

Article

Not peer-reviewed version

---

# Why ROC-AUC Is Misleading for Highly Imbalanced Data: In-Depth Evaluation of MCC, F2-score, H-measure, and AUC-based Metrics across Diverse Classifiers

---

[Mehdi Imani](#)<sup>\*</sup>, [Majid Joudaki](#)<sup>\*</sup>, [Ayoub Bagheri](#), [Hamid R. Arabnia](#)

Posted Date: 13 January 2026

doi: 10.20944/preprints202510.0958.v2

Keywords: SMOTE; ADASYN; H-measure; MCC; PR-AUC; ROC-AUC; XGBoost; CatBoost; random forest; logistic regression



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a [Creative Commons CC BY 4.0 license](#), which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

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.

Article

# Why ROC-AUC Is Misleading for Highly Imbalanced Data: In-Depth Evaluation of MCC, F2-score, H-measure, and AUC-based Metrics across Diverse Classifiers

Mehdi Imani <sup>1,\*</sup>, Majid Joudaki <sup>2,\*</sup>, Ayoub Bagheri <sup>3</sup> and Hamid R. Arabnia <sup>4</sup>

<sup>1</sup> Department of Computer and System Sciences, Stockholm University, 10691 Stockholm, Sweden

<sup>2</sup> Department of Computer Engineering, Faculty of Engineering, Ayatollah Boroujerdi University, 69199-69737, Boroujerd, Iran

<sup>3</sup> Department of Methodology and Statistics, Utrecht University, Padualaan 14, 3584 CH, Utrecht, The Netherlands

<sup>4</sup> School of Computing, University of Georgia, Athens, GA 30602, USA

\* Correspondence: m.imani@gmail.com (M.I.); m.joudaki@abru.ac.ir (M.J.)

## Abstract

This study re-evaluates ROC-AUC for binary classification under severe class imbalance (<3% positives). Despite its widespread use, ROC-AUC can mask operationally salient differences among classifiers when the costs of false positives and false negatives are asymmetric. Using three benchmarks, credit-card fraud detection (0.17%), yeast protein localization (1.35%), and ozone level detection (2.9%), we compare ROC-AUC with Matthews Correlation Coefficient, F2-score, H-measure, and PR-AUC. Our empirical analyses span 20 classifier–sampler configurations per dataset, combined with four classifiers (Logistic Regression, Random Forest, XGBoost, and CatBoost) and four oversampling methods plus a no-resampling baseline (no resampling, SMOTE, Borderline-SMOTE, SVM-SMOTE, ADASYN). ROC-AUC exhibits pronounced ceiling effects, yielding high scores even for underperforming models. In contrast, MCC and F2 align more closely with deployment-relevant costs and achieve the highest Kendall's  $\tau$  rank concordance across datasets; PR-AUC provides threshold-independent ranking, and H-measure integrates cost sensitivity. We quantify uncertainty and differences using stratified bootstrap confidence intervals, DeLong's test for ROC-AUC, and Friedman–Nemenyi critical-difference diagrams, which collectively underscore the limited discriminative value of ROC-AUC in rare-event settings. The findings recommend a shift to a multi-metric evaluation framework: ROC-AUC should not be used as the primary metric in ultra-imbalanced settings; instead, MCC and  $F_2$  are recommended as primary indicators, supplemented by PR-AUC and H-measure where ranking granularity and principled cost integration are required. This evidence encourages researchers and practitioners to move beyond sole reliance on ROC-AUC when evaluating classifiers in highly imbalanced data.

**Keywords:** SMOTE; ADASYN; H-measure; MCC; PR-AUC; ROC-AUC; XGBoost; CatBoost; random forest; logistic regression

## 1. Introduction

Rare-event binary classification remains a significant challenge across various domains, including but not limited to fraud detection, bioinformatics, environmental monitoring, cybersecurity, medical diagnosis, manufacturing fault detection, and autonomous systems. In these settings, the positive class typically constitutes a small fraction of the data, making accurate detection both technically difficult and operationally critical. Standard tools for model evaluation, particularly

Area Under the Receiver Operating Characteristic Curve (ROC-AUC), often yield inflated assessments due to their insensitivity to class imbalance and asymmetric error costs. At the same time, several studies argue that ROC-AUC remains statistically robust to class imbalance and provides a stable, threshold-independent ranking measure, motivating ongoing debate rather than a settled consensus. As machine learning systems become integral to decision-making in high-stakes, real-world environments, it is essential to critically evaluate both the data preprocessing methods and performance metrics used in model assessment. This study presents a comprehensive, metric-centered investigation of classifier evaluation under highly imbalanced data conditions, focusing on resampling strategies and the behavior of metrics. Despite extensive discussion of performance metrics for imbalanced classification, existing studies typically focus on advocating individual alternatives to the ROC-AUC or provide conceptual arguments without a unified empirical protocol. As a result, a lack of statistically rigorous, cross-domain evaluation remains, which jointly examines how different metric families behave and potentially disagree under ultra-imbalanced prevalence and realistic resampling pipelines. This unresolved debate regarding whether the ranking robustness of ROC-AUC outweighs the prevalence sensitivity of PR-AUC directly motivates this study. Specifically, we investigate how different metric families behave, diverge, and ultimately influence model selection under ultra-imbalanced conditions. Accordingly, our analysis tests the hypothesis that no single metric is sufficient in rare-event settings, and that complementary, cost-aware metrics are required for reliable evaluation. The subsections below provide background on the methodological foundations that guide our analysis.

### 1.1. Resampling Strategies in Imbalanced Data

When the minority-class prior satisfies 1,

$$\pi = Pr(y = 1) \ll 0.5, \quad (1)$$

empirical risk minimization with a symmetric loss favors the majority class [1]. One approach is resampling, i.e., constructing a training set whose posterior prior  $\pi^*$  is closer to 0.5. Let  $N_1$ ,  $N_0$  denote the counts of minority (class 1) and majority (class 0) instances, and let  $r$  be an oversampling factor applied to the minority class. After oversampling, the adjusted prior is given by 2:

$$\pi^* = \frac{r N_1}{r N_1 + N_0}, \text{ with } r = \frac{\pi^*}{(1-\pi^*)} \frac{N_0}{N_1}, \quad (2)$$

when  $r$  satisfies 2, the resulting training set achieves a target prior  $\pi^*$  that is closer to 0.5.

A perfectly balanced set, therefore, corresponds to  $r = (N_0/N_1)$ . The subsections below summarise the major families of resampling and highlight their theoretical motivations. Random undersampling, which involves randomly discarding the majority of instances, reduces  $Pr(y = 0)$  to  $\pi^* \approx 0.5$  [2]. While computationally attractive, this approach may eliminate informative majority examples and increase the variance of the estimator. Ensemble variants, such as EasyEnsemble and BalanceCascade [3], mitigate this issue by building multiple classifiers on independently undersampled subsets and aggregating their predictions. Random oversampling, which replicates minority observations to achieve the desired ratio, is unbiased in expectation; however, it can result in exact duplicate rows, leading to overfitting [4]. The expected Bayes risk decreases only if the learner regularises against memorisation. To avoid duplication, synthetic minority over-sampling generates artificial instances according to 3:

$$x_{new} = x_i + \lambda(x_{NN} - x_i), \quad \lambda \sim \mathcal{U}(0,1), \quad (3)$$

where  $x_i$  is a minority point and  $x_{NN}$  one of its  $k$  minority nearest neighbors [5]. Extensions refine the neighborhood criterion:

1. Borderline-SMOTE (Synthetic Minority Oversampling Technique) focuses on minority instances whose nearest neighbors are predominantly from the majority class, thereby increasing density near the decision boundary [6].
2. SVM-SMOTE, a Support Vector Machine-based extension of SMOTE, generates synthetic samples near the decision boundary using support vectors identified by the classifier [7].
3. Safe-Level-SMOTE assigns a safety level defined in 4:

$$SL(x_i) = k^{-1} \sum_{j=1}^k 1(y_j = 1), \quad (4)$$

and chooses the interpolation factor  $\lambda$  so that the synthetic point lies closer to the parent with a higher safe-level score to avoid generating samples in dangerous regions [8].

4. ADASYN (Adaptive Synthetic Sampling) adaptively varies the number of synthetic samples per minority point according to the local imbalance ratio defined in 5:

$$G_i = \frac{\delta_i}{\sum_j \delta_j} G, \quad (5)$$

where  $\delta_i$  is the proportion of majority neighbors [9]. This shifts density toward sparsely represented minority areas.

Combining oversampling with Tomek links deletion [10] or Edited Nearest Neighbors (ENN) [11] removes the majority of points lying within the minority manifold, reducing class overlap [12]. Empirically, SMOTE + ENN often yields smoother decision surfaces than either step alone [13].

Density-aware and generative approaches employ information-theoretic or generative criteria. Gaussian-SMOTE (G-SMOTE) replaces linear interpolation with a Gaussian mixture model of the minority class [14], while Generative Adversarial Network (GAN)-based over-samplers learn  $p(x | y = 1)$  implicitly via adversarial training [15]. Theoretical analyses show that, under a Lipschitz assumption on the Bayes decision boundary, synthetic samples drawn from a contiguous minority manifold can reduce the upper bound on the classification error by tightening the margin [16].

## 1.2. Performance Metrics in Binary Classification

Let the confusion matrix for a binary classifier at threshold  $t$  be as Table 1.

**Table 1.** Confusion matrix for a binary classifier

True\Predicted	$\hat{y} = 1$	$\hat{y} = 0$
$y = 1$	TP <sup>a</sup> ( $t$ )	FN <sup>b</sup> ( $t$ )
$y = 0$	FP <sup>c</sup> ( $t$ )	TN <sup>d</sup> ( $t$ )

<sup>a</sup> TP = True Positive, <sup>b</sup> FN = False Negative, <sup>c</sup> FP = False Positive, <sup>d</sup> TN = True Negative

Any scalar score reduces this 2×2 matrix, or in threshold-free form, the ranking of class scores  $s(x)$ , to a single real number. The following subsections review the principal families of metrics, their mathematical properties, and known limitations.

### 1.2.1. Threshold-independent Discrimination

The receiver-operating-characteristic area (ROC-AUC) is defined in 6:

$$ROC - AUC = P_r(s(x^+) > s(x^-)) = \iint 1(s(x^+) > s(x^-)) dF_+ dF_-, \quad (6)$$

where  $F_+$ ,  $F_-$  are the score distributions for positives and negatives [17, 18]. ROC-AUC is equivalent to the Mann-Whitney  $U$  statistic and is invariant under strictly monotone score transformations. Its major weakness is class-imbalance insensitivity: when  $N_1 \ll N_0$  (with  $N_1 = TP + FN$ ,  $N_0 = FP + TN$ ), significant changes in FP translate into tiny variations of the false-positive rate (FPR) [19]. Consider a binary classifier that assigns each instance a score  $s(x)$  and applies a threshold  $t$  to decide between positive and negative. The two axes of its ROC curve are defined in 7:

$$TPR(t) = \frac{TP(t)}{TP(t) + FN(t)}, \quad FPR(t) = \frac{FP(t)}{FP(t) + TN(t)}, \quad (7)$$

In many practical settings, the number of positive cases  $N_1$  is vanishingly small compared to negatives  $N_0$ . In that extreme imbalance, the false positive rate denominator is effective  $N_0$ , so even a significant absolute change in FP produces only a barely perceptible shift in the ROC curve.

This distortion carries over to the AUC itself. Equivalently expressed as 8:

$$AUC = Pr(s(x^+) > s(x^-)), \quad (8)$$

the statistic is overwhelmed by comparisons among the abundant negative-negative pairs. As long as a model avoids egregious score inversions, moderate numbers of false alarms or misses scarcely register in AUC. The upshot is that, under severe imbalance, nearly any non-degenerate classifier achieves an ROC-AUC in the 0.90–0.99 range, obscuring the errors that drive operational cost. By treating false positives and false negatives (FN) symmetrically, ROC-AUC “wins” without ever “paying” for the mistakes that, in domains like fraud or medical diagnosis, are most consequential. Replacing the x-axis with recall yields the precision–recall (PR) curve; its area, defined in 9, is given by

$$PR - AUC = \int_0^1 P(R) dR, \quad P = \frac{TP}{TP+FP}, \quad (9)$$

has a baseline equal to the minority prevalence  $\pi = \frac{N_1}{N_0+N_1}$ . Davis and Goadrich [20] prove that Area Under the Precision–Recall Curve (PR-AUC) is strictly more informative than ROC-AUC when  $\pi$  is small.

In a similar effort, the H-measure [21] is a coherent alternative to the area under the ROC curve that explicitly incorporates the relative cost of false positives and false negatives via a user-specified distribution over misclassification cost-ratios. Rather than treating all operating points equally, the H-measure defines a weighting density  $u(c)$  on the cost-ratio  $c \in [0,1]$  and computes the expected misclassification loss as defined in 10:

$$\hat{L} = \int_0^1 [\pi_0 c FPR(\tau_c) + \pi_1 (1-c) FNR(\tau_c)] u(c) dc,$$

where  $\pi_1$  and  $\pi_0 = 1 - \pi_1$  are the class priors, and  $\tau_c$  is the threshold minimizing the cost for a given  $c$ . By default, one chooses  $u(c)$  to be the Beta(2,2) density, yielding a neutral prior that neither over- nor under-emphasizes extreme cost ratios. The H-measure is then normalized by the worst-possible expected loss under the same density, producing a summary score in  $[0,1]$  as shown in 11:

$$H = 1 - \frac{\hat{L}}{L_{max}}, \quad (11)$$

In contrast to ROC-AUC, which is dominated by the vast number of negative–negative score-pair comparisons under severe imbalance and therefore remains artificially high even when a classifier makes many costly errors, the H-measure penalizes errors proportionally to their operational importance. In highly skewed scenarios (e.g., fraud detection, rare-disease screening), it provides a more discriminating evaluation: classifiers that sacrifice minority-class sensitivity or incur excessive false alarms receive a substantially lower H-measure, whereas the ROC-AUC remains saturated.

### 1.2.2. Single-threshold Confusion-matrix Scores

The Matthews Correlation Coefficient (MCC) and the  $F_\beta$  score are two widely used scalars. The MCC, defined in 12, is computed as

$$MCC = \frac{TP \, TN - FP \, FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

and ranges within  $[-1,1]$  [22].

MCC is a special case of Pearson’s  $r$  between prediction and truth. MCC treats both classes symmetrically and remains well defined even when one row or column is small [23].

The  $F_\beta$  score, defined in 13, is expressed as

$$F_\beta = (1 + \beta^2) \frac{PR}{\beta^2 P + R}, \quad \beta > 1,$$

where  $P$  and  $R$  denote precision and recall, respectively [24].

Unlike MCC,  $F_\beta$  ignores true negatives; it is therefore sensitive to prevalence and may exaggerate performance in dense negative regions [25]. An alternative that explicitly accounts for prevalence is balanced accuracy, defined in 14:

$$BA = \frac{1}{2}(TPR + TNR) = \frac{1}{2}\left(\frac{TP}{N_1} + \frac{TN}{N_0}\right),$$

which has been recently advocated for class-imbalance evaluation [26].

Another related metric, Cohen's  $\kappa$  (kappa), is a chance-corrected measure of agreement that quantifies how much better a classifier's predictions agree with the true labels than would be expected by random chance [27]. For a two-category problem, let the observed and expected agreements be defined in 15 and 16, respectively:

$$P_o = \frac{\text{number of correct predictions}}{N}, \quad (15)$$

and

$$P_e = \sum_{c \in \{0,1\}} \left(\frac{N_c^{pred}}{N} \times \frac{N_c^{true}}{N}\right), \quad (16)$$

where  $N_c^{pred}$  and  $N_c^{true}$  are the counts of predicted and true instances in class  $c$ . Cohen's  $\kappa$  is then computed as shown in 17:

$$\kappa = \frac{P_o - P_e}{1 - P_e}, \quad (17)$$

which ranges from  $-1$  (complete disagreement) through  $0$  (no better than chance) to  $1$  (perfect agreement).

Unlike raw accuracy,  $\kappa$  corrects for any agreement that would arise simply from the marginal class frequencies, which is an important feature when classes are highly imbalanced. In such settings, a naive classifier can achieve high accuracy (and thus high  $P_o$ ) by always predicting the majority class, yet its  $\kappa$  will remain low because  $P_e$  is large. Interpretive benchmarks [28] suggest that  $\kappa < 0.00$  indicates "poor" agreement,  $0.00-0.20$  "slight,"  $0.21-0.40$  "fair,"  $0.41-0.60$  "moderate,"  $0.61-0.80$  "substantial," and  $> 0.80$  "almost perfect."

Cohen's  $\kappa$  was deliberately excluded from our evaluation because it offers little independent information beyond existing metrics based on the confusion matrix in the context of extreme class imbalance. Our study employs the  $F_2$  score to capture threshold-specific recall-weighted performance, the MCC for chance-corrected balance, ROC-AUC and PR-AUC for threshold-agnostic discrimination, and the H-measure for cost-sensitive integration. Since  $\kappa$  and MCC both correct for class-marginal effects and in practice produce virtually identical classifier rankings on ultra-skewed data, including  $\kappa$  would have been redundant and risked obscuring the clarity of our comparative analysis.

### 1.2.3. Contribution of the Study

Prior work has highlighted metric pitfalls in imbalanced learning [19, 23, 29], arguing that PR-AUC are more informative than ROC on skewed data and that MCC is preferable to accuracy and F1; some even propose replacing ROC-AUC outright with MCC. Yet, we still lack a statistically rigorous, cross-domain comparison of ROC-AUC against MCC,  $F_2$ , PR-AUC, and the H-measure under ultra-imbalanced prevalence and realistic resampling pipelines. Saito and Rehmsmeier [19] focus on PRC vs. ROC in imbalance but do not examine MCC or H-measure or analyze sampler-classifier pipelines. Chicco and Jurman [23] advocate MCC over accuracy and F1 but do not position MCC against ROC-AUC, PR-AUC, and H-measure in rare-event scenarios or under resampling. Chicco and Jurman [29] argue that the MCC vs. ROC-AUC conceptually, without providing an empirical, multi-metric treatment that combines rare-event settings, resampling, and statistical testing. Accordingly, our contribution departs from prior MCC-centered analyses by formalizing a prescriptive, statistically validated multi-metric evaluation protocol rather than advocating a single replacement metric. Meanwhile, Richardson et al. [30] reopen the debate by contending that ROC-AUC remains robust under imbalance while PR-AUC is prevalence-sensitive, an observation that, rather than endorsing a single "best" metric, motivates a multi-metric protocol that separates ranking from cost and thresholding. In response, we present a multi-metric, cost-aligned package that combines MCC and  $F_2$  as primary threshold-dependent metrics with PR-AUC for threshold-free ranking and the H-measure for principled cost integration. This framework is evaluated in ultra-imbalanced ( $< 3\%$ )

regimes across resampling pipelines and supported by bootstrap confidence intervals, DeLong's tests, Kendall's  $\tau$ , and Friedman–Nemenyi analyses. We also include a prevalence-sensitivity check, showing how shifts in class balance affect PR-AUC and the H-measure on our datasets, which reinforces the need to report complementary metrics rather than relying on ROC-AUC alone.

The proposed multi-metric protocol is a prescriptive evaluation framework rather than an informal guideline. It specifies explicit decision rules: MCC and  $F_2$  serve as primary metrics for thresholded, cost-asymmetric deployment; PR-AUC assesses threshold-free ranking; and the H-measure integrates cost assumptions explicitly. Model selection is accepted only when supported by rank concordance (Kendall's  $\tau$ ) and statistical significance testing, preventing decisions based on isolated metric improvements. Empirically, we demonstrate that this protocol yields distinct and operationally superior model choices compared to conventional ROC-AUC-centric reporting.

Importantly, the novelty of this work lies not in proposing new metrics but in establishing a unified, statistically rigorous evaluation protocol that jointly analyzes metric disagreement, cost alignment, and resampling effects under ultra-imbalanced conditions, an aspect largely absent from prior imbalanced-learning studies.

To the best of our knowledge, no prior study has combined cross-domain rare-event datasets, multiple resampling strategies, and formal rank- and significance-based comparisons to assess how different metric families can lead to contradictory model-selection conclusions.

To address this gap, we conduct a cross-domain analysis of 20 classifier-sampler configurations per dataset. Specifically, we evaluate four classifiers, Logistic Regression (LR), Random Forest (RF), XGBoost (XGB), and CatBoost (CB), combined with five sampling strategies, namely no oversampling, SMOTE, Borderline SMOTE, SVM SMOTE, and ADASYN. The analysis is performed on three rare event benchmarks: credit card fraud detection with 0.17% positives, yeast protein localization with 1.35% positives, and ozone exceedance with approximately 3% positives.

Our contributions are fourfold:

Empirical characterization of ROC-AUC in rare-event regimes. We quantify ceiling effects and demonstrate that ROC-AUC can overstate model quality, as it remains relatively insensitive to operationally costly misclassifications when the prevalence is less than 3%, even as false positives and false negatives vary substantially across pipelines.

A pragmatic, cost-aware multi-metric alternative. Using Kendall's  $\tau$  rankings and paired significance testing, we demonstrate that MCC and  $F_2$  more accurately reflect asymmetric error costs and deployment priorities. In contrast, PR-AUC (threshold-free ranking) and the H-measure (principled cost weighting) provide complementary views. We distill this into a portable reporting protocol: use MCC and  $F_2$  as primary metrics, with PR-AUC and H-measure as complementary metrics; report ROC-AUC only with explicit caveats in ultra-imbalanced settings. Specifically, MCC and  $F_2$  are prioritized when deployment involves fixed decision thresholds and asymmetric error costs; candidate models are compared at operationally relevant thresholds using these metrics, with selections accepted only when supported by rank consistency and statistical significance. In practice, metric prioritization within the protocol should reflect domain-specific risk asymmetries:  $F_2$  is emphasized when false negatives are substantially more costly than false positives (e.g., safety or fraud detection), whereas MCC provides a balanced summary when both error types matter under constrained review capacity. The deployment context, such as real-time decision-making versus offline triage, should guide threshold selection rather than relying on metric substitution.

Statistically rigorous, resampling-aware evaluation. We pair model selection with robust inference using stratified bootstrap confidence intervals, DeLong's tests for ROC-AUC, Kendall's  $\tau$  for rank concordance, and Friedman–Nemenyi critical-difference analysis, to reveal practically meaningful differences that ROC-AUC alone can mask.

Actionable guidance for practitioners and researchers. We provide a replicable framework for evaluating classifiers under extreme class imbalance that integrates threshold-dependent (MCC and  $F_2$ ) and threshold-free (PR-AUC and H-measure) metrics, aligns with stakeholder cost asymmetries, and transfers across domains (finance, bioinformatics, environmental monitoring).

These contributions move beyond single-metric advocacy toward a multi-metric, cost-aligned evaluation protocol empirically validated in ultra-imbalanced, real-world scenarios.

## 2. Related Work

A quantitative comparison of performance metrics has attracted sustained interest because the choice of metric can significantly alter scientific conclusions and deployment decisions. Early empirical surveys [31, 32] catalogued divergences among accuracy, ROC-based, and cost-based criteria, noting that overall accuracy, defined in 18,

$$ACC = \frac{TP + TN}{N_1 + N_0}, \quad (18)$$

is dominated by the majority class when  $\pi = N_1/(N_1 + N_0) \ll 0.5$ .

For instance, Hanley and McNeil [17] justified ROC-AUC as the probability that a randomly chosen positive receives a higher score than a randomly chosen negative. However, Hand [21] demonstrated that AUC implicitly assigns unequal, prevalence-dependent misclassification costs, which violates many decision contexts. Building on that critique, Davis and Goadrich [20] derived the monotone transformation that maps any ROC point (FPR, TPR) to (R, P) space and proved that the PR curve dominates ROC when  $\pi < 0.2$ . Saito and Rehmsmeier [19] confirmed the theoretical claim using biomedical data, where the ROC-AUC varied by less than 0.02, while the PR-AUC varied by more than 0.50 for the same algorithms. Very recent work has reopened the debate. Richardson et al. [30] contend, via simulation and an epitope-prediction case study, that ROC-AUC remains robust to imbalance, whereas PR-AUC “over-penalises” legitimate classifiers. Their critique hinges on the fact that precision is a function of both TPR and prevalence, making PR-AUC sensitive to the sampling of the evaluation set. Conversely, Zhang and Geng [33] demonstrate that PR-AUC’s prevalence sensitivity is a feature, not a bug, when the deployment environment shares the same class skew. The persisting disagreement underscores the need for multi-metric reporting.

The MCC was initially proposed for protein secondary-structure prediction [22]. Chicco and Jurman [23] provided simulations and genomics case studies where MCC ranked classifiers more consistently with domain utility than  $F_1$  or balanced accuracy. In 2023, these authors argued that MCC should replace ROC-AUC as the “standard statistic” for binary classification, citing ROC’s hidden cost bias and MCC’s symmetry. Itaya et al. [34] derived asymptotic confidence intervals for single and paired MCC estimates, enabling formal hypothesis testing between classifiers.

Elkan [1] formalized expected cost (EC) as shown in 19:

$$EC(t) = C_{FN} \frac{FN(t)}{N} + C_{FP} \frac{FP(t)}{N},$$

arguing that threshold choice must minimize EC under a user-supplied cost matrix. Hernández-Orallo [35] extended ROC analysis to dominance curves, constructing the convex hull of cost points to identify potentially optimal classifiers under all cost/prevalence pairs. Hand [36] proposed the H-measure, which integrates EC over a beta-distributed cost parameter to mitigate the AUC’s hidden-cost flaw. While discrimination metrics assess ranking, Niculescu-Mizil and Caruana [37] compared log-loss, defined in 20, with the Brier score and AUC, showing that well-calibrated probabilities can be critical for cost-sensitive decisions even when AUC values are identical:

$$L = -\frac{1}{N} \sum_{i=1}^N [y_i \log \hat{p}_i + (1 - y_i) \log(1 - \hat{p}_i)],$$

Flach and Kull [38] further decomposed log-loss into calibration and refinement components, providing diagnostic insight complementary to ROC analysis. He and Garcia [39] reviewed algorithmic and evaluation issues in imbalanced learning, recommending PR-AUC and G-mean. More recently, Blagus and Lusa [40] demonstrated that  $F_\beta$  with  $\beta > 1$  is preferable to  $F_1$  for rare disease prediction, and Imani et al. [41, 42] examined how varying class-imbalance ratios affect classifier performance and the apparent efficacy of resampling (e.g., SMOTE and its variants), evaluating both threshold-dependent and threshold-free metrics, including ROC-AUC, PR-AUC, MCC,  $F_1$ -score, and Cohen’s  $\kappa$ . Complementing these findings, a comprehensive review of churn prediction reports that

the ROC-AUC remains one of the most commonly reported metrics in practice, reflecting established reporting conventions in the literature [43]. This persistence motivates providing clearer guidance on metric selection in the context of class imbalance.

### 3. Datasets

This study evaluates classifier performance on three publicly available benchmark datasets that exhibit extreme class imbalance (see Table 2). The datasets are all real-world operational datasets commonly used as benchmarks in imbalanced learning research. Importantly, they span three orders of magnitude in minority-class prevalence (0.17%, 1.35%, and approximately 3%), enabling a controlled examination of metric behavior across varying degrees of extreme imbalance. Before modelling, all features were standardized to have a zero mean and unit variance. As is common in many operational settings, no instance-level cost annotations were available. Details of each dataset are provided in subsections 3.1–3.3.

**European Credit-card Fraud Detection:** This widely studied dataset comprises 284807 card-transaction records, of which 492 are confirmed frauds (imbalance rate  $\approx 0.17\%$ ). Each observation is represented by 28 principal components derived from the original monetary attributes.

**UCI Yeast Protein Localization (“POX”):** The UCI Yeast benchmark contains 1484 protein sequences described by eight physicochemical descriptors. The minority class “POX” (indicating peroxisomal localization) appears in only 20 instances ( $\approx 1.35\%$ ).

**UCI Ozone Level Detection:** This dataset consists of 2536 hourly measurements of atmospheric conditions, each with 73 features, and 57 recorded ozone-exceedance events ( $\approx 3\%$ ).

**Table 2.** The datasets specifications.

Dataset	Positives	Negatives	Imbalance Rate	Features	Source
Credit-card Fraud	492	284315	0.17%	28 PCA-obscured transaction attributes	Kaggle [44]
Yeast Protein Localisation	20	1464	1.35%	8 physicochemical descriptors	UCI repository [45]
Ozone Level Detection	57	1791	3.00%	72 atmospheric covariates	UCI repository [46]

#### 3.1. Fraud Dataset (Credit-Card Fraud Detection)

The Credit-Card Fraud Detection dataset comprises 284807 credit-card transactions made by European cardholders over a two-day period in September 2013 [43]. Only 492 transactions are labelled as fraudulent, representing approximately 0.17% of the data. Each record consists of 31 features: 28 anonymized principal components (V1–V28), obtained via Principal Component Analysis (PCA) to preserve confidentiality, a “Time” feature (seconds elapsed since the first transaction), and an “Amount” feature (transaction value in Euros). The target column “Class” takes the value 0 for legitimate transactions and 1 for fraud. Because PCA was used to mask the original feature identities, V1–V28 do not have explicit semantic meaning; nevertheless, they capture the essential structure that distinguishes fraudulent from legitimate activity [44]. The extreme rarity of fraud events underscores why this dataset is a standard benchmark in imbalanced-learning research [19].

#### 3.2. Yeast Dataset (UCI version; POX Subset)

The Yeast dataset is available from the UCI Machine Learning Repository [45]. Initially, each of the 1484 instances and 11 features describes characteristics of proteins via eight continuous features and a single “Localization” label that can take one of ten categories (CYT, NUC, MIT, ME1, ME2, ME3, EXC, VAC, POX, ERL). Of particular interest is the POX class, which appears only 20 times ( $\approx 1.35\%$ ) out of 1484 instances [45]. Since the UCI version encodes localization as a single categorical

field, we extract the POX cases and recode them as the positive class (1), with all other localizations merged into a single negative class (0). This one-versus-all transformation yields a binary problem with POX representing 1.35% of instances and therefore serving as a rare-class benchmark in line with prior literature [47].

### 3.3. Ozone Dataset

The one-hour ozone exceedance dataset originates from the UCI Machine Learning Repository's Air Quality Evaluation collection [46], which captures 2536 timed atmospheric chemistry and meteorology observations recorded at an urban monitoring station in California across multiple summer seasons. Each record includes a suite of 73 continuous features, such as temperature, barometric pressure, wind speed, relative humidity, and concentrations of nitrogen oxides (NO<sub>x</sub>), hydrogen cyanide (HCN), and volatile organic compounds, paired with a binary target indicating whether the one-hour average ozone (O<sub>3</sub>) concentration exceeded the regulatory threshold of 0.08 ppm.

Following preprocessing, entries with missing values are discarded, and the original timestamp column is removed to focus purely on predictive measurements. To conform with prevalent imbalance-learning benchmarks, a literature-standard split retains all 57 confirmed class 1 instances and randomly selects the corresponding 1791 class 0 instances, producing an extreme imbalance ratio of 31.4:1. Before model training, all features are standardized to zero mean and unit variance, and the data are partitioned into stratified training and testing subsets that preserve the native class proportions. The pronounced skew toward non-exceedance cases highlights the challenges of predicting rare events. This dataset is particularly suitable for evaluating sampling techniques, ensemble methods, and cost-sensitive learning algorithms.

### 3.4. Data Preprocessing

Data preprocessing was conducted uniformly across all three datasets to facilitate fair comparison under severe class imbalance. To prevent label leakage, the target label was excluded from the test set's feature matrix for each dataset. All preprocessing and resampling steps were fitted on the training split only, and the resulting fitted preprocessors (e.g., scalers, encoders) were then applied to the test features; no statistics were re-estimated on the test data, and the test labels were used solely for final evaluation. In the Credit Card Fraud dataset, the "Time" feature was excluded to avoid confounding with temporal dependence under static resampling. Transaction amounts exhibited extreme right skew and were log-transformed as shown in 21 [48]:

$$Amount_{log} = \ln(1 + Amount) \quad (1)$$

Subsequently, all 29 features were standardized to zero mean and unit variance [49]. A stratified train-test split (70% / 30%) preserved the original 0.17% fraud ratio, with oversampling techniques applied exclusively to the training subset [44].

The multi-class localization labels were recoded into a binary target in the UCI Yeast dataset, as defined in (22), assigning label 1 to the "POX" class and 0 to all other localizations:

$$y = \begin{cases} 1, & \text{if } Localization = POX, \\ 0, & \text{otherwise.} \end{cases}$$

This yielded 20 positive and 1464 negative instances ( $\approx 1.35\%$ ). The eight continuous features were standardized via z-score normalization to ensure equal weighting in distance-based sampling. A stratified train-test split (70% / 30%) was performed, maintaining the class ratio in both sets, with oversampling restricted to the training partition [45].

The UCI Ozone Level Detection dataset underwent complete-case analysis, where any record containing missing values was removed, resulting in a sample size reduction from 2536 to 1848 observations. The timestamp column was dropped to prevent non-numeric data from affecting subsequent analyses. A literature-standard subset was constructed by retaining all 57 positive instances and randomly sampling 1791 negatives to achieve an imbalance ratio of 31.4:1. All 73

features were standardized to zero mean and unit variance. Finally, a stratified split (70% training, 30% testing) preserved the class distribution, with oversampling applied solely to the training set.

## 4. Methods

Our goal was to examine how alternative evaluation metrics behave when the minority class is vanishingly rare and to test whether a small, theory-driven set of metrics can remain intact across disparate application areas. We therefore built a deliberately symmetrical experimental protocol: one script, three datasets, twenty classifier–sampler variants, one set of metrics, and one statistical lens. This section walks through each step.

### 4.1. Classifier-Sampling Framework

All experiments were conducted in Python 3.11 using scikit-learn 1.5, imbalanced-learn 0.12, CatBoost 1.3, and XGBoost 2.0. We evaluated four classifiers, logistic regression with  $L_2$  regularization, Random Forest with 300 trees, XGBoost with depth = 6 and learning rate  $\eta = 0.1$ , and CatBoost with 500 iterations, each under one baseline (no oversampling) and four oversampling techniques (SMOTE, Borderline-SMOTE, SVM-SMOTE, and ADASYN). This yields  $(1 + 4) \times 4 = 20$  unique classifier–sampler configurations per dataset. We applied these 20 configurations independently to each of the three datasets, for a total of 60 model evaluations.

Performance for each configuration was estimated via stratified 10-fold cross-validation. Oversampling was applied only to the training set within each fold to avoid information leakage. No resampling or parameter fitting was performed on test data. We recorded ROC-AUC, PR-AUC, the H-measure, MCC,  $F_2$ -score, and the raw confusion-matrix counts from every run.

For threshold-dependent metrics (MCC and  $F_2$ ), predictions are evaluated at fixed, operationally relevant decision thresholds that are held constant across models and folds, rather than optimized per metric or per fold, to reflect realistic deployment conditions. The H-measure is computed using the standard Beta(2,2) cost distribution, which represents a neutral prior over relative misclassification costs and is commonly used when domain-specific cost information is unavailable. This choice allows cost-aware comparison without imposing asymmetric assumptions that would vary across application domains.

To ensure reproducibility and to avoid implicit optimization bias, all classifiers and resampling methods were configured using standard, literature-consistent hyperparameters, without dataset-specific tuning. The objective of this study is to analyze metric behavior rather than maximize predictive performance.

Logistic Regression employed  $L_2$  regularization with the default inverse regularization strength ( $C = 1.0$ ) and the liblinear solver. Random Forest models were trained using 300 trees, Gini impurity, and the default feature subsampling method. XGBoost used a maximum tree depth of 6, learning rate  $\eta = 0.1$ , and 300 boosting rounds. CatBoost models were trained for 500 iterations with default regularization and symmetric trees.

All resampling techniques were implemented using the *imbalanced-learn* library with standard parameter settings. SMOTE-based methods employed  $k = 5$  nearest neighbors. Borderline-SMOTE used the “borderline-1” strategy. SVM-SMOTE generated samples near support vectors identified by a linear SVM. ADASYN adaptively generated synthetic samples based on local class imbalance using its default neighborhood size.

### 4.2. Rank-correlation Analysis Between Metrics

To quantify how consistently the candidate metrics rank the 20 classifier-sampler configurations, we computed the pair-wise Kendall rank correlation coefficient  $\tau$  ([50]) for every metric pair  $(m, n) \in \{\text{PR} - \text{AUC}, \text{ROC} - \text{AUC}, \text{MCC}, F_2, \text{H} - \text{measure}\}$ . Let  $x_{im}$  and  $x_{in}$  denote the values of metrics  $m$  and  $n$  for configuration  $i$  ( $i = 1, \dots, N$ ,  $N = 20$ ).

Kendall’s statistic is defined in 23:

$$\tau_{mn} = \frac{\sum_{i < j} \text{sgn}(x_{im} - x_{jm}) \text{sgn}(x_{in} - x_{jn})}{\binom{N}{2}},$$

and the sign function is given in 24:

$$\text{sgn}(u) = \begin{cases} 1, & u > 0 \\ 0, & u = 0 \\ -1, & u < 0 \end{cases}, \quad (24)$$

The numerator in (23) counts concordant minus discordant configuration pairs, while the denominator represents the total number of unordered pairs. For illustration, when  $N=20$ , Kendall's statistic reduces to 25:

$$\tau = \frac{\#concordant - \#discordant}{\binom{20}{2}}, \quad (25)$$

Kendall's rank-correlation coefficient ( $\tau$ ) is preferred to the Pearson product-moment coefficient ( $r$ ) for assessing agreement among evaluation metrics in highly imbalanced learning because it aligns with the methodological aim, which is comparing metric-induced rankings rather than raw magnitudes. Kendall  $\tau$  is a non-parametric statistic that depends only on the ordering of observations; it remains invariant under any strictly monotone transformation of the metric scores and is therefore insensitive to the heterogeneous, bounded scales of MCC (-1 ... 1), PR-AUC (0 ... 1), and ROC-AUC (0 ... 1) [51]. Unlike  $r$ , which assumes joint normality and homoscedasticity,  $\tau$  makes no distributional assumptions and is robust to the heavy skew, ceiling effects, and frequent ties (e.g., TP = 0, FP = 0  $\rightarrow$  identical MCC) that characterise rare-event experiments. Furthermore,  $\tau$  admits an intuitive probabilistic interpretation ( $\tau = 0.60$  implies 80% concordant versus 20% discordant pairs), facilitating substantive discussion of metric concordance. These properties render Kendall's coefficient a statistically reliable and conceptually faithful measure for ranking-consistency studies under extreme class imbalance [50, 52].

#### 4.3. Statistical Testing and Confidence Intervals

We generated 95% confidence intervals (CIs) for each metric using a stratified bootstrap (2000 replicates per test fold, preserving class prevalence). ROC-AUC differences between any two classifier-sampler configurations were evaluated with the paired-sample DeLong test ([53]), and p-values are reported. When comparing more than two methods, we applied the Friedman aligned-ranks test followed by the Nemenyi Critical-Difference (CD) procedure ( $\alpha = 0.05$ ).

## 5. Results and Discussions

This section presents a detailed empirical investigation into the performance of twenty classifier-sampler configurations across three highly imbalanced datasets: credit card fraud detection, Yeast protein localization, and Ozone level detection. Under extreme class imbalance, the primary objective is to examine the sensitivity and reliability of five evaluation metrics: ROC-AUC, PR-AUC,  $F_2$ -score, MCC, and H-measure. Rather than comparing classifiers per se, the focus lies on understanding how each metric responds to variations in false positives and false negatives induced by different sampling techniques. Results highlight notable inconsistencies in ROC-AUC's ability to reflect practical performance costs, whereas alternative metrics demonstrate more substantial alignment with operational realities and domain expert expectations.

### 5.1. Detailed Per-dataset Results (Fraud Dataset)

This section compares the behaviour of twenty classifier-sampler configurations on three datasets that are thematically unrelated yet similarly skewed: the credit-card fraud dataset, the Yeast protein localization dataset, and the Ozone Level Detection dataset.

#### 5.1.1. Fraud Dataset Results

Table 3 presents the results of 20 distinct classifier–sampler configurations, including the corresponding confusion matrix components and five evaluation metrics. All evaluations were conducted on the test set (unseen data) of the Fraud dataset. Since the study’s objective is metric evaluation, not model comparison, we examine how each metric responds to the dramatic swings in FP and FN counts that arise under extreme class-imbalance. The empirical evaluation conducted on the Fraud dataset demonstrates clearly the limitations inherent in relying on ROC-AUC as an evaluation metric for rare-event binary classification tasks. Although ROC-AUC scores across various classifiers and sampling methods remain consistently high, a deeper inspection of the performance using alternative metrics reveals significant shortcomings in ROC-AUC’s reliability for highly imbalanced datasets.

Taking the Logistic Regression classifier with ADASYN sampling as a notable example, the ROC-AUC score is observed to be impressively high at 0.968. However, this apparently robust performance contrasts with extremely poor values for other critical metrics: an  $F_2$ -score of just 0.09, MCC of 0.126, and an H-measure of 0.638. Further exacerbating this discrepancy is the notably large number of false positives (FP = 6595), which clearly illustrates that the ROC-AUC cannot adequately penalize the misclassification of negative class instances. Similarly, another striking contradiction is observed when examining Logistic Regression with SMOTE sampling. Despite achieving a high ROC-AUC score of 0.968, this combination demonstrates poor  $F_2$  (0.237), MCC (0.227), and H-measure (0.651) scores, which are further compounded by an extremely high false positive rate (FP = 2019). This trend persists across multiple combinations, highlighting ROC-AUC’s inability to reflect meaningful performance deficiencies in classifiers when dealing with highly imbalanced datasets.

The inconsistency in performance, as indicated by the ROC-AUC compared to more practically relevant metrics, is further exemplified by the Logistic Regression classifier combined with Borderline-SMOTE sampling, which achieves an acceptable ROC-AUC score of 0.935. Nonetheless, substantial performance issues arise, as clearly evidenced by an  $F_2$ -score of 0.448, an MCC of 0.360, and an H-measure of 0.558, coupled with a high false positive count (FP = 647). These results underscore the critical failure of ROC-AUC in accurately capturing and penalizing the actual misclassification costs associated with rare-event classes. We emphasize that the behavior observed for logistic regression under aggressive oversampling is not a failure of the classifier per se, but an expected consequence of training a linear model on heavily resampled data at extreme prevalence; the key finding is that ROC-AUC remains insensitive to the resulting explosion in false positives, whereas MCC,  $F_2$ , PR-AUC, and the H-measure appropriately penalize such configurations.

**Table 3.** The results on the credit-card fraud dataset

Sampling+Model	ROC-AUC	PR-AUC	F2	MCC	H	FP	FN	TP	TN	
Baseline	RF	0.934	0.821	0.796	0.855	0.761	5	35	113	85290
	LR	0.953	0.708	0.658	0.731	0.574	15	56	92	85280
	XGB	0.969	0.84	0.787	0.84	0.756	8	36	112	85287
	CB	0.971	0.836	0.805	0.856	0.749	7	33	115	85288
SMOTE	RF	0.966	0.819	0.793	0.831	0.727	13	34	114	85282
	LR	0.968	0.705	0.237	0.227	0.651	2019	18	130	83276
	XGB	0.976	0.836	0.814	0.8	0.752	35	26	122	85260
	CB	0.969	0.823	0.783	0.733	0.706	65	26	122	85230
Borderline-SMOTE	RF	0.93	0.818	0.795	0.851	0.732	6	35	113	85289
	LR	0.935	0.67	0.448	0.36	0.558	647	26	122	84648
	XGB	0.976	0.823	0.82	0.845	0.727	15	29	119	85280
	CB	0.959	0.815	0.787	0.756	0.702	50	28	120	85245
SVM-SMOTE	RF	0.933	0.827	0.828	0.843	0.739	18	27	121	85277
	LR	0.936	0.687	0.522	0.416	0.587	481	23	125	84814
	XGB	0.971	0.834	0.817	0.827	0.751	22	28	120	85273
	CB	0.954	0.827	0.804	0.768	0.751	50	25	123	85245

ADASYN	RF	0.96	0.822	0.783	0.814	0.727	17	35	113	85278
	LR	0.968	0.711	0.09	0.126	0.638	6595	16	132	78700
	XGB	0.972	0.827	0.814	0.8	0.717	35	26	122	85260
	CB	0.969	0.807	0.766	0.713	0.698	71	28	120	85224

Conversely, metrics such as MCC,  $F_2$ , and H-measure exhibit greater consistency in identifying performance inadequacies, effectively distinguishing between well-performing and poorly performing models. For instance, the baseline Random Forest classifier achieves strong, stable performance across MCC (0.855),  $F_2$  (0.796), and H-measure (0.761) with low FP (5), clearly indicative of genuine classification effectiveness.

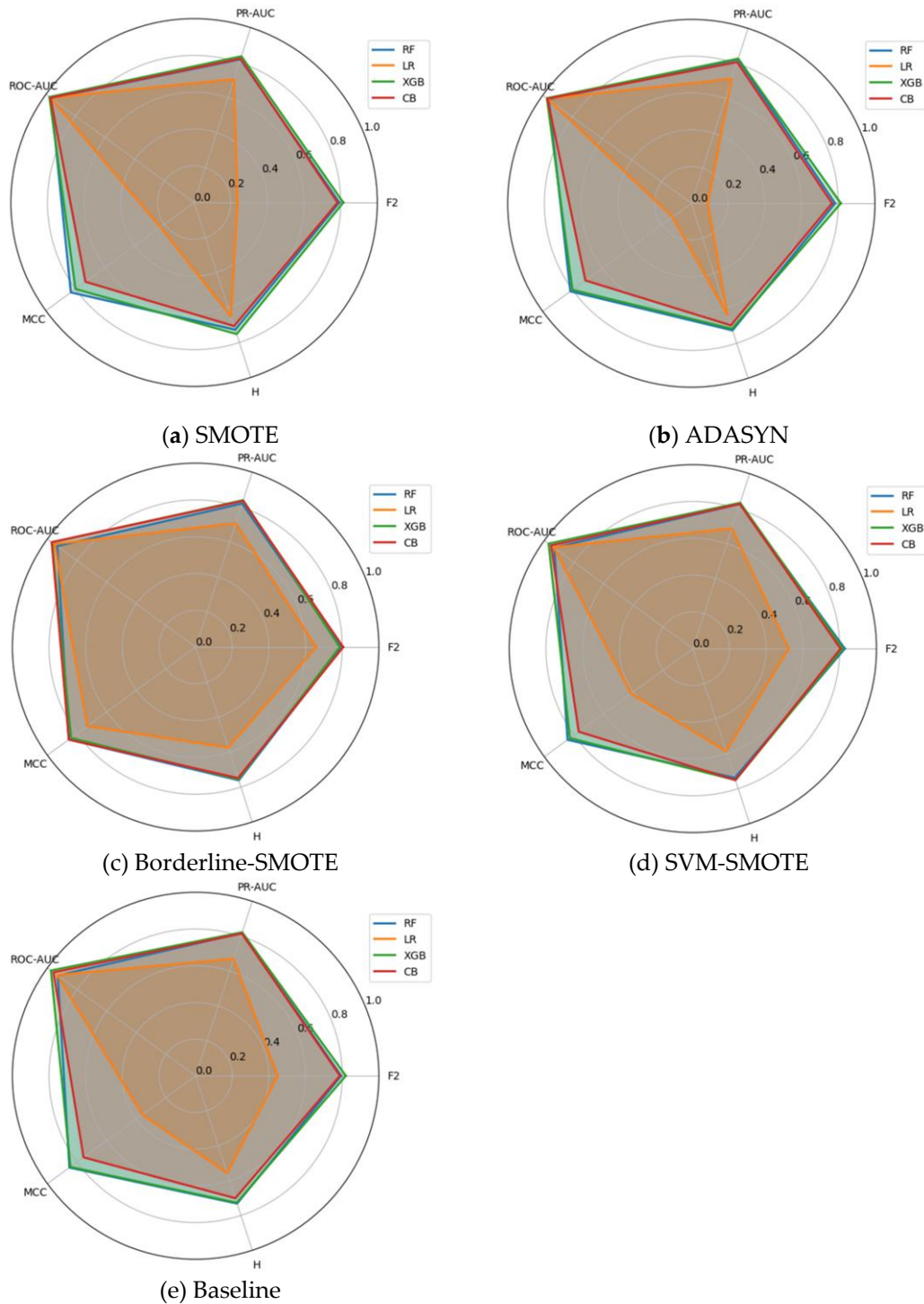
In summary, the empirical evidence firmly establishes that, despite its widespread use, ROC-AUC often provides an overly optimistic and misleading assessment of classifier performance in highly imbalanced contexts. Alternative metrics, specifically MCC,  $F_2$ , and H-measure, are more effective and accurate indicators of genuine predictive performance and should be preferred in evaluation methodologies involving rare-event classification.

Table 4 summarizes the analysis conducted on the Fraud dataset, encapsulating the observed performance ranges, sensitivity to variations in false positives and false negatives, and key observations for ROC-AUC, PR-AUC,  $F_2$ -score, MCC, and H-measure. This comparative overview highlights significant discrepancies between ROC-AUC and alternative metrics, underscoring ROC-AUC's limited sensitivity to misclassification costs in highly imbalanced datasets.

**Table 4.** The summary of the analysis on the fraud dataset

Metric	Observed range	Sensitivity to FP/FN variations	Key observations
ROC-AUC	0.930 – 0.976 ( $\Delta \approx 0.046$ )	Minimal. ROC-AUC uses the empirical FPR denominator ( $\approx 85$ k) and therefore changes by $< 0.01$ when FP rises from 5 (Random Forest baseline) to 6595 (LR+ADASYN).	Logistic Regression baseline vs. LR+SMOTE: FP increases by a factor of 135 (from 15 to 2019) yet ROC-AUC increases (from 0.953 to 0.968).
PR-AUC	0.669 – 0.839 ( $\Delta \approx 0.17$ )	Moderate. Precision penalises each additional FP, so PR-AUC drops from 0.821 to 0.704 when Logistic Regression baseline is oversampled with SMOTE (FP increases from 15 to 2019). However, the metric is threshold-free and does not reflect the absolute alarm burden in the deployed cut-off.	CatBoost baseline (FP = 7) vs. CB+ADASYN (FP = 71): PR-AUC falls from 0.836 to 0.807, a visible but still modest decline given the ten-fold FP increase.
$F_2$	0.000 – 0.827 ( $\Delta \approx 0.83$ )	High. By quadrupling recall weight, $F_2$ rewards FP-heavy configurations if they gain enough TP, but collapses when precision implodes. LR+SMOTE attains the highest TP (130) and lowest FN (18) yet $F_2 = 0.237$ —demonstrating severe precision penalty.	Random Forest baseline (TP = 113, FP = 5) vs. RF+ADASYN (TP = 113, FP = 17): identical recall, FP increases by a factor of 3.4, $F_2$ drops from 0.795 to 0.782.
MCC	0.125 – 0.855 ( $\Delta \approx 0.73$ )	Very high and symmetric. MCC falls almost linearly with either FP or FN explosions. It ranks LR+SMOTE (MCC = 0.227) and LR+ADASYN (0.126) near the bottom despite top-quartile ROC-AUC values, exposing their high alarm costs.	MCC and $F_2$ exhibit Kendall $\tau \approx 0.90$ across the grid, confirming consistent ordering once a threshold is fixed.

Complementing the scalar summaries in Table 4, a concise cross-metric visualization aids interpretation. Figures 1 (a–e) provide small-multiples radar plots that compare five evaluation criteria, including  $F_2$ , H-measure, MCC, ROC-AUC, and PR-AUC, for Random Forest, Logistic Regression, XGBoost, and CatBoost under each resampling strategy. Axes are fixed across panels and scaled to  $[0, 1]$ ; polygons report fold-wise means. The purpose is illustrative: to visualize pattern and separation across metrics, complementing the confidence-interval and rank-based analyses reported later.



**Figure 1. (a–e).** Multi-metric radar plots for the fraud dataset across five resampling strategies. Axes are scaled to  $[0, 1]$ ; polygons show mean performance across cross-validation folds, with larger and more uniformly shaped polygons indicating stronger overall performance across the metric bundle.

Two consistent regularities are apparent across all samplers. First, ROC-AUC exhibits a ceiling effect: for every classifier and sampler, the ROC-AUC spoke lies close to the outer ring, producing minimal model separation. Second, the threshold-dependent/cost-aligned metrics (MCC and  $F_2$ ) reveal substantial differences that the ROC-AUC curve masks. In particular, Logistic Regression deteriorates sharply under synthetic minority schemes: under SMOTE and ADASYN, the Logistic Regression polygon collapses on the MCC and  $F_2$  axes, while remaining near-maximal on ROC-AUC, indicating a severe precision loss (inflated false positives) that does not materially affect rank-based AUC. The tree/boosted models (Random Forest, XGBoost, CatBoost) remain comparatively stable in terms of MCC and  $F_2$  across samplers, with XGBoost and Random Forest typically forming the largest polygons (i.e., the strongest across the bundle).

The Baseline panel serves as a reference: ensembles dominate on MCC and  $F_2$ , while Logistic Regression exhibits weaker performance but does not collapse. Moving to SMOTE and ADASYN, the Logistic Regression degradation intensifies (MCC and  $F_2$  shrink markedly) even though PR-AUC and H-measure decline only moderately, and ROC-AUC stays saturated. This pattern is consistent with decision-boundary distortion and score miscalibration induced by aggressive oversampling at a prevalence of 0.17%, which disproportionately inflates false positives at practically relevant thresholds. Borderline-SMOTE and SVM-SMOTE exhibit the same qualitative behavior, albeit with milder Logistic Regression degradation; ensembles, however, retain broad, well-rounded polygons, reflecting their robustness to these resampling variants.

Taken together, the radars visualize the complementarity within the proposed metric bundle. PR-AUC and H-measure track the MCC and  $F_2$  separations (though less dramatically), reinforcing their role as threshold-free and cost-sensitive companions, respectively. Conversely, the near-constant ROC-AUC across panels underscores its limited diagnostic value in this ultra-imbalanced setting. These visual regularities align with our Kendall- $\tau$  concordance results (strong agreement among MCC,  $F_2$ , H-measure, and PR-AUC; weak with ROC-AUC) and the critical-difference rankings that favor tree/boosted models. We therefore use the radars as an intuitive summary of sampler-classifier interactions and as corroborating evidence for the central claim: relying solely on ROC-AUC can misrepresent practical performance, whereas a multi-metric, cost-aligned protocol reveals operationally meaningful differences.

### 5.1.2. Yeast Dataset Results

Table 5 presents the results of 20 distinct classifier-sampler configurations, including the corresponding confusion matrix components and five evaluation metrics. All evaluations were conducted on the test set (unseen data) of the Yeast dataset. The empirical evaluation conducted on the Yeast dataset clearly demonstrates the limitations inherent in relying on ROC-AUC as an evaluation metric for rare-event binary classification tasks. Although ROC-AUC scores across various classifiers and sampling methods frequently appear stable or relatively high, a deeper analysis using alternative metrics reveals significant shortcomings in the reliability of ROC-AUC for highly imbalanced datasets.

**Table 5.** The results on the Yeast dataset

Sampling+Model		ROC-AUC	PR-AUC	$F_2$	MCC	H	FP	FN	TP	TN
Baseline	RF	0.882	0.657	0.536	0.608	0.577	1	3	3	439
	LR	0.899	0.679	0.69	0.727	0.677	1	2	4	439
	XGB	0.966	0.299	0.192	0.283	0.127	1	5	1	439
	CB	0.934	0.722	0.69	0.727	0.572	1	2	4	439
SMOTE	RF	0.85	0.543	0.441	0.377	0.4	7	3	3	433
	LR	0.908	0.679	0.207	0.174	0.677	92	1	5	348
	XGB	0.884	0.375	0.484	0.455	0.353	4	3	3	436
	CB	0.79	0.526	0.455	0.398	0.501	6	3	3	434
	RF	0.917	0.66	0.536	0.608	0.534	1	3	3	439

Borderline-SMOTE	LR	0.888	0.809	0.455	0.351	0.727	26	1	5	414
	XGB	0.902	0.45	0.536	0.608	0.411	1	3	3	439
	CB	0.797	0.6	0.536	0.608	0.53	1	3	3	439
SVM-SMOTE	RF	0.921	0.633	0.69	0.727	0.533	1	2	4	439
	LR	0.888	0.809	0.781	0.717	0.727	3	1	5	437
	XGB	0.897	0.468	0.69	0.727	0.511	1	2	4	439
	CB	0.791	0.599	0.69	0.727	0.53	1	2	4	439
ADASYN	RF	0.918	0.451	0.441	0.377	0.303	7	3	3	433
	LR	0.908	0.679	0.146	0.125	0.677	142	1	5	298
	XGB	0.882	0.582	0.429	0.358	0.505	8	3	3	432
	CB	0.772	0.518	0.417	0.341	0.501	9	3	3	431

For instance, the Logistic Regression classifier combined with SMOTE sampling yields an apparently high ROC-AUC score of 0.908. However, this performance is sharply contradicted by considerably lower scores in crucial alternative metrics, such as  $F_2$  (0.207), MCC (0.174), and H-measure (0.677). The substantial false positive rate observed in this scenario (FP = 92) further highlights the inability of ROC-AUC to reflect the practical costs associated with increased false alarms effectively. Similarly, the XGBoost classifier combined with SMOTE sampling yields a ROC-AUC score of 0.884, which initially appears moderate. However, detailed metrics, including  $F_2$  (0.484), MCC (0.455), and H-measure (0.353), reveal critical weaknesses in performance, particularly when considering that even a relatively modest increase in false positives (FP = 4) can negatively impact the model's practical effectiveness.

Additionally, analysis of the Logistic Regression classifier with ADASYN sampling provides further evidence of ROC-AUC's limitations. Despite maintaining a high ROC-AUC score (0.908), this combination demonstrates poor performance in alternative metrics:  $F_2$  at 0.146, MCC at 0.125, and H-measure at 0.677. Moreover, this classifier configuration suffers from an extremely high false positive count (FP = 142), further underscoring the ROC-AUC's inadequate sensitivity to misclassification costs.

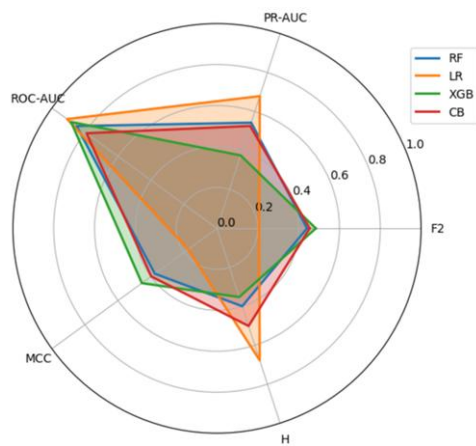
Conversely, metrics such as MCC,  $F_2$ , and H-measure consistently provide a more accurate representation of classifier performance, effectively distinguishing between models that perform well and those that do not. For example, the baseline Random Forest classifier achieves stable and relatively high scores across MCC (0.608),  $F_2$  (0.536), and H-measure (0.577), while maintaining a low false positive count (FP = 1), clearly signalling robust classification capability.

In summary, the empirical evidence from the Yeast dataset clearly demonstrates that ROC-AUC often presents a misleadingly optimistic view of classifier performance in highly imbalanced scenarios. Alternative metrics, such as MCC,  $F_2$ , and H-measure, emerge as more reliable and practically meaningful indicators of model performance in rare-event classification problems. Table 6 summarizes the detailed analysis conducted on the Yeast dataset, presenting the performance range, sensitivity to false positives and false negatives, and key observations for ROC-AUC, PR-AUC,  $F_2$ -score, MCC, and H-measure. This summary clearly highlights the inadequacy of ROC-AUC and supports the practical relevance and greater accuracy of alternative metrics for highly imbalanced datasets.

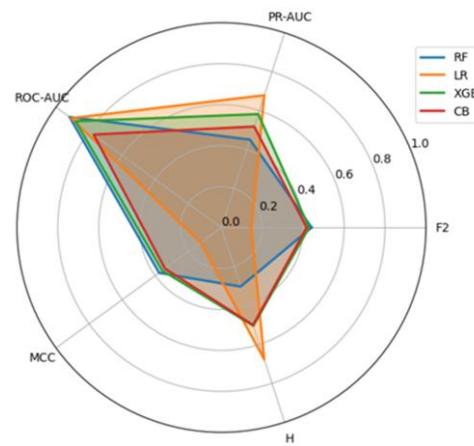
**Table 6.** The summary of the analysis on the Yeast dataset

Metric	Observed range	Sensitivity to FP/FN variations	Key observations
ROC-AUC	0.772 – 0.966 ( $\Delta \approx 0.194$ )	Minimal-to-moderate. Due to the dataset's high negative class size ( $\approx 439$ ), ROC-AUC scores exhibit modest sensitivity despite significant false positives variation.	Logistic Regression baseline (FP=1, ROC-AUC=0.899) vs. LR + ADASYN (FP=142, ROC-AUC=0.908): ROC-AUC slightly increases despite an extreme 142-fold FP rise.

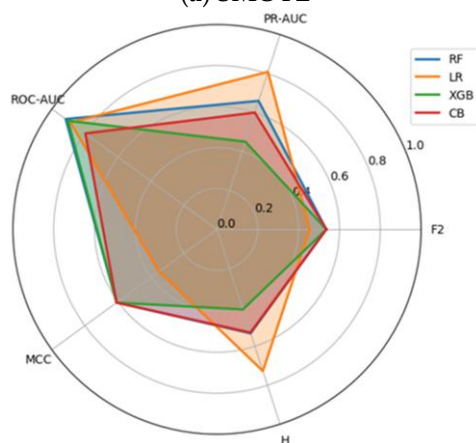
PR-AUC	0.299 – 0.809 ( $\Delta \approx 0.510$ )	Moderate-to-high. Precision directly penalizes false positives, clearly reflecting severe FP increases.	Logistic Regression baseline (FP=1, PR-AUC=0.679) vs. LR+ADASYN (FP=142, PR-AUC=0.679): limited numeric change despite significant FP escalation, indicating threshold-free limitation.
F2	0.146 – 0.781 ( $\Delta \approx 0.635$ )	High. Heavily sensitive to false positives, significantly penalizing classifiers with precision deterioration.	LR+SMOTE: achieves high ROC-AUC (0.908) but very poor F <sub>2</sub> (0.207) due to high FP (92), clearly demonstrating sensitivity to precision collapse.
MCC	0.125 – 0.727 ( $\Delta \approx 0.602$ )	Very high and symmetric. Significantly penalizes both false positives and false negatives, clearly reflecting overall performance deterioration.	LR+ADASYN yields MCC=0.125 despite ROC-AUC=0.908, accurately reflecting severe classification cost due to FP (142). MCC consistently ranks high-FP scenarios lower.



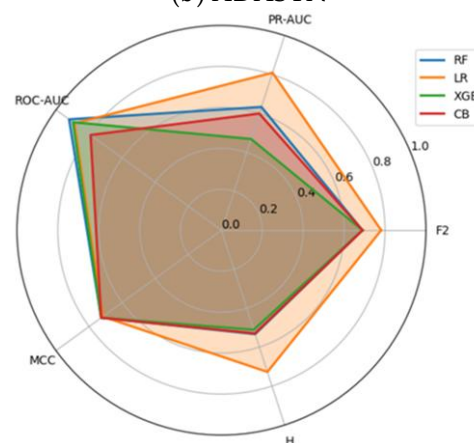
(a) SMOTE



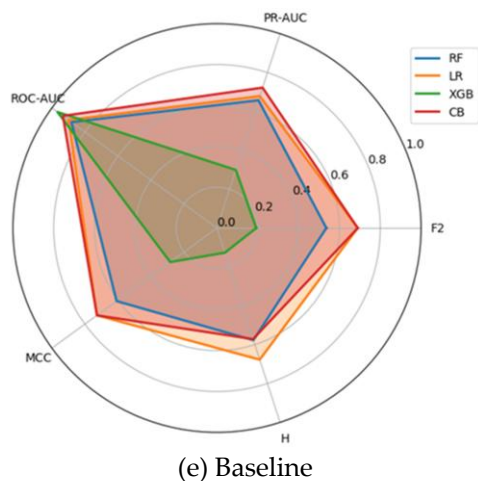
(b) ADASYN



(c) Borderline-SMOTE



(d) SVM-SMOTE



**Figure 2. (a-e).** Multi-metric radar plots for the fraud dataset across five resampling strategies. Axes are scaled to  $[0,1]$ ; polygons show mean performance across cross-validation folds, with larger and more uniformly shaped polygons indicating stronger overall performance across the metric bundle.

In addition to the scalar results in Table 6, a compact cross-metric perspective provides an integrated view. Figures 2 (a–e) present small-multiples radar plots for the Yeast dataset (1.35% positives), comparing F2, H-measure, MCC, ROC-AUC, and PR-AUC for Random Forest, Logistic Regression, XGBoost, and CatBoost under each resampling strategy. Axes are fixed across panels, scaled to  $[0,1]$ , and polygons report fold-wise means. As with the Fraud radars, the goal is illustrative: to visualize patterns and separation across metrics, complementing the following confidence-interval and rank-based analyses.

Two regularities again emerge. First, ROC-AUC remains near the outer ring for all models and samplers, yielding limited separation. Second, threshold-dependent/cost-aligned metrics (MCC and F2) reveal material differences that ROC-AUC alone obscures, with PR-AUC and H-measure generally moving in the same direction, albeit less sharply.

Dataset-specific nuances are notable. In the Baseline panel, XGBoost exhibits a pronounced collapse on F2, MCC, PR-AUC, and H, despite a high ROC-AUC score, an archetypal instance of AUC saturation masking practically relevant errors. CatBoost and Logistic Regression form larger, more rounded polygons, and Random Forest sits in between. Under SMOTE and ADASYN, Logistic Regression exhibits a mixed profile: PR-AUC and H increase substantially, whereas MCC (and at times F2) decreases, indicating that oversampling improves ranking and cost-weighted separation while simultaneously inflating false positives at decision-useful thresholds (score–threshold miscalibration). Borderline-SMOTE mitigates this tension, with milder degradation in Logistic Regression on MCC and F2, and stable ensemble performance. SVM-SMOTE yields the most balanced polygons overall (especially for Logistic Regression and CatBoost), suggesting that margin-aware synthesis can enhance both ranking-based and threshold-dependent metrics on the Yeast dataset.

Taken together, these radars (i) make the ROC-AUC ceiling effect visually explicit; (ii) highlight sampler–classifier interactions that matter operationally (e.g., Baseline XGBoost collapse; Logistic Regression oversampling trade-offs); and (iii) show PR-AUC and H-measure qualitatively tracking the MCC and F2 separations. The visual patterns are consistent with the Kendall- $\tau$  concordance and critical-difference rankings reported for Yeast, reinforcing the central claim that relying solely on ROC-AUC is insufficient. In contrast, a multi-metric, cost-aligned protocol reveals differences of practical consequence.

### 5.1.3. Ozone Dataset Results

Table 7 presents the results of 20 distinct classifier–sampler configurations, including the corresponding confusion matrix components and five evaluation metrics. All evaluations were

conducted on the Ozone dataset's test set (unseen data). The empirical evaluation conducted on the Ozone dataset provides further compelling evidence of the limitations inherent in using ROC-AUC as an evaluation metric for rare-event binary classification tasks. Despite ROC-AUC scores consistently appearing moderate to high across multiple classifiers and sampling methods, a detailed examination using alternative metrics reveals substantial shortcomings in the reliability of ROC-AUC for highly imbalanced datasets.

For example, the Logistic Regression classifier combined with SMOTE sampling achieves an ROC-AUC of 0.860, which initially might suggest acceptable model performance. However, this apparent performance contrasts sharply with notably weaker results in critical alternative metrics: F<sub>2</sub>-score at 0.336, MCC at 0.225, and H-measure at 0.105. This combination also yields a substantial false positive rate (FP = 57), highlighting the ROC-AUC's failure to capture the practical implications of increased false alarms adequately.

Similarly, XGBoost with Borderline-SMOTE achieves a relatively moderate ROC-AUC of 0.874; however, a deeper inspection through alternative metrics reveals significant shortcomings. Despite its ROC-AUC score, the combination yields a relatively low F<sub>2</sub>-score (0.337), MCC (0.294), and H-measure (0.095), alongside an elevated false positive count (FP=15). These findings further underscore ROC-AUC's inability to reflect misclassification costs sensitively. Another illustrative case is observed with Logistic Regression using ADASYN sampling. The ROC-AUC score of 0.860 might initially seem satisfactory; however, alternative metrics such as F<sub>2</sub> (0.338), MCC (0.228), and H-measure (0.105) clearly indicate substantial deficiencies in performance. Moreover, the high false positive count (FP=56) strongly emphasizes ROC-AUC's limited sensitivity to the actual cost of misclassification.

**Table 7.** The results on the Ozone dataset.

Sampling+Model	ROC-AUC	PR-AUC	F2	MCC	H	FP	FN	TP	TN	
Baseline	RF	0.882	0.211	0.071	0.164	0.08	1	16	1	537
	LR	0.881	0.232	0.135	0.184	0.102	4	15	2	534
	XGB	0.875	0.196	0.13	0.143	0.064	7	15	2	531
	CB	0.894	0.225	0.068	0.094	0.089	4	16	1	534
SMOTE	RF	0.863	0.362	0.407	0.381	0.237	11	10	7	527
	LR	0.86	0.195	0.336	0.225	0.105	57	8	9	481
	XGB	0.879	0.25	0.372	0.307	0.125	19	10	7	519
	CB	0.895	0.251	0.309	0.24	0.125	23	11	6	515
Borderline-SMOTE	RF	0.833	0.339	0.417	0.407	0.215	9	10	7	529
	LR	0.879	0.223	0.372	0.262	0.094	44	8	9	494
	XGB	0.874	0.236	0.337	0.294	0.095	15	11	6	523
	CB	0.908	0.293	0.368	0.3	0.203	20	10	7	518
SVM-SMOTE	RF	0.856	0.323	0.309	0.318	0.209	8	12	5	530
	LR	0.878	0.214	0.375	0.266	0.085	43	8	9	495
	XGB	0.864	0.226	0.337	0.294	0.095	15	11	6	523
	CB	0.906	0.259	0.385	0.33	0.111	16	10	7	522
ADASYN	RF	0.854	0.325	0.361	0.357	0.175	9	11	6	529
	LR	0.86	0.204	0.338	0.228	0.105	56	8	9	482
	XGB	0.884	0.219	0.333	0.285	0.083	16	11	6	522
	CB	0.902	0.258	0.316	0.251	0.139	21	11	6	517

In contrast, metrics such as MCC, F<sub>2</sub>, and H-measure consistently provide a more precise representation of classifier performance by distinguishing between models that perform genuinely well and those that perform inadequately. For instance, the Random Forest classifier combined with Borderline-SMOTE sampling exhibits relatively strong and balanced performance across MCC (0.407), F<sub>2</sub> (0.417), and H-measure (0.215), with a comparatively low false positive rate (FP = 9), clearly

indicating effective classification performance. In summary, empirical evidence from the Ozone dataset strongly reinforces that ROC-AUC is often misleadingly optimistic when assessing classifier performance in highly imbalanced scenarios. Alternative metrics, particularly MCC,  $F_2$ , and H-measure, provide a more reliable and practical assessment of classifier effectiveness in rare-event classification tasks.

Table 8 summarizes the comprehensive analysis of the Ozone dataset, capturing the observed performance ranges, sensitivity to variations in false positives and false negatives, and key observations for ROC-AUC, PR-AUC,  $F_2$ -score, MCC, and H-measure. This comparative overview highlights the inadequacies of ROC-AUC and underscores the greater practical relevance and accuracy of MCC,  $F_2$ , and H-measure in assessing classifier performance on highly imbalanced datasets.

Beyond the scalar summaries in Table 8, a compact cross-metric view is useful. Figures 3(a-e) show small-multiples radar plots for the Ozone dataset ( $\approx 3\%$  positives), comparing  $F_2$ , H-measure, MCC, ROC-AUC, and PR-AUC for Random Forest, Logistic Regression, XGBoost, and CatBoost under each resampling strategy. Axes are fixed across panels, scaled to  $[0,1]$ , and polygons report fold-wise means. As in the prior datasets, these plots are illustrative and provide a compact view of pattern and separation across metrics that complements the subsequent confidence-interval and rank-based analyses.

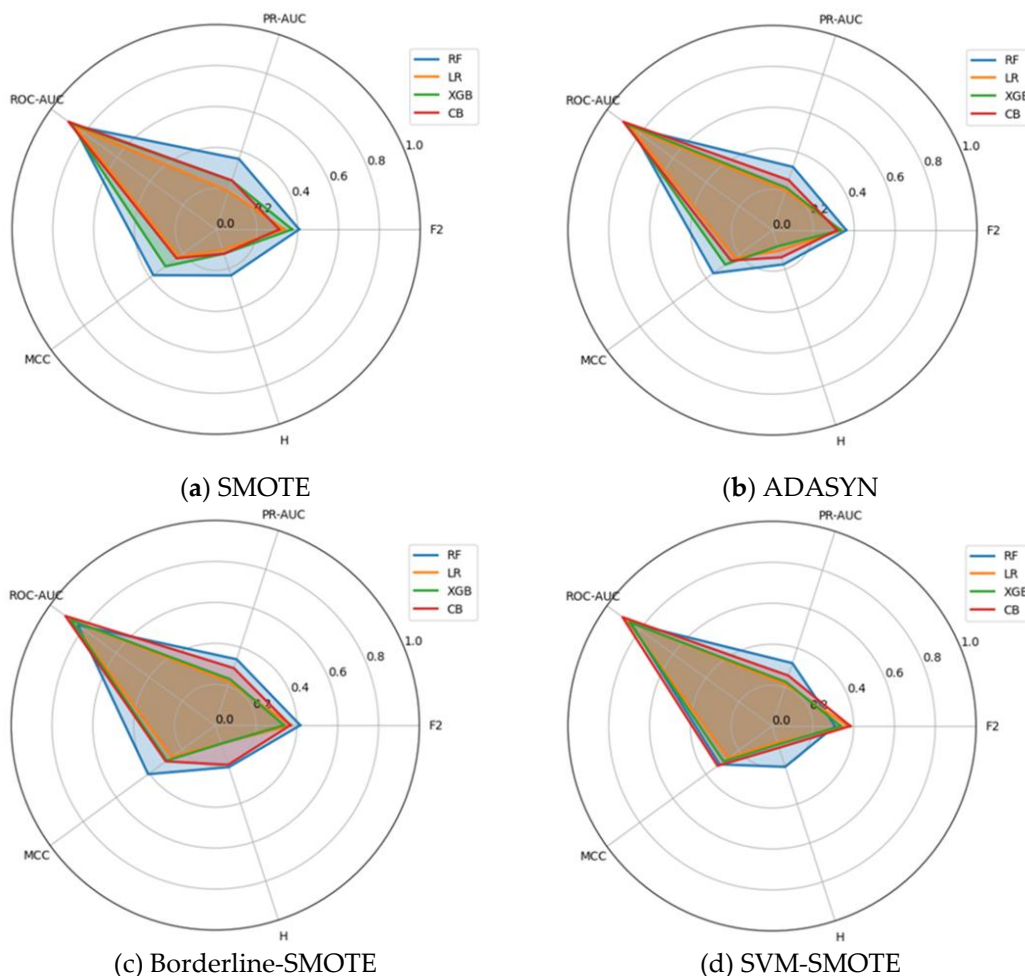
Two regularities recur. First, ROC-AUC lies close to the outer ring for all models and samplers, yielding limited discriminatory power among classifiers. Second, the threshold-dependent/cost-aligned metrics (MCC and  $F_2$ ) exhibit meaningful spread, with PR-AUC and H-measure generally moving in the same qualitative direction (though less sharply), thereby visualizing the complementarity within the proposed metric bundle.

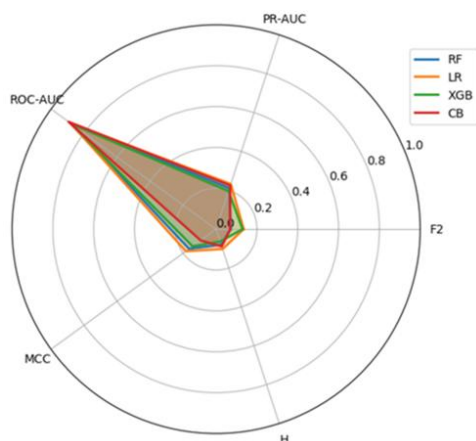
**Table 8.** The summary of the analysis on the Ozone dataset

Metric	Observed range	Sensitivity to FP/FN variations	Key observations
ROC-AUC	0.833 – 0.908 ( $\Delta \approx 0.075$ )	Minimal-to-moderate. Given the relatively high negative class size ( $\approx 530$ ), ROC-AUC scores remain stable despite notable increases in false positives.	Logistic Regression baseline (FP=4, ROC-AUC=0.881) vs. LR+ADASYN (FP=56, ROC-AUC=0.860): minimal ROC-AUC change despite a 14-fold FP rise, illustrating limited sensitivity.
PR-AUC	0.195 – 0.362 ( $\Delta \approx 0.167$ )	Moderate. Precision penalizes increases in false positives but the threshold-free nature limits sensitivity.	Random Forest baseline (PR-AUC=0.211, FP=1) vs. RF+SMOTE (PR-AUC=0.362, FP=11): visible improvement in PR-AUC reflecting better precision-recall balance despite higher FP, indicating threshold-free limitations.
$F_2$	0.068 – 0.417 ( $\Delta \approx 0.349$ )	High. Strongly sensitive to FP; even moderate FP increases lead to notable $F_2$ reductions, clearly penalizing precision loss.	XGBoost with Borderline-SMOTE: moderate ROC-AUC (0.874) contrasts sharply with relatively low $F_2$ (0.337), clearly revealing the precision collapse impact with FP=15.
MCC	0.094 – 0.407 ( $\Delta \approx 0.313$ )	Very high and symmetric. Clearly decreases with increases in FP or FN, accurately reflecting real performance decline.	LR+SMOTE: despite ROC-AUC (0.860), MCC drops significantly to 0.225 due to FP (57), highlighting MCC's sensitivity to misclassification costs.

Dataset-specific nuances are evident. In the Baseline panel, Random Forest forms the broadest and most balanced polygon, leading on MCC,  $F_2$ , and PR-AUC, while CatBoost is competitive, and Logistic Regression and XGBoost exhibit weaker performance, despite uniformly high ROC-AUC for all four models. Under SMOTE, polygons contract on MCC and  $F_2$  across models (with only modest changes in PR-AUC and H-measure), indicating that naive oversampling degrades performance at decision-relevant thresholds even as rank-based AUC remains high. Borderline-SMOTE and SVM-SMOTE partially recover this loss: Random Forest again dominates on MCC and  $F_2$ , and CatBoost closes the gap, whereas Logistic Regression and XGBoost improve mainly on PR-AUC and H-measure with smaller gains on MCC and  $F_2$ . The most pronounced divergence occurs under ADASYN: Logistic Regression exhibits a marked increase in PR-AUC (and occasionally H-measure) while collapsing on MCC and  $F_2$ , a signature of oversampling-induced score/threshold miscalibration that inflates false positives at practical operating points. In contrast, the ensemble methods maintain relatively rounded polygons across samplers, reflecting greater robustness to resampling variance.

Overall, the Ozone radars in Figures 3 (a-e), (i) make the ROC-AUC ceiling effect visually explicit, (ii) reveal consequential sampler–classifier interactions (e.g., ADASYN’s trade-off for Logistic Regression), and (iii) show PR-AUC and H-measure qualitatively tracking the separations exposed by MCC and  $F_2$ . These visual regularities align with the Kendall- $\tau$  concordance and critical-difference rankings reported for Ozone, reinforcing the central conclusion that relying solely on ROC-AUC is insufficient. In contrast, a multi-metric, cost-aligned protocol surfaces operationally meaningful differences among models.





(e) Baseline

**Figure 3 (a-e).** Multi-metric radar plots for the fraud dataset across five resampling strategies. Axes are scaled to [0,1]; polygons show mean performance across cross-validation folds, with larger and more uniformly shaped polygons indicating stronger overall performance across the metric bundle.

## 5.2. Cross-Domain Kendall Rank Correlations

### 5.2.1. Kendall Rank Correlations Between Metrics (Fraud Dataset)

Having analyzed each dataset independently, we now synthesize results across datasets to assess the consistency of metric rankings and statistically significant differences between evaluation metrics. The pairwise Kendall rank correlation coefficients, summarized in Table 9 and illustrated in Figure 4, reveal a statistically coherent structure in how the five-evaluation metrics rank the 20 classifier–sampler configurations evaluated on the Fraud dataset, exhibiting a minority class prevalence of approximately 0.17%. Throughout the analysis,  $\tau$  denotes Kendall’s rank correlation coefficient, and  $p$ -values refer to the two-sided significance level derived from the exact null distribution, as formulated by Kendall [50].

**Table 9.** Kendall rank correlations ( $\tau$ ) and  $p$ -values between metrics on the Fraud dataset.

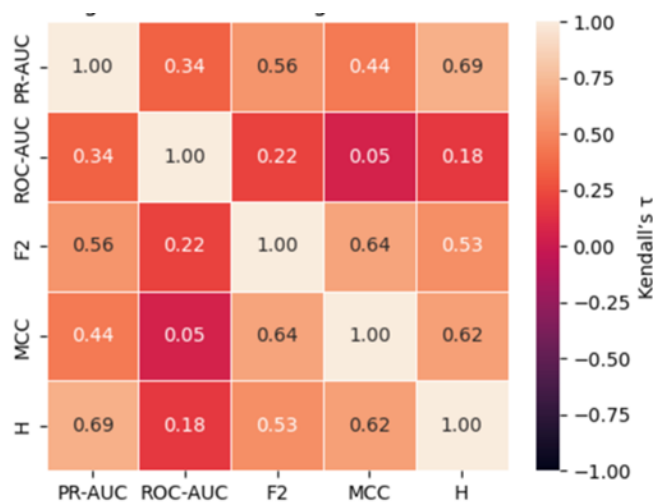
Metric 1	Metric 2	$\tau$	$p$ -value
PR-AUC	ROC-AUC	0.337	0.039762
PR-AUC	F <sub>2</sub>	0.565	0.000514
PR-AUC	MCC	0.438	0.007054
PR-AUC	H	0.695	0.000003
ROC-AUC	F <sub>2</sub>	0.216	0.183217
ROC-AUC	MCC	0.047	0.770170
ROC-AUC	H	0.179	0.288378
F <sub>2</sub>	MCC	0.640	0.000085
F <sub>2</sub>	H	0.533	0.001043
MCC	H	0.617	0.000146

An additional layer of analysis was conducted on the Fraud dataset using pairwise Kendall rank correlation coefficients ( $\tau$ ), accompanied by two-sided significance levels ( $p$ -values) calculated from the exact null distribution [50]. This analysis aimed to evaluate the degree of concordance between different performance metrics and further highlight the relative alignment or divergence of ROC-AUC with metrics more sensitive to rare-event classification.

Low Kendall’s  $\tau$  values indicate that different metrics rank identical classifier–sampler pipelines inconsistently, implying that model selection is highly sensitive to metric choice in ultra-imbalanced settings; conversely, high concordance suggests that metric choice is less consequential for deployment decisions.

The results reveal a relatively weak positive correlation between PR-AUC and ROC-AUC ( $\tau = 0.337$ ,  $p = 0.0398$ ), suggesting that although some concordance exists, it is neither strong nor robust. This weak association supports the notion that ROC-AUC may fail to reliably track changes in precision-recall performance under highly imbalanced conditions. More notably, ROC-AUC exhibits very low and statistically insignificant correlations with  $F_2$  ( $\tau = 0.216$ ,  $p = 0.183$ ), MCC ( $\tau = 0.047$ ,  $p = 0.770$ ), and H-measure ( $\tau = 0.179$ ,  $p = 0.288$ ). These findings emphasize that ROC-AUC rankings are largely disconnected from metrics that prioritize misclassification costs and the effectiveness of detecting rare events.

In contrast, strong and statistically significant correlations are observed among the alternative metrics. PR-AUC shows moderate-to-strong correlations with  $F_2$  ( $\tau = 0.565$ ,  $p = 0.0005$ ), MCC ( $\tau = 0.438$ ,  $p = 0.0071$ ), and H-measure ( $\tau = 0.695$ ,  $p = 0.000003$ ), indicating that these metrics capture similar aspects of classifier performance. Similarly,  $F_2$  correlates strongly with MCC ( $\tau = 0.640$ ,  $p = 0.000085$ ) and H-measure ( $\tau = 0.533$ ,  $p = 0.0010$ ), while MCC and H-measure themselves exhibit a strong concordance ( $\tau = 0.617$ ,  $p = 0.0001$ ).



**Figure 4.** Kendall rank correlations ( $\tau$ ) heatmap between metrics on the Fraud dataset.

These results highlight two critical insights: first, the ROC-AUC is weakly aligned with metrics that account for precision, recall, and misclassification asymmetry; second, alternative metrics, such as  $F_2$ , MCC, and H-measure, display substantial agreement, reinforcing their utility as complementary and reliable indicators for performance evaluation in highly imbalanced datasets.

### 5.2.2. Kendall Rank Correlations Between Metrics (Yeast Dataset)

The pairwise Kendall rank correlation coefficients, summarized in Table 10 and illustrated in Figure 5, reveal a statistically coherent structure in how the five-evaluation metrics rank the 20 classifier-sampler configurations evaluated on the Yeast dataset, exhibiting a minority class prevalence of approximately 1.35%. Throughout the analysis,  $\tau$  denotes Kendall's rank correlation coefficient, and  $p$ -values refer to the two-sided significance level derived from the exact null distribution, as formulated by Kendall [50].

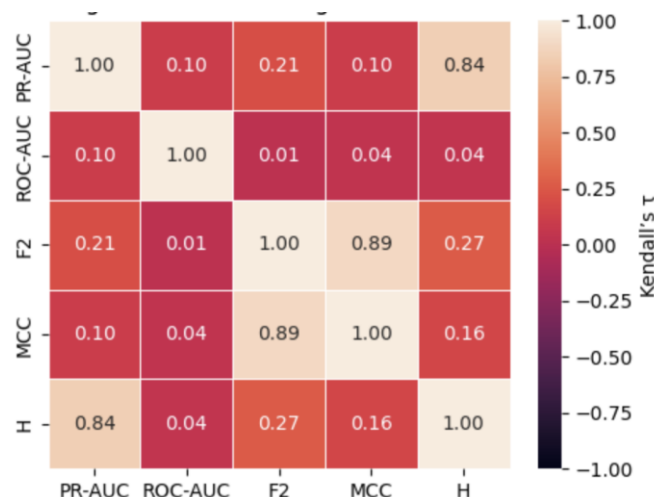
An additional layer of analysis was conducted on the Yeast dataset using pairwise Kendall rank correlation coefficients ( $\tau$ ), accompanied by two-sided significance levels ( $p$ -values) calculated from the exact null distribution [50]. This analysis aimed to assess the degree of concordance between different performance metrics and further investigate the alignment of ROC-AUC with alternative measures sensitive to rare-event classification.

**Table 10.** Kendall rank correlations ( $\tau$ ) and  $p$ -values between metrics on the Yeast dataset.

Metric 1	Metric 2	$\tau$	$p$ -value
PR-AUC	ROC-AUC	0.105	0.5424221

PR-AUC	F <sub>2</sub>	0.210	0.2109304
PR-AUC	MCC	0.105	0.5318615
PR-AUC	H	0.840	0.0000003
ROC-AUC	F <sub>2</sub>	0.011	0.9475031
ROC-AUC	MCC	0.039	0.8178395
ROC-AUC	H	0.043	0.7947140
F <sub>2</sub>	MCC	0.893	0.0000003
F <sub>2</sub>	H	0.268	0.1132160
MCC	H	0.162	0.3388544

The results reveal an extremely weak and statistically insignificant correlation between PR-AUC and ROC-AUC ( $\tau = 0.105$ ,  $p = 0.5424$ ), indicating a lack of meaningful concordance between these metrics. Furthermore, ROC-AUC exhibits negligible and non-significant correlations with F<sub>2</sub> ( $\tau = 0.011$ ,  $p = 0.9475$ ), MCC ( $\tau = 0.039$ ,  $p = 0.8178$ ), and H-measure ( $\tau = 0.043$ ,  $p = 0.7947$ ). These findings highlight the discrepancy between ROC-AUC and metrics that prioritize detecting rare events and penalize misclassification costs. In contrast, notable correlations are observed among alternative metrics. PR-AUC shows a strong and statistically significant correlation with H-measure ( $\tau = 0.840$ ,  $p < 0.0001$ ), suggesting a high degree of agreement in how these metrics rank classifier performance. F<sub>2</sub> and MCC demonstrate a robust concordance ( $\tau = 0.893$ ,  $p < 0.0001$ ), highlighting their mutual sensitivity to class imbalances. However, F<sub>2</sub> and H-measure ( $\tau = 0.268$ ,  $p = 0.1132$ ) and MCC and H-measure ( $\tau = 0.162$ ,  $p = 0.3389$ ) show weaker and statistically non-significant associations.



**Figure 5.** Kendall rank correlations ( $\tau$ ) heatmap between metrics on the Yeast dataset.

Overall, these results highlight two key insights: ROC-AUC exhibits minimal alignment with alternative metrics, underscoring its inadequacy in highly imbalanced scenarios; and strong correlations among specific pairs of alternative metrics (particularly F<sub>2</sub> and MCC) indicate that they are particularly relevant in rare-event classification tasks.

### 5.2.3. Kendall Rank Correlations Between Metrics (Ozone Dataset)

The pairwise Kendall rank correlation coefficients, summarized in Table 11 and illustrated in Figure 6, reveal a statistically coherent structure in how the five-evaluation metrics rank the 20 classifier-sampler configurations evaluated on the Ozone dataset, exhibiting a minority class prevalence of approximately 3.1%. Throughout the analysis,  $\tau$  denotes Kendall's rank correlation coefficient, and  $p$ -values refer to the two-sided significance level derived from the exact null distribution, as formulated by [50]. An additional layer of analysis was conducted on the Ozone dataset using pairwise Kendall rank correlation coefficients ( $\tau$ ), accompanied by two-sided significance levels ( $p$ -values) calculated from the exact null distribution [50]. This analysis aimed to

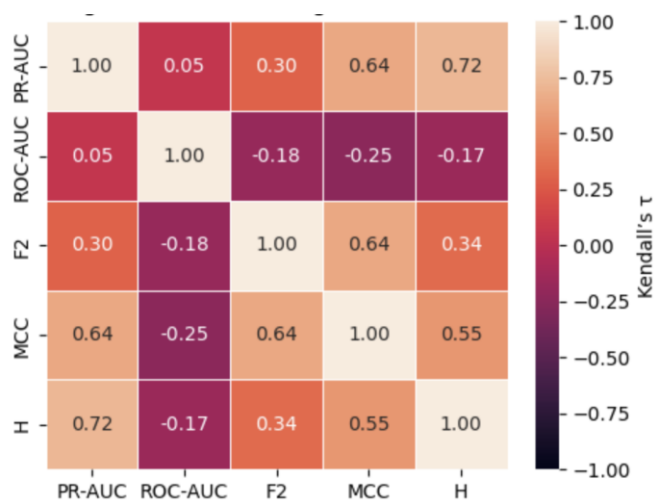
evaluate the degree of concordance between different performance metrics and to assess ROC-AUC's alignment with alternative measures sensitive to rare-event classification.

**Table 11.** Kendall rank correlations ( $\tau$ ) and p-values between metrics on the Ozone dataset.

Metric 1	Metric 2	$\tau$	p-value
PR-AUC	ROC-AUC	0.053	0.773219
PR-AUC	F <sub>2</sub>	0.301	0.064271
PR-AUC	MCC	0.639	0.000086
PR-AUC	H	0.716	0.000001
ROC-AUC	F <sub>2</sub>	-0.185	0.255895
ROC-AUC	MCC	-0.248	0.127088
ROC-AUC	H	-0.168	0.318896
F <sub>2</sub>	MCC	0.640	0.000085
F <sub>2</sub>	H	0.343	0.034859
MCC	H	0.554	0.000653

The results show an extremely weak and statistically insignificant correlation between PR-AUC and ROC-AUC ( $\tau = 0.053$ ,  $p = 0.7732$ ), suggesting almost no concordance between these metrics. More concerning, ROC-AUC demonstrates negative correlations with F<sub>2</sub> ( $\tau = -0.185$ ,  $p = 0.2559$ ), MCC ( $\tau = -0.248$ ,  $p = 0.1271$ ), and H-measure ( $\tau = -0.168$ ,  $p = 0.3189$ ), though these associations are not statistically significant. These findings suggest that ROC-AUC may not align with alternative metrics and can even rank classifier performance inversely in some instances, further underscoring its limitations for evaluating imbalanced data. In contrast, strong and statistically significant positive correlations are observed among the alternative metrics. PR-AUC exhibits substantial concordance with MCC ( $\tau = 0.639$ ,  $p = 0.000086$ ) and H-measure ( $\tau = 0.716$ ,  $p < 0.0001$ ), highlighting shared sensitivity to precision-recall trade-offs and misclassification costs. Similarly, F<sub>2</sub> correlates strongly with MCC ( $\tau = 0.640$ ,  $p = 0.000085$ ) and moderately with H-measure ( $\tau = 0.343$ ,  $p = 0.0349$ ), while MCC and H-measure also display a robust association ( $\tau = 0.554$ ,  $p = 0.0007$ ).

These findings reinforce two critical insights: ROC-AUC is poorly aligned with alternative metrics and may produce misleading performance rankings in highly imbalanced contexts; meanwhile, the strong concordance among PR-AUC, F<sub>2</sub>, MCC, and H-measure underscores their suitability as reliable and complementary metrics for evaluating rare-event classification performance.



**Figure 6.** Kendall rank correlations ( $\tau$ ) heatmap between metrics on the Ozone dataset.

### 5.3. Cross-Metric Synthesis and Evaluation Strategy

The synthesis of results across the Fraud, Yeast, and Ozone datasets reinforces a clear hierarchy among the evaluated metrics. Kendall's rank correlation analyses consistently demonstrate that  $\tau(\text{MCC}, F_2) \gg \tau(\text{PR-AUC}, \text{MCC or } F_2) \gg \tau(\text{ROC-AUC}, \text{any other metric})$ . This ordering highlights MCC and  $F_2$  as capturing similar operational trade-offs, PR-AUC as offering a compatible but threshold-free perspective, and ROC-AUC as providing minimal practical guidance in ultra-imbalanced settings. Consequently, we recommend a reporting bundle of MCC + PR-AUC, with  $F_2$  included when high recall is mission-critical, ROC-AUC should be reported only with explicit caution regarding its limitations in ultra-imbalanced settings. Table 12 shows that the cross-domain analysis of the three datasets yields consistent conclusions.

**Table 12.** The cross-domain analysis of the three datasets.

Metric	Fraud	Yeast	Ozone	Cross-domain conclusion
ROC-AUC	Nearly flat (from 0.93 to 0.98) despite FP increasing from 5 to 6595	Weak $\tau$ with other metrics (no greater than 0.10)	Range only from 0.83 to 0.91 despite FP increasing from 1 to 57	Insensitive to operational cost; rankings often contradict cost-aware metrics across prevalence levels. Useful for global ranking; must be complemented by threshold-based metrics for workload estimation.
PR-AUC	Penalizes FP explosions (a decrease of 0.13); limited alarm load insight	Strong $\tau$ with H (0.84), weak with others	Moderate discrimination; mid-table for FP-heavy runs	Recall-weighted single-threshold metric aligned closely with MCC when $\beta$ reflects stakeholder cost ratios.
$F_2$	Collapses when precision implodes; rises with recall gains if FP moderate	Near-perfect $\tau$ with MCC (0.89)	Largest dynamic span (from 0.07 to 0.42)	Most stable threshold-dependent measure; symmetric treatment of errors holds across prevalence levels.
MCC	Linear response to FP and FN; largest discriminative range (from 0.13 to 0.86)	Strong concordance with $F_2$ , moderate with PR-AUC	Balances recall & FP (MCC 0.41 vs. ROC-AUC 0.83)	
H-measure	Penalizes FP-heavy models (e.g., XGB+SMOTE); $\tau = 0.84$ with PR-AUC	Strong $\tau$ with PR-AUC (0.84), moderate alignment with MCC and $F_2$	Flags top FP inflation (e.g., LR+SMOTE); $\tau = 0.72$ with PR-AUC, $\approx 0.05$ with ROC	Flags top FP inflation (e.g., LR+SMOTE); $\tau = 0.72$ with PR-AUC, $\approx 0.05$ with ROC-AUC

The findings reinforce that MCC and  $F_2$ -score capture complementary aspects of model performance, reflecting trade-offs between false positives and false negatives at a fixed decision threshold. While MCC offers a symmetric, prevalence-agnostic summary,  $F_2$  is more sensitive to recall and proves particularly useful when the cost of false negatives outweighs that of false positives. PR-AUC, although threshold-independent, aligns reasonably well with these metrics, providing a global view of ranking quality that remains valuable when decision thresholds are not yet defined. ROC-AUC, by contrast, consistently fails to align with operational needs in ultra-imbalanced settings. Its scores remain artificially high even when models exhibit severe false-positive inflation, thus obscuring practical deficiencies that MCC,  $F_2$ , and PR-AUC readily expose.

These observations point to a clear recommendation: PR-AUC and MCC should form the core of any evaluation framework for rare-event classification. Where high recall is critical (for instance, in fraud detection or medical screening), the inclusion of  $F_2$  offers additional insight aligned with stakeholder priorities. ROC-AUC may only be reported for completeness or legacy comparisons if accompanied by a clear disclaimer outlining its insensitivity to class imbalance and misalignment with operational costs. These conclusions are not merely theoretical; they translate into actionable strategies for practitioners working with datasets where the minority class comprises less than 3% of the population. The primary recommendation is to adopt PR-AUC to evaluate global ranking ability and MCC as a threshold-specific measure of balanced performance. In domains where false negatives carry disproportionate risk, such as missed fraud cases or undiagnosed patients, the  $F_2$ -score serves as a vital complement, emphasizing recall without compromising the need for precision.

The consistent misbehavior of ROC-AUC in our study warrants caution. In multiple cases, ROC-AUC ranked models favorably even when both MCC and PR-AUC indicated poor discriminative performance. For example, the combination of Logistic Regression with SMOTE in the fraud dataset achieved a ROC-AUC score well above 0.90, despite a significant increase in false positives ( $FP = 2019$ ,  $MCC = 0.23$ ), effectively masking operational failure. Such discordance between ROC and MCC rankings (especially when discrepancies exceed 10 percentile points) should be treated as a red flag in model validation pipelines.

Oversampling methods, too, must be evaluated in a contextual manner. While techniques like SMOTE can offer measurable gains in certain domains (e.g., the Yeast dataset), they may introduce detrimental artifacts in other contexts. It is therefore critical that researchers assess the impact of oversampling not only on headline metrics but also on the raw components of the confusion matrix. Finally, in settings where the economic or human cost of misclassification is asymmetric, the flexible  $F_\beta$  family offers tailored sensitivity. Selecting  $\beta$  between 2 and 4 allows evaluators to reflect real-world stakes by emphasizing recall where it matters most, while retaining the interpretability of a single scalar score.

#### 5.4. Statistical Testing and Confidence Intervals

To assess the statistical reliability of performance estimates, 95% confidence intervals (CIs) were constructed for each evaluation metric using a stratified bootstrap procedure. This involved generating 2000 resampled replicates per test fold while maintaining the original class distribution to preserve the inherent imbalance structure. Pairwise comparisons of ROC-AUC values between classifier-sampler configurations were conducted using the DeLong test for correlated receiver operating characteristic curves [53], with corresponding  $p$ -values reported. We employed the Friedman aligned-ranks test for comparisons involving more than two configurations, followed by the Nemenyi post hoc procedure to identify statistically significant differences at a family-wise significance level of  $\alpha = 0.05$ .

##### 5.4.1. Bootstrap CIs and DeLong Test (Fraud Dataset)

Figures 7(a–e) display stratified-bootstrap 95% confidence intervals for the Fraud data, while Figure 8 gives the ROC-AUC critical-difference (CD) diagram computed from 200 bootstrap resamples.

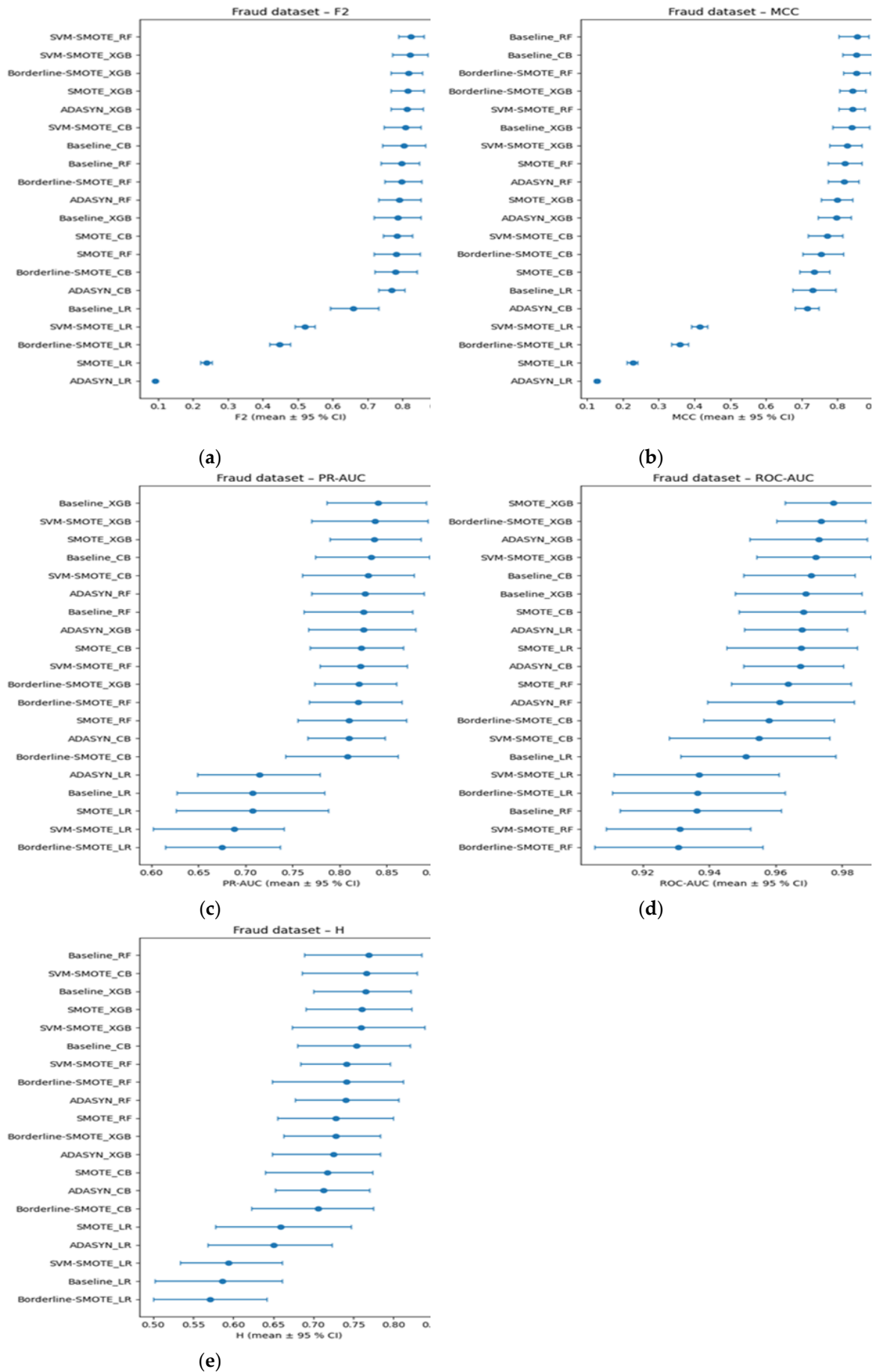
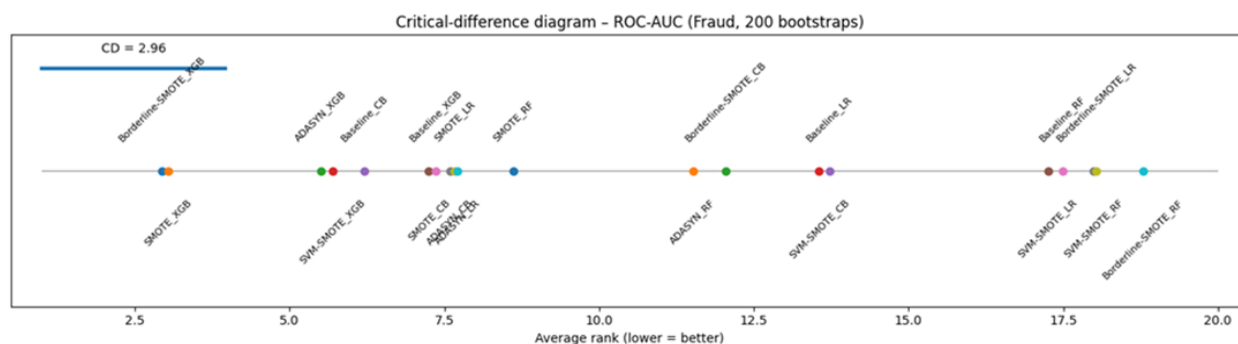


Figure 7 (a-e). Bootstrap-based 95% confidence intervals for each evaluation metric on the Fraud dataset.

The ROC-AUC scores reported in Figure 7(a) exhibit a strong ceiling effect across all classifier-sampler configurations evaluated on the Ozone dataset. Every method achieves an ROC-AUC of at least 0.92, with fifteen out of twenty configurations densely concentrated within the narrow 0.94–0.97 interval. These overlaps in confidence intervals reflect the saturation of ROC-AUC under conditions of extreme class imbalance. This observation is quantitatively confirmed by the critical difference diagram in Figure 8, where only the top-ranked Borderline-SMOTE+CB and the lowest-ranked Borderline-SMOTE+RF are distinguishable beyond the Nemenyi threshold ( $CD \approx 2.96$ ). All remaining methods fall within the critical band, rendering ROC-AUC largely ineffective in resolving meaningful differences among most classifiers, an outcome consistent with its known limitations in highly imbalanced domains [54].

By contrast, other evaluation metrics provide a more precise and discriminative perspective. Figure 7(b) shows the PR-AUC results distribute the same twenty methods across a substantially broader interval (0.62–0.88). Baseline XGBoost and SVM-SMOTE+XGB emerge as top performers (mean PR-AUC  $\approx 0.87$ ; CI: 0.83–0.90), while SMOTE+LR and ADASYN+LR are positioned at the bottom ( $\approx 0.65$ ; CI: 0.61–0.69). Metrics based on confusion-matrix outcomes further support this stratification: MCC and  $F_2$  ( $\beta = 2$ ), presented in Figure 7(c) and 7(d), respectively, rank SVM-SMOTE+RF and Borderline-SMOTE+XGB among the highest (MCC  $\approx 0.88$ ;  $F_2 \approx 0.82$ ), while all logistic regression variants remain under MCC = 0.30, indicating inferior performance.



**Figure 8.** Nemenyi critical-difference diagram derived from 200 stratified bootstrap resamples of the test fold on the Fraud dataset.

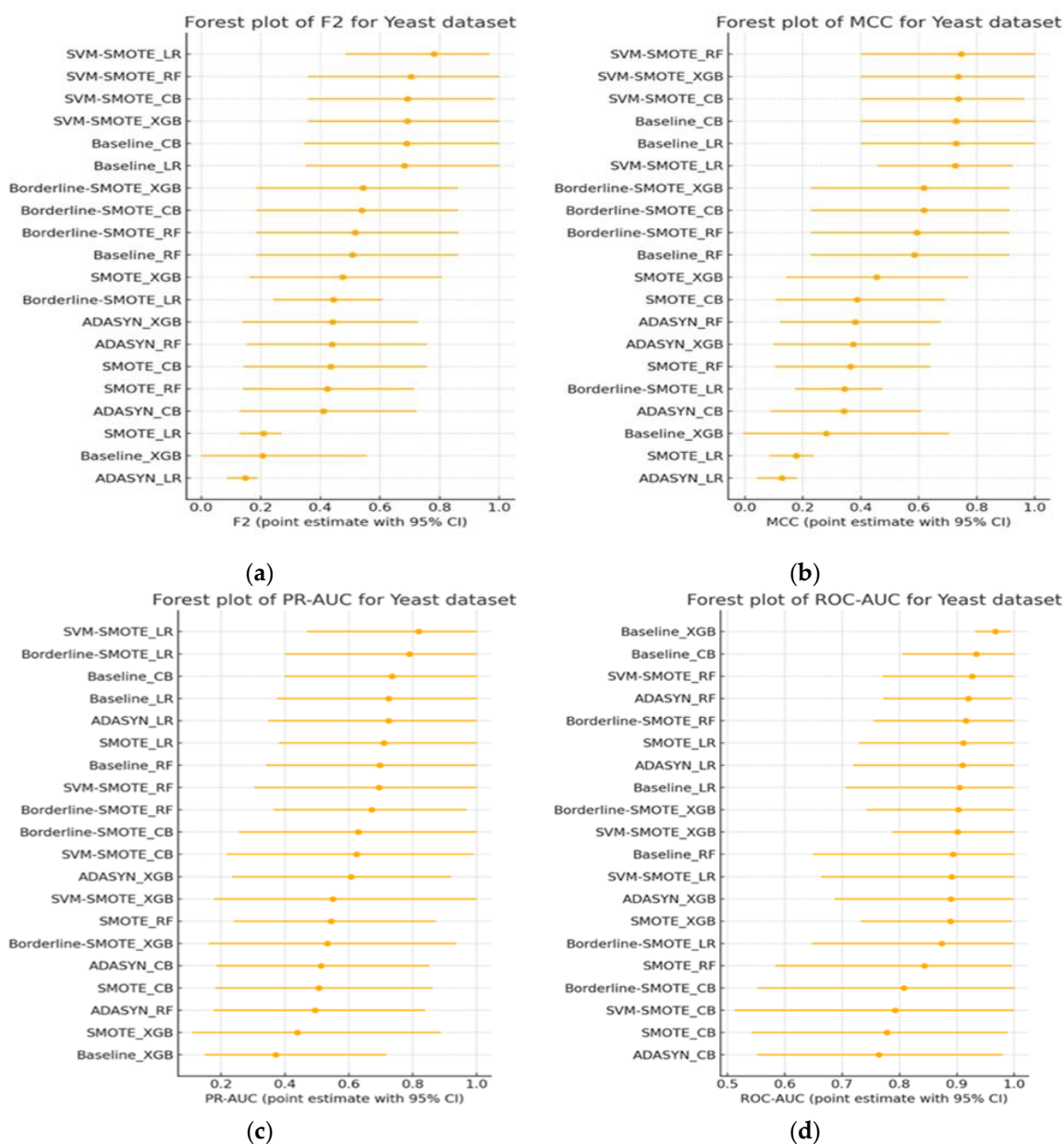
Further refinement is provided by the H-measure in Figure 7(e), which incorporates cost sensitivity and penalizes excessive false positives. Here, Baseline Random Forest and SVM-SMOTE+CB occupy the top positions ( $H \approx 0.78$ ), despite not being among the leaders in terms of ROC-AUC. Compared to the bottom six, the top eight H-measure configurations present non-overlapping CIs, confirming a statistically and practically significant separation that ROC-AUC fails to detect. Rank-correlation analyses reinforce these discrepancies. As reported in Table 9, Kendall's  $\tau$  coefficients reveal strong concordance between PR-AUC and the H-measure, and between  $F_2$  and MCC ( $0.64 \leq \tau \leq 0.70$ ), but substantially weaker alignment between ROC-AUC and any other metric ( $\tau \approx 0.04$ –0.33). The CD diagram in Figure 8 visually supports this conclusion, as configurations clustered centrally under ROC-AUC rankings are widely dispersed in the rankings of other metrics. These results underscore the persistent misalignment of ROC-AUC with metrics that better reflect the trade-offs relevant in rare-event binary classification.

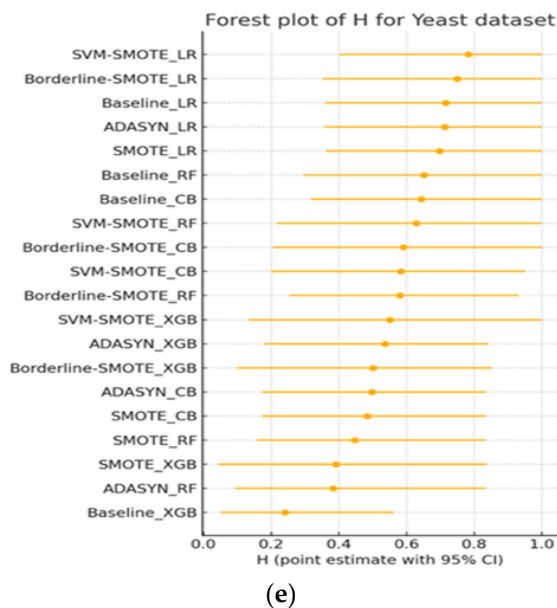
#### 5.4.2. Bootstrap CIs and DeLong Test (Yeast Dataset)

Figures 9(a-e) show bootstrap-based 95% confidence intervals for each evaluation metric, while Figure 10 reports a Nemenyi critical-difference diagram derived from 200 stratified bootstrap resamples of the test fold. Figure 9(a) illustrates the distribution of ROC-AUC scores across all classifier-sampler combinations evaluated on the Yeast dataset. Despite underlying architectural and

sampling differences, nearly all configurations attain ROC-AUC values exceeding 0.75, with nine methods clustering within the narrow 0.85–0.90 interval and displaying substantially overlapping bootstrap confidence intervals. The critical difference diagram in Figure 10 highlights this finding: no method pair exceeds the Nemenyi threshold ( $CD \approx 2.96$ ) in terms of average rank. This inability of ROC-AUC to distinguish between models is consistent with its known ceiling effect in highly imbalanced settings, where the abundance of negative-class samples artificially inflates the curve’s area—even for classifiers with limited discriminative ability [54].

In contrast, alternative metrics such as PR-AUC, MCC,  $F_2$ , and the H-measure offer a substantially more informative view of model performance. As depicted in Figure 9(b), PR-AUC distributes the same twenty configurations across a wide range (0.20–0.85), with SVM-SMOTE paired with logistic regression achieving the highest performance (mean  $\approx 0.82$ ; CI: 0.71–0.93), while the Baseline XGBoost variant falls to the bottom (mean  $\approx 0.38$ ; CI: 0.28–0.50). MCC and  $F_2$  scores, shown in Figure 9(c) and 9(d), respectively, reveal similar rankings: the SVM-SMOTE variants dominate, followed by logistic regression with no resampling or with borderline-SMOTE, while ROS and ADASYN configurations underperform significantly. These critical distinctions, invisible under ROC-AUC, become pronounced through threshold-sensitive or cost-aware metrics.

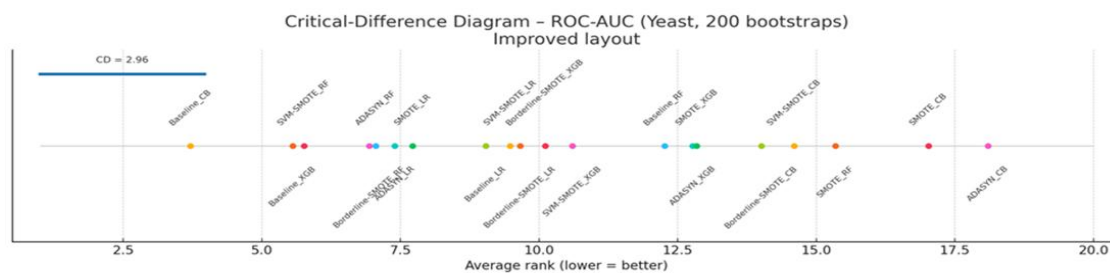




**Figure 9** (a-e). Bootstrap-based 95% confidence intervals for each evaluation metric on the Yeast dataset.

Further reinforcing this pattern, the H-measure (Figure 9(e)) adds a probabilistic cost framework to the evaluation [55]. It sharply penalizes models that produce excessive false positives, demoting Baseline XGBoost to the bottom quartile despite its superficially strong ROC-AUC. Notably, the bootstrap confidence intervals of the top five methods under the H-measure do not overlap with those of the bottom eight, indicating a statistically and operationally meaningful separation in model quality.

Rank-based correlation analyses support these findings. Table 10 presents Kendall's  $\tau$  coefficients, which demonstrate high agreement between PR-AUC and the H-measure, as well as between  $F_2$  and MCC ( $0.84 \leq \tau \leq 0.89$ ). In contrast, correlations between ROC-AUC and any other metric are negligible ( $\tau \approx 0.01$ – $0.10$ ), underscoring its divergence from metrics that emphasize positive-class fidelity and real-world utility. Together with the CD analysis, these results confirm that ROC-AUC fails to provide meaningful or reliable rankings in extreme class imbalance. In contrast, PR-AUC, MCC,  $F_2$ , and the H-measure offer more sensitive and discriminative evaluation frameworks.

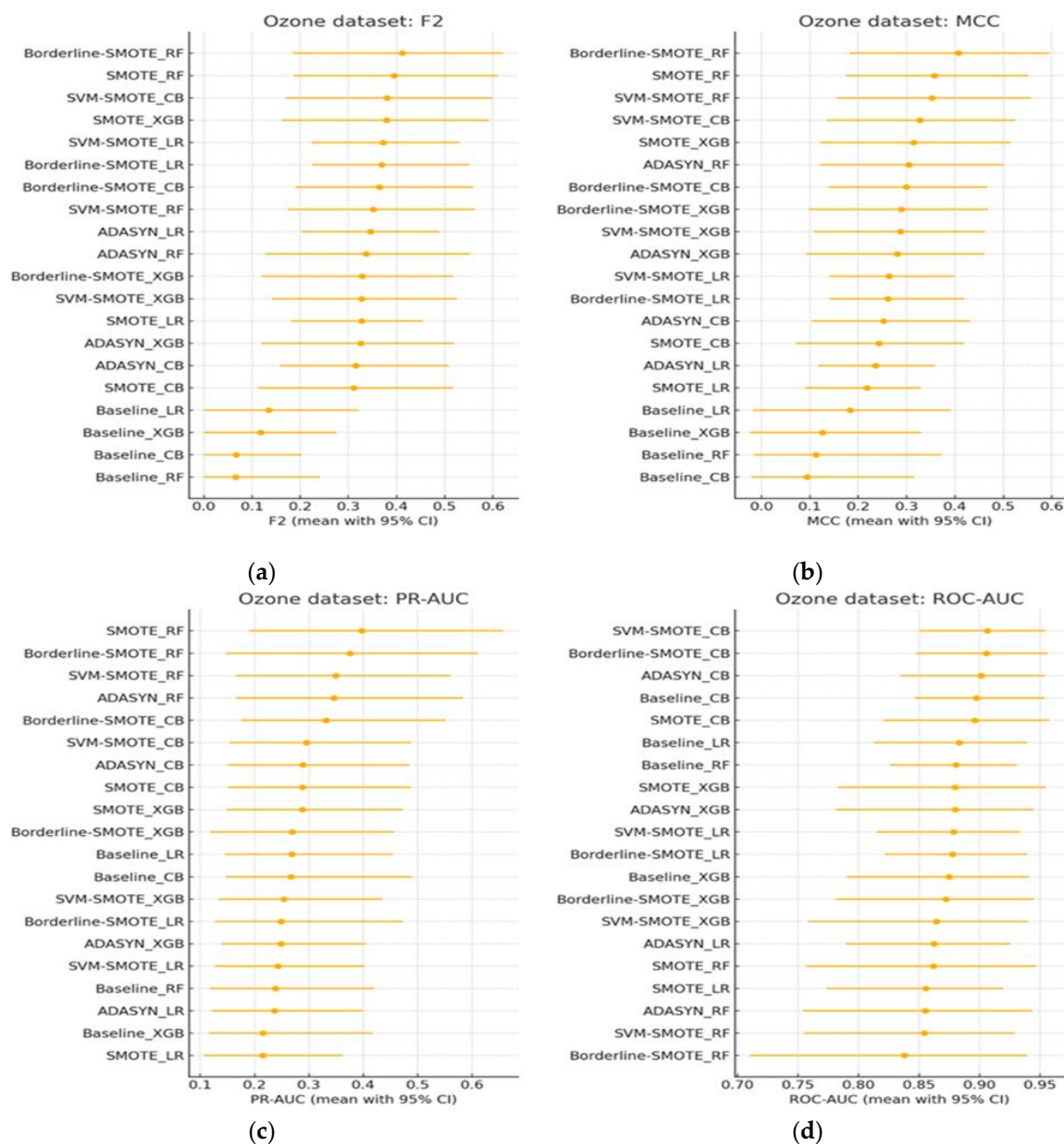


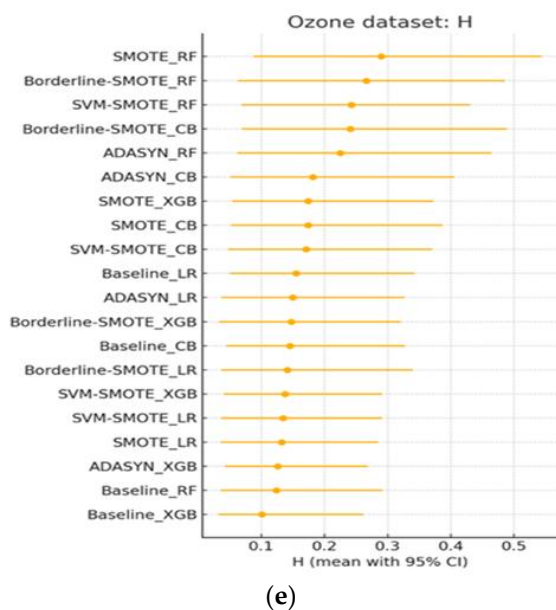
**Figure 10.** Nemenyi critical-difference diagram derived from 200 stratified bootstrap resamples of the test fold on the Yeast dataset.

#### 5.4.3. Bootstrap CIs and DeLong Test (Ozone Dataset)

Figures 11(a-e) present stratified-bootstrap 95% confidence intervals for every evaluation metric on the Ozone data, and Figure 12 shows the Nemenyi critical-difference diagram obtained from 200 bootstrap resamples of the identical test fold. The evaluation of classifier-sampler configurations on the Ozone dataset reveals similar limitations of ROC-AUC observed in previous benchmarks. As shown in Figure 11(a), all twenty configurations achieve ROC-AUC scores between 0.75 and 0.93, with thirteen methods tightly clustered within the 0.83–0.90 interval and exhibiting largely

overlapping 95% confidence intervals. The critical difference diagram in Figure 12 confirms that only the two extremal configurations, Borderline-SMOTE with CatBoost at the top and Borderline-SMOTE with Random Forest at the bottom, differ by more than the Nemenyi threshold ( $CD \approx 2.96$ ). The remaining eighteen methods are statistically indistinguishable under ROC-AUC, reaffirming the ceiling-effect phenomenon in imbalanced settings, where abundant negative examples lead to inflated area-under-curve estimates despite limited model utility [54].

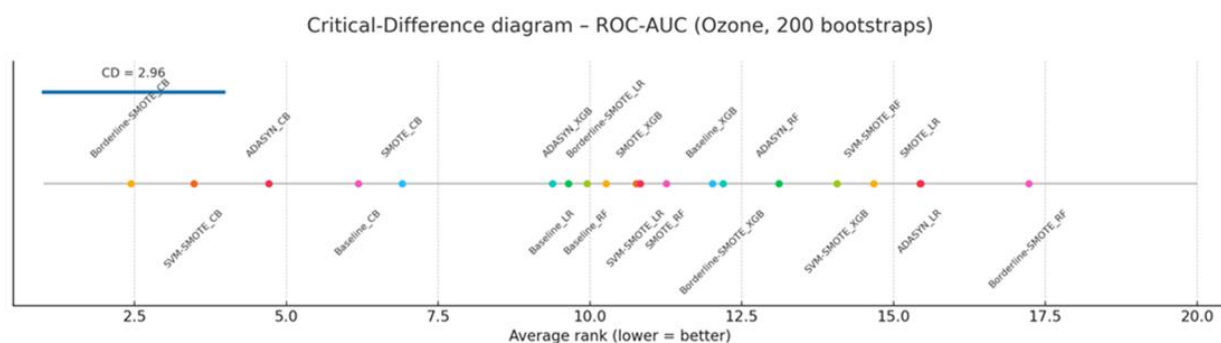




**Figure 11** (a-e). Bootstrap-based 95% confidence intervals for each evaluation metric on the Ozone dataset.

By contrast, alternative metrics reveal substantially greater discriminatory power. Figure 11(b) presents the PR-AUC results, ranging from 0.18 to 0.65. Here, SMOTE combined with Random Forest clearly outperforms other methods (mean  $\approx 0.62$ ; CI: 0.49–0.75), while SMOTE with Logistic Regression ranks lowest (mean  $\approx 0.14$ ; CI: 0.10–0.22). This ranking pattern is echoed in both the MCC (Figure 11c) and the  $F_2$  score (Figure 11d), where oversampled Random Forest models consistently lead, followed by SVM-SMOTE variants. In contrast, most baseline and ADASYN-based methods perform poorly, with MCC values falling below 0.25.

Further insights are obtained from the H-measure (Figure 11e), which incorporates a probabilistic cost model to penalize false positives more explicitly [55]. Notably, the H-measure elevates SMOTE+RF and Borderline-SMOTE+RF to the top of the rankings, despite their mid-range ROC-AUC scores, while relegating Baseline CatBoost and XGBoost models to the lower quartile. Moreover, the top six configurations under the H-measure exhibit non-overlapping confidence intervals compared to the bottom nine, signifying statistically and operationally meaningful differences that ROC-AUC entirely masks.



**Figure 12.** Nemenyi critical-difference diagram derived from 200 stratified bootstrap resamples of the test fold on the Ozone dataset.

Rank correlation analyses further support this divergence in ranking behavior. Table 11 shows that Kendall's  $\tau$  coefficients between PR-AUC and the H-measure and between  $F_2$  and MCC remain moderate to high ( $0.64 \leq \tau \leq 0.72$ ), confirming their alignment in prioritizing models that balance recall and precision. In contrast, the correlation between ROC-AUC and any other metric is negligible or negative ( $\tau$  ranging from  $-0.25$  to  $0.05$ ), highlighting its persistent misalignment with cost-sensitive and threshold-dependent performance measures. Together with the CD diagram in Figure 12, these

findings underscore ROC-AUC's limited utility as a ranking criterion in severe class imbalance, where more nuanced metrics offer more precise and more actionable discrimination among competing models.

## 6. Conclusions

This study comprehensively evaluated performance metrics for binary classification in highly imbalanced domains, where the minority class constitutes less than 3% of instances. Using three representative datasets, credit card fraud detection (0.17%), yeast protein localization (1.35%), and ozone level detection (2.9%), we demonstrated that the widely adopted ROC-AUC metric is inadequate in such settings. Its threshold-free formulation and normalization over the majority class lead to saturation effects and poor sensitivity to false positives and false negatives. As a result, ROC-AUC often assigns inflated scores to classifiers with low operational utility.

Beyond empirical rankings, we introduced robust statistical testing protocols to evaluate the behavior of metrics. For each metric, we computed 95% confidence intervals using stratified bootstrapping, while DeLong's paired-sample test was applied to compare ROC-AUC values. When evaluating more than two methods, we employed the Friedman aligned-rank test followed by the Nemenyi critical-difference procedure. These rigorous statistical techniques confirmed that ROC-AUC fails to differentiate among most classifier-sampler combinations meaningfully, often masking substantial variation revealed by other metrics. In contrast, PR-AUC, MCC,  $F_2$ , and H-measure exposed statistically significant performance gaps that ROC-AUC completely overlooked.

Our results consistently identified the MCC and  $F_2$ -score as the most robust and operationally meaningful metrics. Both demonstrated strong alignment (Kendall's  $\tau \approx 0.89$ ), balancing precision and recall under fixed thresholds. The H-measure contributed a cost-sensitive and decision-theoretic dimension to model evaluation, offering valuable nuance despite some sensitivity to parameter assumptions. PR-AUC, although threshold-free, provided complementary insights by ranking models based on the trade-offs between positive-class precision and recall.

We note that strong ROC-AUC performance does not necessarily imply well-calibrated predicted probabilities. Calibration quality (e.g., assessed via the Brier score or expected calibration error) represents an orthogonal consideration that may be critical in cost-sensitive, rare-event deployments, but is beyond the scope of the present evaluation-focused analysis.

These findings offer a clear recommendation: ROC-AUC should no longer be the default evaluation metric in rare-event classification. Instead, researchers and practitioners should adopt a multi-metric reporting strategy, led by MCC and  $F_2$  for threshold-based evaluation, with PR-AUC and H-measure used to provide additional perspectives on model ranking and cost trade-offs. This approach enables a statistically sound and operationally relevant understanding of model performance, particularly in high-stakes domains where misclassification costs are asymmetric and minority detection is critical.

### *Limitations and Future Work*

Despite the strength of the evidence presented, several limitations warrant acknowledgment. (i) The empirical analysis is restricted to three publicly available datasets spanning finance, bioinformatics, and environmental monitoring; evaluating additional domains, such as cybersecurity intrusion detection, clinical event prediction, and autonomous driving, would better assess external validity under diverse operational constraints. (ii) The study focuses exclusively on tabular data. The recommended metric bundle, MCC and  $F_2$  as threshold-dependent metrics, complemented by PR-AUC and H-measure as threshold-free and cost-sensitive views, is theoretically model-agnostic and modality-agnostic, as it operates on predicted scores and confusion matrices rather than raw features.

However, metric behavior may vary in high-dimensional or unstructured domains (e.g., computer vision and natural language processing) due to differences in score calibration, class-conditional distributions, and training practices such as focal loss, class-balanced reweighting, or data augmentation. Future studies on non-tabular benchmarks employing contemporary architectures

(e.g., CNNs/ViTs for imaging; transformers for text) and modality-appropriate imbalance treatments, with explicit attention to calibration and clinically/operationally relevant operating regions (e.g., low-FPR screening), could test generalizability and reveal any modality-specific adjustments (e.g., alternative thresholding policies or H-measure cost priors). (iii) Adaptive threshold-selection procedures and cost-sensitive loss functions were not considered; integrating such mechanisms may further align MCC and  $F_2$  with stakeholder risk tolerances and deployment objectives. (iv) Dynamic settings, including streaming data and concept drift, were outside the scope; examining how MCC,  $F_2$ , H-measure, and PR-AUC perform under temporal and distributional shifts would inform use in evolving systems. (v) Finally, while the analysis employed bootstrap confidence intervals and rank-based statistical tests, future work could leverage more advanced inferential frameworks, such as Bayesian ranking models or multi-metric decision analysis, to strengthen the reliability of metric comparisons in extremely imbalanced regimes.

**Author Contributions:** Conceptualization, M.I.; Methodology, M.I.; Software, M.I.; Validation, M.I., M.J., and A.B.; Formal analysis, M.I.; Investigation, M.I. and M.J.; Resources, M.I. and M.J.; Data curation, M.I. and M.J.; Writing—original draft preparation, M.I.; Writing—review and editing, M.J., A.B., and H.R.A.; Visualization, M.I.; Supervision, A.B., and H.R.A.; Project administration, M.I. 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.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The datasets analyzed in this study are openly available from Kaggle—Credit Card Fraud Detection [44]—and from the UCI Machine Learning Repository: Yeast [45] and Ozone Level Detection [46].

**Acknowledgments:** The authors used ChatGPT solely for improving clarity and grammar during manuscript preparation. No part of the scientific content, data analysis, or conclusions was generated by AI. All interpretations and results are the work of the authors.

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

## References

1. Elkan, C. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, Seattle, WA, USA, 4–10 August 2001.
2. Kubat, M.; Matwin, S. Addressing the curse of imbalanced training sets: One-sided selection. In *Proceedings of the 14th International Conference on Machine Learning*, Nashville, TN, USA, 8–12 July 1997.
3. Liu, X.-Y.; Wu, J.; Zhou, Z.-H. Exploratory undersampling for class-imbalance learning. *IEEE Trans. Syst., Man, Cybern., Part B (Cybern.)* **2009**, *39*, 539–550. <https://doi.org/10.1109/TSMCB.2008.2007853>.
4. Drummond, C.; Holte, R.C. C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *Proceedings of the ICML 2003 Workshop on Learning from Imbalanced Data Sets II*, Washington, DC, USA, 21–24 August 2003.
5. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. <https://doi.org/10.1613/jair.953>.
6. Han, H.; Wang, W.-Y.; Mao, B.-H. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning. In *Proceedings of the International Conference on Intelligent Computing (ICIC 2005)*, Hefei, China, 23–26 August 2005. [https://doi.org/10.1007/11538059\\_91](https://doi.org/10.1007/11538059_91).
7. Nguyen, H.M.; Cooper, E.W.; Kamei, K. Borderline over-sampling for imbalanced data classification. *Int. J. Knowl. Eng. Soft Data Paradigms* **2011**, *3*, 4–21. <https://doi.org/10.1504/IJKESDP.2011.039875>.

8. Bunkhumpornpat, C.; Sinapiromsaran, K.; Lursinsap, C. Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem. In *Proceedings of the 13th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2009)*, Bangkok, Thailand, 27–30 April 2009. [https://doi.org/10.1007/978-3-642-01307-2\\_43](https://doi.org/10.1007/978-3-642-01307-2_43).
9. He, H.; Bai, Y.; Garcia, E.A.; Li, S. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN 2008)*, Hong Kong, China, 1–6 June 2008. <https://doi.org/10.1109/IJCNN.2008.4633969>.
10. Tomek, I. Two modifications of CNN. *IEEE Trans. Syst., Man, Cybern.* **1976**, *6*, 769–772. <https://doi.org/10.1109/TSMC.1976.4309452>.
11. Wilson, D.L. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans. Syst., Man, Cybern.* **1972**, *2*, 408–421. <https://doi.org/10.1109/TSMC.1972.4309137>.
12. Batista, G.E.A.P.A.; Prati, R.C.; Monard, M.C. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl.* **2004**, *6*, 20–29. <https://doi.org/10.1145/1007730.1007735>.
13. Imani, M.; Arabnia, H.R. Hyperparameter optimization and combined data sampling techniques in machine learning for customer churn prediction: A comparative analysis. *Technologies* **2023**, *11*, 167. <https://doi.org/10.3390/technologies11060167>.
14. Douzas, G.; Bacao, F. Geometric SMOTE: A geometrically enhanced drop-in replacement for SMOTE. *Inf. Sci.* **2019**, *501*, 118–135. <https://doi.org/10.1016/j.ins.2019.06.007>.
15. Mariani, G.; Scheidegger, F.; Istrate, R.; Alakuijala, J.; Bekas, C.; Malossi, A. BAGAN: Data augmentation with balancing GAN. *arXiv* **2018**, preprint. <https://doi.org/10.48550/arXiv.1803.09655>.
16. Wei, Q.; Liu, J. Theoretical analysis of synthetic sampling for class imbalance. *Mach. Learn.* **2023**, *112*, 4073–4102.
17. 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. <https://doi.org/10.1148/radiology.143.1.7063747>.
18. Fawcett, T. An introduction to ROC analysis. *Pattern Recognit. Lett.* **2006**, *27*, 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>.
19. Saito, T.; Rehmsmeier, M. The precision–recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* **2015**, *10*, e0118432. <https://doi.org/10.1371/journal.pone.0118432>.
20. Davis, J.; Goadrich, M. The relationship between precision–recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML 2006)*, Pittsburgh, PA, USA, 25–29 June 2006. <https://doi.org/10.1145/1143844.1143874>.
21. Hand, D.J. Measuring classifier performance: A coherent alternative to the area under the ROC curve. *Mach. Learn.* **2009**, *77*, 103–123. <https://doi.org/10.1007/s10994-009-5119-5>.
22. Matthews, B.W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta Protein Struct.* **1975**, *405*, 442–451. [https://doi.org/10.1016/0005-2795\(75\)90109-9](https://doi.org/10.1016/0005-2795(75)90109-9).
23. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **2020**, *21*, 6. <https://doi.org/10.1186/s12864-019-6413-7>.
24. Christen, P.; Ranbaduge, T.; Vatsalan, A. A review of the F-measure: Its history, properties, criticism, and alternatives. *ACM Comput. Surv.* **2023**, *56*, 73. <https://doi.org/10.1145/3606367>.
25. Powers, D.M.W. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *J. Mach. Learn. Technol.* **2011**, *2*, 37–63.
26. Brodersen, K.H.; Ong, C.S.; Stephan, K.E.; Buhmann, J.M. The balanced accuracy and its posterior distribution. In *Proceedings of the 20th International Conference on Pattern Recognition*, Istanbul, Turkey, 23–26 August 2010. <https://doi.org/10.1109/ICPR.2010.764>.
27. Cohen, J. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. <https://doi.org/10.1177/001316446002000104>.
28. Landis, J.R.; Koch, G.G. The measurement of observer agreement for categorical data. *Biometrics* **1977**, *33*, 159–174. <https://doi.org/10.2307/2529310>.

29. 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. <https://doi.org/10.1186/s13040-023-00322-4>.
30. Richardson, E.; Trevizani, R.; Greenbaum, J.A.; Carter, H.; Nielsen, M.; Peters, B. The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns* **2024**, *5*, 100994. <https://doi.org/10.1016/j.patter.2024.100994>.
31. Provost, F.; Fawcett, T. Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, Newport Beach, CA, USA, 14–17 August 1997.
32. Sokolova, M.; Lapalme, G. A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* **2009**, *45*, 427–437. <https://doi.org/10.1016/j.ipm.2009.03.002>
33. McDermott, M.B.A.; Zhang, H.; Hansen, L.H.; Angelotti, G.; Gallifant, J. A closer look at AUROC and AUPRC under class imbalance. In *Adv. Neural Inf. Process. Syst.* *37*, Vancouver, Canada, 10–15 December 2024.
34. Itaya, Y.; Tamura, J.; Hayashi, K.; Yamamoto, K. Asymptotic properties of Matthews correlation coefficient. *Stat. Med.* **2025**, *44*, e10303.
35. Hernández-Orallo, J. ROC curves for regression. *Pattern Recognit.* **2013**, *46*, 3395–3411. <https://doi.org/10.1016/j.patcog.2013.06.014>
36. Hand, D.J. Evaluating diagnostic tests: The area under the ROC curve and the balance of errors. *Stat. Med.* **2010**, *29*, 1502–1510. <https://doi.org/10.1002/sim.3859>.
37. Niculescu-Mizil, A.; Caruana, R. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, Bonn, Germany, 7–11 August 2005. <https://doi.org/10.1145/1102351.1102430>.
38. Flach, P.A.; Kull, M. Precision–recall–gain curves: PR analysis done right. In *Adv. Neural Inf. Process. Syst.* *28*, Montréal, Canada, 7–12 December 2015.
39. He, H.; Garcia, E.A. Learning from imbalanced data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284. <https://doi.org/10.1109/TKDE.2008.239>.
40. Blagus, R.; Lusa, L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinform.* **2013**, *14*, 106. <https://doi.org/10.1186/1471-2105-14-106>
41. Imani, M.; Ghaderpour, Z.; Joudaki, M.; Beikmohammadi, A. The impact of SMOTE and ADASYN on random forest and advanced gradient boosting techniques in telecom customer churn prediction. In *Proceedings of the 10th International Conference on Web Research (ICWR)*, Tehran, Iran, 24–25 April 2024. <https://doi.org/10.1109/ICWR61162.2024.10533320>.
42. Imani, M.; Beikmohammadi, A.; Arabnia, H.R. Comprehensive analysis of random forest and XGBoost performance with SMOTE, ADASYN, and GNUS under varying imbalance levels. *Technologies* **2025**, *13*, 88. <https://doi.org/10.3390/technologies13030088>.
43. Imani, M.; Joudaki, M.; Beikmohamadi, A.; Arabnia, H.R. Customer churn prediction: A systematic review of recent advances, trends, and challenges in machine learning and deep learning. *Mach. Learn. Knowl. Extr.* **2025**, *7*, 105. <https://doi.org/10.3390/make7030105>.
44. Dal Pozzolo, A.; Boracchi, G.; Caelen, O.; Alippi, C.; Bontempi, G. Credit card fraud detection: A realistic modeling and a novel learning strategy. *IEEE Trans. Neural Netw. Learn. Syst.* **2018**, *29*, 3784–3797. <https://doi.org/10.1109/TNNLS.2017.2736643>.
45. Nakai, K. *Yeast* [dataset]. *UCI Machine Learning Repository*. Available online: <https://doi.org/10.24432/C5KG68> (accessed on 4 November 2025).
46. Zhang, K.; Fan, W.; Yuan, X. *Ozone Level Detection* [dataset]. *UCI Machine Learning Repository*. Available online: <https://doi.org/10.24432/C5NG6W> (accessed on 4 November 2025).
47. Alcalá-Fdez, J.; Sánchez, L.; García, S.; del Jesús, M.J.; Ventura, S.; Garrell, J.M.; Otero, J.; Romero, C.; Bacardit, J.; Rivas, V.M.; Fernández, J.C.; Herrera, F. KEEL: A software tool to assess evolutionary algorithms for data mining problems. *Soft Comput.* **2009**, *13*, 307–318. <https://doi.org/10.1007/s00500-008-0323-y>.

48. Aggarwal, C.C. *Data Mining: The Textbook*; Springer: Cham, Switzerland, **2015**; pp. 1–734. <https://doi.org/10.1007/978-3-319-14142-8>.
49. Liu, Y.; Wang, Y.; Chen, X.; Woon, Y.H.; Wong, K.Y. Standardization and normalization of quantitative data in medical research. *J. Biomed. Res.* **2008**, *22*, 193–200.
50. Kendall, M.G. A new measure of rank correlation. *Biometrika* **1938**, *30*, 81–93. <https://doi.org/10.1093/biomet/30.1-2.81>.
51. Abdi, H. The Kendall rank correlation coefficient. In *Encyclopedia of Measurement and Statistics*; Salkind, N.J., Ed.; Sage Publications: Thousand Oaks, CA, USA, **2007**; pp. 508–510. <https://doi.org/10.4135/9781412952644.n239>.
52. Conover, W.J. *Practical Nonparametric Statistics*, 3rd ed.; John Wiley & Sons: New York, NY, USA, **1999**; pp. 1–608.
53. 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. <https://doi.org/10.2307/2531595>.
54. Demšar, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.* **2006**, *7*, 1–30.
55. Hand, D.J.; Anagnostopoulos, C. A better Beta for the H-measure of classification performance. *Pattern Recognit. Lett.* **2014**, *40*, 41–46. <https://doi.org/10.1016/j.patrec.2013.12.011>.

**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.