_b(X_{inf})$. This leads directly to the approximation of the PoBP during inference:

$$\theta_b(X_{inf}) \approx \theta_b(X_{train}) \quad (25)$$

Method 3: PoBP identification based on the Actual Label Probabilistic Confusion Matrix (ALP CM) [33]. The ALP CM approximates the actual CM during inference even though the actual class labels are unknown. A good approximation holds for calibrated class probabilities (based on Platt's method [34] in this paper) and for a substantial input dataset.

This approximation can be used to identify the PoBP during inference by adjusting the decision threshold to achieve the following condition:

$$PP(X_{inf}) = \hat{P}_{ALP}(X_{inf}) \approx P(X_{inf}) \quad (26)$$

Note that Methods 2 and 3 do not rely on the assumption that the prevalence in the inference dataset is identical to that of the training dataset.

7.2. Evaluation of the PoBP Identification Methods During Inference

To evaluate the PoBP identification methods, we split the available datasets into training and test sets and use the test set to simulate the performance of the proposed methods during inference. We consider two different methods for deriving the test datasets: (a) Unbiased Class Mix method: the test datasets are produced based on random selection of samples from the available dataset in a way that they have the same prevalence as the training dataset and (b) Biased Class Mix method: the unbiased mix test datasets are further processed to produce test datasets of shorter size with a prevalence that deviates from the one of the training dataset. Table 7 provides the resulting mean absolute error for estimating the number of positives in the test dataset (based on 400 iterations of generating unbiased and biased class-mix test sets) for all problems and for one different classifier per problem (see Appendix A). It is evident that Method 1 results in a higher mean absolute error when the class mix is biased. Methods 2 and 3 appear to be performing better in the presence of bias. It should be noted that the performance of Method 3 depends on the degree to which calibrated class probabilities are in agreement with actual class frequencies.

Table 7. The mean absolute error in the estimation of the actual positives in the test set.

Problem	Classifier	Test Set	Set Size	Method 1	Method 2	Method 3
1	SVM	Unbiased	8238	0%	2.8%	1.5%
		Biased	5945	14.3%	5.6%	6.1%
2	RF	Unbiased	200	0%	2.1%	0.8%
		Biased	172	8.4%	5.8%	6.8%
3	NB	Unbiased	6513	0%	0.4%	0.1%
		Biased	5564	6.3%	3.6%	5.2%
4	XGB	Unbiased	29,092	0%	0.5%	0.3%
		Biased	21,847	6.2%	2.9%	3.5%
5	LogReg	Unbiased	5581	0%	3%	1.6%
		Biased	4383	13.2%	7.3%	7.6%

Figure 4 compares a set of performance metrics corresponding to the actual PoBP with those corresponding to the estimated PoBP using each of the above methods. The following indicative set of metrics is considered: Accuracy, Precision, Recall, F1-score, TNR, NPV, and MCC. The figure refers to Problem 4 and the XGB classifier. Figure 4a refers to an unbiased class mix test set of size 29,092 with prevalence 75.8%, and Figure 4b to a biased test set of size 17,096 with prevalence 68.9%. As shown in Table 7, all methods yield accurate PoBP approximations for unbiased input datasets. For that reason, Figure 4a shows that the

metrics for the PoBP identified using Methods 1, 2, and 3 are almost identical to those for the actual PoBP. Figure 4b indicates that, for a biased input dataset, the error in identifying the PoBP (see Table 7) is reflected in the deviation between the metrics corresponding to the estimated PoBP and those corresponding to the actual PoBP. Moreover, as expected from Table 7, Figure 4b shows that Methods 2 and 3 perform better than Method 1.

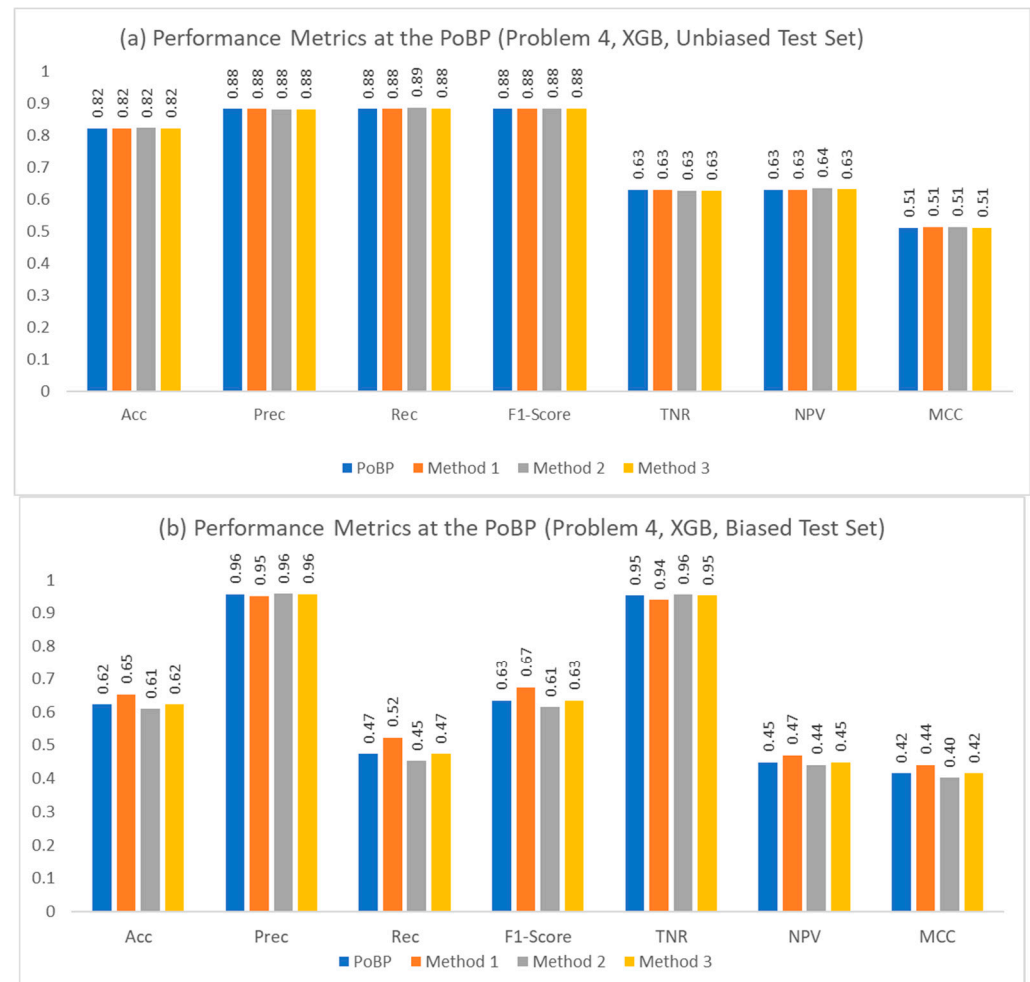


Figure 4. Performance metrics for the actual vs. the estimated PoBP: (a) Unbiased, (b) Biased Test set (Problem 4, XGB classifier).

8. Performance at the PoBP vs. Typical Characteristic Points

Taking into account that at the PoBP all metrics are equivalent, the selection of the PoBP as the reference point for the analysis of a classification problem appears to be attractive, especially for the identification of the best-performing classifier. On the other hand, as already mentioned, selecting the PoBP does not optimize commonly used performance metrics such as F1-score or MCC. In this section, we compare the performance at the PoBP vs. the following characteristic points typically selected during the study of a classification problem: (a) classification threshold of $\theta = 0.5$, (b) maximum Accuracy, (c) maximum MCC, (d) maximum F1-score.

Figure 5a presents the performance of the PoBP and the above characteristic points for Problem 3 (unbiased test set of size $S = 6513$, prevalence 76%), XGB classifier, and for a set of metrics: Accuracy, TPR, PPV, F1-Score, TNR, NPV, and MCC. As can be seen, the differences between PoBP and the rest in terms of Accuracy, F1-score, and MMC are quite limited. On the other hand, we observe tradeoffs between TPR and TNR as well as between PPV and NPV. In particular, TPR at the PoBP is 0.90, while at the rest points, it ranges

from 0.94 to 0.95. However, TNR is 0.67 at the PoBP, while it ranges from 0.54 to 0.60 at the other performance points. Similarly, the PPV at the PoBP is 0.90 and drops for the rest points, ranging from 0.87 to 0.88, while the NPV at the PoBP is 0.67 and increases for the rest points, ranging from 0.75 to 0.76.



Figure 5. Selected performance metrics at the PoBP vs. other characteristic points of performance: (a) Problem 3, XGB, test dataset of size $S = 6513$, (b) Problem 5, LogReg, test dataset of size $S = 5581$.

Figure 5b presents the performance of the PoBP and the above characteristic points for Problem 5, LogReg classifier (unbiased test set of size $S = 5581$, prevalence 41%). Again, the observed difference between PoBP and the rest of the characteristic performance points for Accuracy, F1-score, and MMC is quite limited. Also, we observe tradeoffs between PoBP and the rest points of performance (i.e., TPR vs. TNR, and PPV vs. NPV). However, compared to the example of Figure 5a, in this case, we observe that PoBP has lower TPR and higher TNR than the rest, except at the maximum F1-score, where the opposite applies.

The above presented examples cannot apparently be generalized for all binary classification problems. Instead, they illustrate that the selection of the PoBP can be a viable option for the study of a specific problem. However, methodologically, alternative performance points can be considered, and depending on the importance of certain aspects of the problem (e.g., it may be preferable to have a higher TPR while a lower TNR is not an issue), a different performance point may eventually be selected by a data scientist.

9. Discussion and Conclusions

The paper addresses the binary classification problem and defines the PoBP, the classification threshold for which the number of predicted positives equals the actual

number of positive samples, a property of the ideal classifier. It has been proven that at the PoBP, all binary classification performance metrics are equivalent, i.e., they yield the same results when comparing the performance of different binary classifiers. This is a significant advantage for analyzing real-world problems, as the wide variety of binary classification performance metrics in the literature often yields different conclusions about the best-performing classifier.

The literature [7–29] has thus far focused on the advantages and disadvantages of certain performance metrics, proposed new metrics, or considered methods to assess the efficiency of different metrics. From this viewpoint, the current paper takes a different direction by introducing simple concepts, such as equivalent performance metrics and the definition of the PoBP. The paper shows that the observed differences in the performance of the metrics are eliminated at the PoBP. This property allows for comparing different classifiers at their PoBP using a single metric, thereby simplifying the problem of selecting the most appropriate metric.

Given that the ROC Curve [30] is another common method for analyzing the performance of different classifiers, the paper explores the geometrical representation of the PoBP in the ROC space. The PoBP of a classifier is a single point on the ROC curve, while it is proven that the PoBP of any classifier of the same binary classification problem belongs to a straight line called the PoBP ROC curve line. As expected, the better classifier has better performance across both metrics of the ROC space.

The identification of the PoBP during the training phase is trivial, whereas an approximation method is required during inference, since the number of actual positives is unknown in that case. The paper provides three alternative methods using: (a) the prevalence in the training set, (b) the PoBP classification threshold during the training phase, and (c) the ALP Confusion Matrix. Based on the analysis of real-world binary classification problems, it is shown that the last two methods provide a good approximation of the PoBP performance, even when the prevalence in the inference set differs from that in the training dataset.

The paper then considers whether PoBP should be the preferred point of performance for analyzing a binary classification problem. To address this issue, we present examples of a classifier's performance at the PoBP vs. other commonly used points of performance (i.e., classification threshold of 0.5, maximum Accuracy, maximum MCC, and maximum F1-score). The example results indicate that PoBP may perform similarly across certain metrics (Accuracy, MCC, and F1-score), while tradeoffs may occur in other metrics (TPR, TNR, PPV, and NPV). Although these results cannot be considered applicable to all binary classification problems, the paper provides a method for comparing PoBP with other performance metrics. Depending on the specific problem requirements, the data scientist may select either the PoBP or an alternative point that best fits the problem's needs.

The next research steps include various topics such as: (a) the analysis of a multi-class classification problem aiming at the generalization of the PoBP concept, (b) the definition of equivalent performance metrics will be exploited to investigate different characteristic points of performance, and (c) the investigation of the applicability of the PoBP concept for classifiers with non-probabilistic output (e.g., k-NN).

In conclusion, the paper, through the introduction of the PoBP concept, provides new insights and a relevant methodology that simplifies the performance analysis of binary classification.

Author Contributions: Conceptualization, I.M. and G.M.; methodology, I.M.; software, G.M.; validation, I.M. and G.M.; formal analysis, I.M.; investigation, G.M.; data curation, G.M.; writing—original draft preparation, I.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in the study are included in the article.

Conflicts of Interest: The authors declare no conflicts of interest. Author I. Markoulidakis was employed by the company Public Power Company. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Appendix A. Binary Classification Problems and Applied Classifiers

Appendix A.1. Binary Classification Problems

The following table lists real-world binary classification problems used as examples in the analysis presented in the paper.

Table A1. List of example binary classification problems.

Problem	Problem Title	Number of Features	Dataset Size	Prevalence
Problem 1	Bank Analysis [35]	20	41,188	89%
Problem 2	German Data [36]	24	1000	70%
Problem 3	Adult Income [37]	14	32,561	76%
Problem 4	The Rain Problem [38]	22	145,461	76%
Problem 5	Student Depression [39]	16	27,901	41%

Appendix A.2. Applied Binary Classifiers

Decision Trees (DT): The DT [40,41] parameters used in this paper are the following: Function measuring quality of split: Entropy, Maximum depth of tree: 3, Weights associated with classes: 1.

Support Vector Machines (SVM): The following SVM [42–44] parameters have been adopted: Kernel type: Linear, Degree of polynomial kernel function: 3, Weights associated with classes: 1. Note that probabilities in this case are produced based on Python (3.13.5) scikit-learn (1.8.0) where Logistic Regression is applied over the resulting SVM scores and probabilities are then calibrated based on Platt’s method.

Random Forest (RF): The following RF [45,46] parameters have been adopted: Number of Trees: 100, Max Tree Depth Limit: 2, Measurements of split quality: Gini index.

Artificial Neural Networks (ANNs): In this paper, an ANN [47–49] with one input, one hidden, and one output layer is adopted. The ReLU is the activation function for the input and hidden layers, and the Softmax is the activation function for the output layer. The ANN consists of six hidden neurons. The optimizer is Adam, the Calculated Loss is based on sparse categorical cross-entropy, the number of Epochs is 100, and the Batch size is 10.

Convolutional Neural Networks (CNNs): The CNN [50–52] adopted in this paper is based on a Sequential mode (array of Keras Layers), the Kernel size is 3, the Pool size is 4, the activation function is ReLU, the Calculation loss is Categorical Cross-Entropy, the number of Epochs is 100, and the Batch size is 128.

Naïve Bayes (NB) [53]: A Gaussian Naïve Bayes model is used for the study performed in this paper.

Logistic Regression (LogReg): The LogReg [54] model in this paper is based on the following parameters: Max number of iterations: 5000; Optimization algorithm: L-BFGS; Weights associated with classes: 1.

Extreme Gradient Boosting (XGB) [55]: The following model parameters have been adopted in this paper: Learning Rate: 0.01, Max depth: 3, Subsample: 1.

Appendix B

Appendix B.1. Matthews Correlation Coefficient (MCC) Analysis

MCC can be expressed as follows:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{P \cdot PP \cdot N \cdot PN}} \quad (A1)$$

$$MCC = \frac{TP \cdot TN - (P - TP) \cdot (N - TN)}{\sqrt{P \cdot PP \cdot N \cdot PN}} = \left(\frac{TP}{N} + \frac{TP}{P} - 1 \right) \cdot \sqrt{\frac{P \cdot N}{PP \cdot PN}} \quad (A2)$$

$$MCC = (TPR + TNR - 1) \cdot \sqrt{\frac{P \cdot N}{PP \cdot PN}} \quad (A3)$$

Therefore, MCC can be expressed in relation to Informedness (BM) as follows:

$$MCC = BM \cdot \sqrt{\frac{P \cdot N}{PP \cdot PN}} \quad (A4)$$

Appendix B.2. Cohen's Kappa Analysis

Cohen's kappa metric is defined as follows:

$$k = \frac{p_0 - p_c}{1 - p_c}, \quad p_0 = Acc = \frac{TP + TN}{P + N}, \quad p_c = \frac{(TN + FP)(TN + FN) + (TP + FN)(TP + FP)}{(P + N)^2} \quad (A5)$$

where p_0 is the Accuracy and p_c is the expected Accuracy metric. Taking into account the relations between the CM elements (Table 1), we obtain the following:

$$k = \frac{2(TP \cdot TN - FN \cdot FP)}{(TP + FP)(TN + FP) + (TP + FN)(TN + FN)} = \frac{2(TP \cdot TN - FN \cdot FP)}{PP \cdot N + P \cdot PN} = \frac{2(TPR \cdot TNR - FNR \cdot FPR)}{PP/P + PN/N} \quad (A6)$$

And taking into account that $FPR = 1 - TPR$ and $FNR = 1 - TNR$ we obtain the relation between Cohen's kappa and Informedness (BM):

$$k = \frac{2 \cdot (TPR + TNR - 1)}{PP/P + PN/N} = \frac{2 \cdot BM}{PP/P + PN/N} \quad (A7)$$

Appendix C. Pairs of Equivalent Metrics

Appendix C.1. Equivalence of Jaccard Index (TS) and F1-Score

According to the F1-score definition, we have the following:

$$F_1 = \frac{2 \cdot TPR \cdot PPV}{TPR + PPV} = \frac{2 \cdot \frac{TP}{P} \cdot \frac{TP}{PP}}{\frac{TP}{P} + \frac{TP}{PP}} = \frac{2 \cdot TP}{PP + P} = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (A8)$$

Assuming that F1-score and TS are different than zero, the latter equation can be reformed as follows, taking into account the TS definition:

$$\frac{2}{F_1} = \frac{2 \cdot TP + FP + FN}{TP} = 1 + \frac{TP + FP + FN}{TP} = 1 + \frac{1}{TS} \quad (A9)$$

This leads to the following equation:

$$TS = \frac{F_1}{2 - F_1} \quad (A10)$$

Let us assume that comparing the TS of two classifiers, f_1, f_2 , we obtain the following:

$$TS(f_1, \theta_1) \geq TS(f_2, \theta_2) \quad (\text{A11})$$

Since F1-score ranges between 0 and 1 (so $2 - F_1 > 0$) we obtain the following inequality, which proves the fact that TS and F1-score are equivalent metrics:

$$\frac{F_1(f_1, \theta_1)}{2 - F_1(f_1, \theta_1)} \geq \frac{F_1(f_2, \theta_2)}{2 - F_1(f_2, \theta_2)} \Leftrightarrow F_1(f_1, \theta_1) \geq F_1(f_2, \theta_2) \quad (\text{A12})$$

Appendix C.2. Equivalence of Balanced Accuracy (BA) and Informedness (BM)

Considering the definitions of BA and BM, their relation is derived as follows:

$$BA = \frac{TPR + TNR}{2} = \frac{BM + 1}{2} \quad (\text{A13})$$

Let us assume that comparing the BA of two classifiers, f_1, f_2 , we obtain the following:

$$BA(f_1, \theta_1) \geq BA(f_2, \theta_2) \quad (\text{A14})$$

Based on the relation between BA and BM, it is evident that the metrics are equivalent:

$$\frac{BM(f_1, \theta_1) + 1}{2} \geq \frac{BM(f_2, \theta_2) + 1}{2} \Leftrightarrow BM(f_1, \theta_1) \geq BM(f_2, \theta_2) \quad (\text{A15})$$

Appendix C.3. Equivalence of EI to FiC

According to the EI definition (assuming that $FP + FN \neq 0$):

$$EI = \frac{TP + TN}{FP + FN} = \frac{PP - FP + PN - FN}{FP + FN} = \frac{S}{FP + FN} - 1 = \frac{1}{FiC} - 1 \quad (\text{A16})$$

For two classifiers f_1, f_2 , it is easy to prove that EI is equivalent to FiC:

$$FiC(f_1, \theta_1) \geq FiC(f_2, \theta_2) \Leftrightarrow \frac{1}{FiC(f_1, \theta_1)} \leq \frac{1}{FiC(f_2, \theta_2)} \Leftrightarrow EI(f_1, \theta_1) \leq EI(f_2, \theta_2) \quad (\text{A17})$$

Appendix C.4. Equivalence of Prevalence Threshold to LR+

Based on the PT definition and assuming that $FPR \neq 0$, we obtain its relation with LR+:

$$PT = \frac{\sqrt{TPR \cdot FPR} - FPR}{TPR - FPR} = \frac{\sqrt{TPR/FPR} - 1}{TPR/FPR - 1} = \frac{1}{\sqrt{TPR/FPR} + 1} = \frac{1}{1 + \sqrt{LR+}} \quad (\text{A18})$$

For two classifiers, f_1, f_2 , with different PT, we derive the equivalence of PT and LR+:

$$PT(f_1, \theta_1) \geq PT(f_2, \theta_2) \Leftrightarrow LR + (f_1, \theta_1) \leq LR + (f_2, \theta_2) \quad (\text{A19})$$

Appendix D. Equivalence of Metrics at the PoBP

Appendix D.1. Equivalence of Accuracy with PPV at the Point-of-Balanced Performance

The relation of Accuracy and PPV becomes ($TN = N - FP$ and $FP = PP - TP$):

$$Acc = \frac{TP + TN}{S} = \frac{TP + (TP + N - PP)}{S} = \frac{PP}{S} \left(2 \cdot PPV - 1 + \frac{N}{PP} \right) \quad (\text{A20})$$

Since at the PoBP, $PP = P$; hence, we obtain the following:

$$Acc(f, \theta_b) = \frac{P}{S} \left(2 \cdot PPV(f, \theta_b) - 1 + \frac{N}{P} \right) \quad (A21)$$

For classifiers f_1, f_2 , Equation (A21) indicates that at the PoBP, Accuracy is equivalent to PPV:

$$Acc(f_1, \theta_{b1}) \geq Acc(f_2, \theta_{b2}) \Leftrightarrow PPV(f_1, \theta_{b1}) \geq PPV(f_2, \theta_{b2}) \quad (A22)$$

Appendix D.2. Equivalence of TNR to TPR at the Point-of-Balanced Performance

The relation between TNR and TPR is derived as follows:

$$TNR = \frac{TN}{N} = \frac{N - FP}{N} = \frac{N - (PP - TP)}{N} = 1 - \frac{PP - TP}{N} = 1 - \frac{PP}{N} + \frac{P}{N} \cdot TPR \quad (A23)$$

At the PoBP (where $PP = P$), we obtain the following:

$$TNR(f, \theta_b) = 1 - \frac{P}{N} + \frac{P}{N} \cdot TPR(f, \theta_b) \quad (A24)$$

For two classifiers f_1, f_2 , Equation (A24) indicates that at the PoBP, TNR is equivalent to TPR:

$$TNR(f_1, \theta_{b1}) \geq TNR(f_2, \theta_{b2}) \Leftrightarrow TPR(f_1, \theta_{b1}) \geq TPR(f_2, \theta_{b2}) \quad (A25)$$

Appendix D.3. Equivalence of Informedness to TPR at the Point-of-Balanced Performance

Assuming that two classifiers f_1, f_2 have a different TPR at their PoBP and considering the equivalence of TPR with TNR at the PoBP (Equation (A25)), we obtain the following:

$$\begin{aligned} TPR(f_1, \theta_{b1}) \geq TPR(f_2, \theta_{b2}) &\Leftrightarrow TPR(f_1, \theta_{b1}) + TNR(f_1, \theta_{b1}) \\ &\geq TPR(f_2, \theta_{b2}) + TNR(f_2, \theta_{b2}) \end{aligned} \quad (A26)$$

From the definition of BM, it is easy to show that it is equivalent to TPR at the PoBP:

$$TPR(f_1, \theta_{b1}) \geq TPR(f_2, \theta_{b2}) \Leftrightarrow BM(f_1, \theta_{b1}) \geq BM(f_2, \theta_{b2}) \quad (A27)$$

Appendix D.4. Equivalence of P4 to TPR at the Point-of-Balanced Performance

For two classifiers f_1, f_2 with different TPR at their PoBP (for $TPR \neq 0$), we obtain the following:

$$TPR(f_1, \theta_{b1}) \geq TPR(f_2, \theta_{b2}) \Leftrightarrow \frac{1}{TPR(f_1, \theta_{b1})} \leq \frac{1}{TPR(f_2, \theta_{b2})} \quad (A28)$$

Since TPR and TNR are equivalent at their PoBP (Equation (A25)) (for $TNR \neq 0$), we obtain the following:

$$TPR(f_1, \theta_{b1}) \geq TPR(f_2, \theta_{b2}) \Leftrightarrow \frac{1}{TNR(f_1, \theta_{b1})} \leq \frac{1}{TNR(f_2, \theta_{b2})} \quad (A29)$$

Based on the above inequalities, we achieve the following:

$$\frac{1}{TPR(f_1, \theta_{b1})} + \frac{1}{TNR(f_1, \theta_{b1})} \leq \frac{1}{TPR(f_2, \theta_{b2})} + \frac{1}{TNR(f_2, \theta_{b2})} \quad (A30)$$

Considering the P4 metric definition and the fact that at the PoBP, TPR becomes equal to PPV (Equation (9)) and TNR becomes equal to NPV (Equation (11)), we obtain the following:

$$P4(f, \theta_b) = \frac{1}{\frac{1}{TPR(f, \theta_b)} + \frac{1}{PPV(f, \theta_b)} + \frac{1}{TNR(f, \theta_b)} + \frac{1}{NPV(f, \theta_b)}} = \frac{1}{2 \left[\frac{1}{TPR(f, \theta_b)} + \frac{1}{TNR(f, \theta_b)} \right]} \quad (A31)$$

From the above analysis:

$$TPR(f_1, \theta_{b1}) \geq TPR(f_2, \theta_{b2}) \Leftrightarrow \frac{1}{2 \cdot P4(f_1, \theta_{b1})} \leq \frac{1}{2 \cdot P4(f_2, \theta_{b2})} \quad (A32)$$

The latter inequality implies that TPR becomes equivalent to the P4 metric at the PoBP:

$$TPR(f_1, \theta_{b1}) \geq TPR(f_2, \theta_{b2}) \Leftrightarrow P4(f_1, \theta_{b1}) \geq P4(f_2, \theta_{b2}) \quad (A33)$$

Appendix D.5. Equivalence of LR+ to TPR at the Point-of-Balanced Performance

According to LR+ definition:

$$LR+ = \frac{TPR}{FPR} = \frac{TPR}{1 - TNR} \quad (A34)$$

Assuming that two classifiers f_1, f_2 have a different TPR at their PoBP and considering the equivalence between TPR and TNR at the same point (Equation (A25)), we obtain the following:

$$TPR(f_1, \theta_{b1}) \geq TPR(f_2, \theta_{b2}) \Leftrightarrow 1 - TNR(f_1, \theta_{b1}) \leq 1 - TNR(f_2, \theta_{b2}) \quad (A35)$$

Assuming that $TNR \neq 1$, the latter inequality becomes the following:

$$TPR(f_1, \theta_{b1}) \geq TPR(f_2, \theta_{b2}) \Leftrightarrow \frac{TPR(f_1, \theta_{b1})}{1 - TNR(f_1, \theta_{b1})} \geq \frac{TPR(f_2, \theta_{b2})}{1 - TNR(f_2, \theta_{b2})} \quad (A36)$$

The above inequality indicates that TPR is equivalent to LR+ at the PoBP:

$$TPR(f_1, \theta_{b1}) \geq TPR(f_2, \theta_{b2}) \Leftrightarrow LR+(f_1, \theta_{b1}) \geq LR+(f_2, \theta_{b2}) \quad (A37)$$

Appendix D.6. Equivalence of LR− to TPR at the Balanced Point of Performance

According to LR− definition, the following applies:

$$LR- = \frac{FNR}{TNR} = \frac{1 - TPR}{TNR} \quad (A38)$$

Assuming that two classifiers f_1, f_2 have a different TPR at their PoBP, we obtain the following:

$$TPR(f_1, \theta_{b1}) \geq TPR(f_2, \theta_{b2}) \Leftrightarrow 1 - TPR(f_1, \theta_{b1}) \leq 1 - TPR(f_2, \theta_{b2}) \quad (A39)$$

Since at the PoBP, TPR becomes equivalent to TNR (see Equation (A25)), we achieve the following:

$$TPR(f_1, \theta_{b1}) \geq TPR(f_2, \theta_{b2}) \Leftrightarrow \frac{1 - TPR(f_1, \theta_{b1})}{TNR(f_1, \theta_{b1})} \leq \frac{1 - TPR(f_2, \theta_{b2})}{TNR(f_2, \theta_{b2})} \quad (A40)$$

The above inequality indicates that TPR is equivalent to LR− at the PoBP:

$$TPR(f_1, \theta_{b1}) \geq TPR(f_2, \theta_{b2}) \Leftrightarrow LR - (f_1, \theta_{b1}) \leq LR - (f_2, \theta_{b2}) \quad (A41)$$

Appendix D.7. Equivalence of DOR to TPR at the Point-of-Balanced Performance

Let us assume a comparison of two classifiers, f_1 and f_2 , at their PoBP. Based on the equivalence of TPR with LR+ and LR− metrics at the PoBP (see Equations (A37) and (A41)), we obtain the following:

$$TPR(f_1, \theta_{b1}) \geq TPR(f_2, \theta_{b2}) \Leftrightarrow \frac{LR + (f_1, \theta_{b1})}{LR - (f_1, \theta_{b1})} \geq \frac{LR + (f_2, \theta_{b2})}{LR - (f_2, \theta_{b2})} \quad (A42)$$

Therefore, TPR and DOR are equivalent at the PoBP:

$$TPR(f_1, \theta_{b1}) \geq TPR(f_2, \theta_{b2}) \Leftrightarrow DOR(f_1, \theta_{b1}) \geq DOR(f_2, \theta_{b2}) \quad (A43)$$

Appendix D.8. Equivalence of G-Mean Score with TPR at the Point-of-Balanced Performance

If two classifiers f_1, f_2 have a different TPR at their PoBP, then based on the equivalence of TPR with TNR at the PoBP (Equation (A44)) and since TPR and TNR are positive, we achieve the following:

$$\begin{aligned} TPR(f_1, \theta_{b1}) \geq TPR(f_2, \theta_{b2}) &\Leftrightarrow TPR(f_1, \theta_{b1}) \cdot TNR(f_1, \theta_{b1}) \\ &\geq TPR(f_2, \theta_{b2}) \cdot TNR(f_2, \theta_{b2}) \end{aligned} \quad (A44)$$

Considering that both TPR and TNR range between 0 and 1, we obtain the following:

$$\begin{aligned} TPR(f_1, \theta_{b1}) \geq TPR(f_2, \theta_{b2}) &\Leftrightarrow \sqrt{TPR(f_1, \theta_{b1}) \cdot TNR(f_1, \theta_{b1})} \\ &\geq \sqrt{TPR(f_2, \theta_{b2}) \cdot TNR(f_2, \theta_{b2})} \end{aligned} \quad (A45)$$

From this inequality, it is evident that the G-Mean Score is equivalent to TPR at the PoBP:

$$TPR(f_1, \theta_{b1}) \geq TPR(f_2, \theta_{b2}) \Leftrightarrow GM(f_1, \theta_{b1}) \geq GM(f_2, \theta_{b2}) \quad (A46)$$

References

1. Alpaydin, E. *Introduction to Machine Learning*; The MIT Press: Cambridge, MA, USA, 2010. ISBN 978-0-262-01243-0.
2. Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Comput. Sci.* **2021**, *2*, 160. [\[CrossRef\]](#)
3. Garg, A.; Roth, D. Understanding probabilistic classifiers. In Proceedings of the ECML 2001 12th European Conference on Machine Learning, LNAI 2167, Freiburg, Germany, 5–7 September 2001; pp. 179–191.
4. Stehman, S.V. Selecting and interpreting measures of thematic classification accuracy. *Remote Sens. Environ.* **1997**, *62*, 77–89. [\[CrossRef\]](#)
5. Ting, K.M. Confusion Matrix. In *Encyclopedia of Machine Learning and Data Mining*; Springer: Berlin/Heidelberg, Germany, 2010.
6. Uddin, S.; Khan, A.; Hossain, M.; Moni, M.A. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 281. [\[CrossRef\]](#)
7. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [\[CrossRef\]](#)
8. Cao, C.; Chicco, D.; Hoffman, M.M. The MCC-F1 curve: A performance evaluation technique for binary classification. *arXiv* **2020**, arXiv:2006.11278. [\[CrossRef\]](#)
9. Chicco, D.; Jurman, G. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Min.* **2023**, *16*, 4. [\[CrossRef\]](#) [\[PubMed\]](#) [\[PubMed Central\]](#)
10. Chicco, D.; Warrens, M.; Jurman, G. The Matthews Correlation Coefficient (MCC) is More Informative Than Cohen’s Kappa and Brier Score in Binary Classification Assessment. *IEEE Access* **2021**, *9*, 78368–78381. [\[CrossRef\]](#)
11. Chicco, D.; Jurman, G. A statistical comparison between Matthews correlation coefficient (MCC), prevalence threshold, and Fowlkes–Mallows index. *J. Biomed. Inform.* **2023**, *144*, 104426. [\[CrossRef\]](#)

12. Itaya, Y.; Tamura, J.; Hayashi, K.; Yamamoto, K. Asymptotic Properties of Matthews Correlation Coefficient. *Stat. Med.* **2025**, *44*, e10303. [\[CrossRef\]](#)
13. Zhu, Q. On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset. *Pattern Recognit. Lett.* **2020**, *136*, 71–80. [\[CrossRef\]](#)
14. Shao, G.; Tang, L.; Liao, J. Overselling overall map accuracy misinforms about research reliability. *Landsc. Ecol.* **2019**, *34*, 2487–2492. [\[CrossRef\]](#)
15. Leeflang, M.M.; Bossuyt, P.M.; Irwig, L. Diagnostic test accuracy may vary with prevalence: Implications for evidence-based diagnosis. *J. Clin. Epidemiol.* **2009**, *62*, 5–12. [\[CrossRef\]](#)
16. Powers, D. The problem with kappa. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12)*; Association for Computational Linguistics: Avignon, France, 2012; pp. 345–355.
17. Powers, D.M.W. Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation. *arXiv* **2011**, arXiv:2010.16061.
18. Larner, A.J. Efficiency Index for Binary Classifiers: Concept, Extension, and Application. *Mathematics* **2023**, *11*, 2435. [\[CrossRef\]](#)
19. Kuncheva, L.I.; Arnaiz-González, Á.; Díez-Pastor, J.-F.; Gunn, I.A.D. Instance selection improves geometric mean accuracy: A study on imbalanced data classification. *Prog. Artif. Intell.* **2019**, *8*, 215–228. [\[CrossRef\]](#)
20. Marra, A. G4 & the balanced metric family—A novel approach to solving binary classification problems in medical device validation & verification studies. *BioData Min.* **2024**, *17*, 43. [\[CrossRef\]](#)
21. Lipton, Z.C.; Elkan, C.; Naryanaswamy, B. Optimal Thresholding of Classifiers to Maximize F1 Measure. In *Proceedings of the Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD*; Springer: Berlin/Heidelberg, Germany, 2014; Volume 8725, pp. 225–239.
22. Canbek, G.; Sagioglu, S.; Temizel, T.T.; Baykal, N. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In *2017 International Conference on Computer Science and Engineering (UBMK)*; IEEE: New York, NY, USA, 2017; pp. 821–826. [\[CrossRef\]](#)
23. Hernández-Orallo, J.; Flach, P.; Ferri, C. A unified view of performance metrics: Translating threshold choice into expected classification loss. *J. Mach. Learn. Res.* **2012**, *13*, 2813–2869.
24. Christen, P.; Hand, D.J.; Kirielle, N. A Review of the F-Measure: Its History, Properties, Criticism, and Alternatives. *ACM Comput. Surv.* **2023**, *56*, 73. [\[CrossRef\]](#)
25. Hand, D.J.; Christen, P.; Kirielle, N. F*: An interpretable transformation of the F-measure. *Mach. Learn.* **2021**, *110*, 451–456. [\[CrossRef\]](#) [\[PubMed\]](#) [\[PubMed Central\]](#)
26. Kim, H.; Park, Y. Calibrating F1 Scores for Fair Performance Comparison of Binary Classification Models with Application to Student Dropout Prediction. *IEEE Access* **2025**, *13*, 136554–136567. [\[CrossRef\]](#)
27. Walauskis, M.A.; Khoshgoftaar, T.M. Choosing the Right Metrics: A Study of Performance Measurement for Binary Classification in Imbalanced and Big Data. *Int. FLAIRS Conf. Proc.* **2025**, *38*. [\[CrossRef\]](#)
28. Canbek, G.; Taskaya Temizel, T.; Sagioglu, S. BenchMetrics: A systematic benchmarking method for binary classification performance metrics. *Neural Comput. Applic* **2021**, *33*, 14623–14650. [\[CrossRef\]](#)
29. Shirdel, M.; Di Mauro, M.; Liotta, A. Worthiness Benchmark: A novel concept for analyzing binary classification evaluation metrics. *Inf. Sci.* **2024**, *678*, 120882. [\[CrossRef\]](#)
30. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. [\[CrossRef\]](#)
31. Hanley, J.A.; McNeil, B.J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **1982**, *143*, 29–36. [\[CrossRef\]](#) [\[PubMed\]](#)
32. Refaeilzadeh, P.; Tang, L.; Liu, H. Cross-Validation. In *Encyclopedia of Database Systems*; Liu, L., Özsu, M.T., Eds.; Springer: Boston, MA, USA, 2009. [\[CrossRef\]](#)
33. Markoulidakis, I.; Markoulidakis, G. Probabilistic Confusion Matrix: A Novel Method for Machine Learning Algorithm Generalized Performance Analysis. *Technologies* **2024**, *12*, 113. [\[CrossRef\]](#)
34. Niculescu-Mizil, A.; Caruana, R. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML' 05)*, Bonn, Germany, 7–11 August 2005; Association for Computing Machinery: New York, NY, USA, 2005; pp. 625–632.
35. Moro, S.; Cortez, P.; Rita, P. A data-driven approach to predict the success of bank telemarketing. *Decis. Support Syst.* **2014**, *62*, 22–31. [\[CrossRef\]](#)
36. Hofmann, H. Statlog (German Credit Data). Machine Learning Repository. 1994, Volume 53. Available online: <https://archive.ics.uci.edu/dataset/144/statlog+german+credit+data> (accessed on 20 February 2026).
37. Becker, B.; Kohavi, R. Adult [Dataset]. UCI Machine Learning Repository. 1996. Available online: <https://archive.ics.uci.edu/dataset/2/adult> (accessed on 20 February 2026).
38. Australia Meteorology Government Bureau. Available online: <http://www.bom.gov.au/> (accessed on 20 February 2026).

39. Garcia-Ceja, E.; Riegler, M.; Jakobsen, P.; Tørresen, J.; Nordgreen, T.; Oedegaard, K.J.; Fasmer, O.B. A Motor Activity Database of Depression Episodes in Unipolar and Bipolar Patients. In Proceedings of the MMSys'18 9th ACM on Multimedia Systems Conference, Amsterdam, The Netherlands, 12–15 June 2018. Available online: <https://dl.acm.org/doi/pdf/10.1145/3204949.3208125> (accessed on 20 February 2026).
40. Quinlan, J.R. Induction of decision trees. *Mach. Learn.* **1986**, *1*, 81–106. [[CrossRef](#)]
41. Rokach, L.; Maimon, O.Z. *Data Mining with Decision Trees: Theory and Applications*; World Scientific: Singapore, 2008; Volume 69.
42. Basak, D.; Srimanta, P.; Patranabis, D.C. Support Vector Regression. *Neural Inf. Process.-Lett. Rev.* **2007**, *11*, 203–224.
43. Abe, S. *Support Vector Machines for Pattern Classification*, 2nd ed.; Advances in Computer Vision and Pattern Recognition; Springer: London, UK, 2010. [[CrossRef](#)]
44. Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. [[CrossRef](#)]
45. Pal, M. Random forest classifier for remote sensing classification. *Int. J. Remote Sens.* **2005**, *26*, 217–222. [[CrossRef](#)]
46. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [[CrossRef](#)]
47. Walczak, S.; Cerpa, N. *Artificial Neural Networks, Encyclopedia of Physical Science and Technology*, 3rd ed.; Academic Press: Cambridge, MA, USA, 2003; pp. 631–645. ISBN 9780122274107. [[CrossRef](#)]
48. Doulamis, A.; Doulamis, N.; Kollias, S. On-line retrainable neural networks: Improving the performance of neural networks in image analysis problems. *IEEE Trans. Neural Netw.* **2000**, *11*, 137–155. [[CrossRef](#)] [[PubMed](#)]
49. Haykin, S. *Neural Networks: A Comprehensive Foundation*; Prentice-Hall Inc.: Upper Anhe, NJ, USA, 2007.
50. Doulamis, A.; Doulamis, N.; Protopapadakis, E.; Voulodimos, A. Combined Convolutional Neural Networks and Fuzzy Spectral Clustering for Real Time Crack Detection in Tunnels. In Proceedings of the 2018 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 4153–4157. [[CrossRef](#)]
51. Albawi, S.; Mohammed, T.A.; Al-Zawi, S. Understanding of a convolutional neural network. In Proceedings of the 2017 International Conference on Engineering and Technology (ICET), Antalya, Turkey, 21–23 August 2017; pp. 1–6. [[CrossRef](#)]
52. Chauhan, R.; Ghanshala, K.K.; Joshi, R.C. Convolutional Neural Network (CNN) for Image Detection and Recognition. In Proceedings of the 2018 1st International Conference on Secure Cyber Computing and Communication (ICSCCC), Jalandhar, India, 15–17 December 2018; pp. 278–282. [[CrossRef](#)]
53. Haouari, B.; Amor, N.B.; Elouedi, Z.; Mellouli, K. Naïve possibilistic network classifiers. *Fuzzy Sets Syst.* **2009**, *160*, 3224–3238. [[CrossRef](#)]
54. Hosmer, D.W., Jr.; Lemeshow, S.; Sturdivant, R.X. *Applied Logistic Regression*; JohnWiley & Sons: Hoboken, NJ, USA, 2013; Volume 398.
55. Chen, T.; Guestrin, C. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.

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.