International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5174 -5186
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0677

When Does Cost-Sensitive Weighting Matter? A Classifier-Capacity Analysis for Imbalanced Classification

Swati Satpute[1], Ajit More[2]
[1]Research Scholar, Dept. of Computer Science, Yashwantrao Mohite College of Arts, Science and Commerce, Bharati Vidyapeeth (Deemed to be University), Pune, Maharashtra, India
[2]Professor, Institute of Management and Entrepreneurship Development, Bharati Vidyapeeth (Deemed to be University), Pune, Maharashtra, India
Emails: *swati.satpute@gmail.com[1], ajit.more@bharatividyapeeth.edu[2]*

Abstract
Class weighting is the most widely used cost-sensitive remedy for class imbalance, and inverse-frequency "balanced" weights are often applied as an unquestioned default. This paper asks whether the choice among competing weighting mechanisms actually matters, and for which classifiers. Seven mechanisms, namely Balanced, Cost-Matrix, Sqrt-InverseFreq, Log-Damped, Effective-Number, a Search-Tuned power scheme, and focal loss, were evaluated on twenty-four benchmark datasets whose imbalance ratios reach 85:1, under three base classifiers of increasing representational capacity: logistic regression, a soft-voting ensemble of logistic regression, random forest, and a support vector machine, and XGBoost. Using leakage-free nested cross-validation and the Friedman–Nemenyi procedure applied separately to each classifier, mechanism choice produced large, highly significant differences under logistic regression ($p < 0.0001$), differences that shrank to non-significance under the voting ensemble ($p = 0.179$) and became indistinguishable from a random ordering under XGBoost ($p = 0.955$, average-rank spread of 0.46). At the same time, the two high-capacity classifiers outperformed logistic regression under every mechanism tested. The evidence supports a simple two-tier policy: tune the weighting scheme carefully for linear learners, and keep the plain balanced default for ensemble and boosting learners, where additional tuning yields no benefit that survives a significance test.

Keywords: *Imbalanced classification, cost-sensitive learning, class weighting, classifier capacity, ensemble learning, XGBoost, Friedman test, Matthews correlation coefficient*

1. Introduction
Class imbalance remains one of the most persistent practical obstacles in applied machine learning: when one class dominates the training sample, standard learners drift toward majority-class predictions and the minority class, which is usually the class of interest, is the one that suffers [1]. The problem appears in fraud detection, rare-disease screening, fault diagnosis, and many other settings, and a large body of techniques has grown around it [2]. Among these techniques, cost-sensitive weighting is probably the cheapest to deploy. The training loss is reweighted so that minority errors cost more, and most software stacks ship a ready-made heuristic for the purpose; scikit-learn's class_weight='balanced' option, which assigns each class a weight inversely proportional to its frequency, traces back to the rare-events correction of King and Zeng [3]. Yet this heuristic is a single point in a much broader design space. Weights can be damped, tuned, derived from a cost matrix, or replaced altogether by a dynamic per-example scheme, and it is not obvious that any one choice should win consistently. Even less obvious is whether the winner depends on the classifier underneath it, a question that, to the best of our knowledge, has not been examined systematically at the mechanism-selection level. This paper takes up both questions empirically. Starting from an earlier comparison of seven weighting mechanisms on fourteen datasets with logistic regression as the only base learner, the study is extended to twenty-four public benchmark datasets and three base classifiers that span a wide capacity range: logistic regression, a soft-voting ensemble that combines logistic regression, a random forest, and a support vector

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5174 -5186
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0677

machine, and XGBoost. The headline finding is an interaction effect that is easy to state: mechanism choice matters a great deal for the linear model and essentially not at all for the two high-capacity learners. The contributions are fourfold. First, seven cost-sensitive mechanisms are compared on twenty-four imbalanced benchmarks under a leakage-free nested cross-validation protocol. Second, the comparison is repeated under three base classifiers through a single mechanism-agnostic per-instance weighting interface, so that classifier families are compared on an equal footing. Third, the Friedman–Nemenyi procedure, run separately per classifier, confirms a monotonic collapse of mechanism sensitivity as classifier capacity grows. Fourth, these observations are distilled into a two-tier, classifier-conditional selection rule that removes an entire tuning dimension for practitioners who already use strong ensemble or boosting models.

2. Related Work

The theoretical footing of cost-sensitive learning was laid by Elkan [4], who showed that optimal decisions follow from the true misclassification cost matrix rather than from class frequencies. Zadrozny et al. [5] established that any classifier can be made cost-sensitive by weighting examples in proportion to their cost, which is the route taken here through the per-instance weighting interface. Interest in the approach has not faded; a recent review of cost-sensitive learning for imbalanced medical data covers 173 studies published since 2010 and reports a marked acceleration after 2020 [6]. On the loss-design side, Cui et al. [7] damp inverse-frequency weights through an effective number of samples, while focal loss reshapes the objective per example rather than per class [8]. The second type of literature is related to the sensitivity of various classifier families to imbalance, in the first place. Random forests also have been shown to degrade more gracefully than linear models when it comes to imbalance, requiring smaller adjustments to recover minority performance, Chen et al. [9] noted. Even though there is no explicit reweighting, Sun et al. [10] observed that boosting already is doing some implicit adaptive reweighting, because it focuses on the misclassified examples round after round. Furthermore, tree ensembles continue to be the best general-purpose learners for tabular data [11] and increasingly, it has been shown that extreme imbalance corrections may not be needed, or even beneficial, for well-specified models [12]. These findings provide a basis for a hypothesis which is directly addressed in this paper: flexible learners may compensate for imbalance structurally, making the choice among external weighting mechanisms less important to them than to a linear model. According to Demšar [13] the Friedman rank test with the Nemenyi post-hoc procedure should be used when many methods are compared on many data sets. That same protocol is followed here, except for one intentional modification: it has been applied once for each base classifier, rather than once for the entire ensemble, and it is this that allows a classifier-conditional effect to manifest itself rather than to be averaged out.

3. Cost-Sensitive Weighting Mechanisms

Throughout, N denotes the number of training instances, K the number of classes, and n_j the size of class j. Each mechanism produces one weight per class, which is then attached to every instance of that class and passed to the learner at fit time. The Balanced scheme [3] is the familiar inverse-frequency rule,

$$w_j = \frac{N}{K n_j}$$

and the Cost-Matrix scheme sets the weight of class j directly to its misclassification cost following Elkan [4]; in the absence of externally validated costs, the empirical frequency ratio serves as a proxy here. Two damped variants temper the correction that (1) applies under severe imbalance: Sqrt-InverseFreq,

$$w_j \propto \frac{1}{\sqrt{n_j}}$$

and Log-Damped,

$$w_j \propto \ln\left(\frac{N}{n_j}\right)$$

The Effective-Number scheme of Cui et al. [7] replaces raw counts with the expected coverage of each class,

The Search-Tuned scheme, introduced in this work, treats the strength of the balanced correction as a

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5174 -5186
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0677

hyperparameter by raising the balanced weight to a power γ selected on training data only,

$$w_j \propto \frac{1-\beta}{1-\beta^{n_j}}, \quad \beta = 0.999$$

with γ drawn from {0.25, 0.5, 0.75, 1, 1.25, 1.5, 2, 3} by an inner cross-validation loop. Setting γ = 1 recovers (1), while γ < 1 damps the correction, so the search interpolates between no correction and an amplified one. Finally, focal loss [8] abandons fixed class weights altogether,

$$\mathrm{FL}(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$$

where p_t is the predicted probability of the true class. Because focal loss is a training objective rather than a weighting, it is implemented as its own binary logistic model, is identical no matter which base classifier the other mechanisms use, and is therefore reported separately in Section VI-F. In tables and figures the six weighting mechanisms are abbreviated Bal, CM, SIF, LD, EN, and ST, in the order introduced above.

4. Experimental Setup

4.1.Datasets

Table I lists the twenty-four public benchmark datasets, an expansion from the fourteen of the initial study. Most originate from the KEEL imbalanced-data repository [14] or the UCI archive, and their imbalance ratios (IR) were verified against the published values wherever available. Twenty problems are binary; four are multiclass and are marked with an asterisk in the results tables. PIMA, with an IR below 2, is retained deliberately as a mild-imbalance control, and Forest_Cover_Type was subsampled to a fixed long-tailed profile. The twenty binary datasets also serve as the test bed for focal loss. As shown in Table 1 Benchmark Datasets (IR = Majority:Minority Ratio).

Table 1 Benchmark Datasets (IR = Majority:Minority Ratio)

Dataset	Type	Instances	IR
Yeast4	Binary	1,484	28.10
Ecoli3	Binary	336	8.60
Page-Blocks0	Binary	5,472	8.79
Abalone9-18	Binary	731	16.40
Glass5	Binary	214	22.78
Mammography	Binary	11,183	≈42
Shuttle-c0-vs-c4	Binary	1,829	13.87
Vowel0	Binary	988	9.98
Poker (8 vs 6)	Binary	1,477	≈85.9
KDD-Cup (buffer_overflow)	Binary	2,233	≈73.4
Pendigits	Binary	10,992	≈9.4
Solar_Flare (F)	Binary	1,066	23.79
PIMA	Binary	768	1.87
Oil_Spill	Binary	937	21.85
Car_Eval_4	Binary	1,728	25.58
Ozone_Level	Binary	2,536	33.74
Chess	Binary	2,901	26.63
USCrime	Binary	1,994	12.29
CreditCard_Mini	Binary	5,000	9.16
Scene	Binary	2,407	12.60
Dermatology	Multi (6)	366	5.60
DryBean	Multi (7)	4,000	22.00

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5174 -5186
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0677

WineQuality	Multi (7)	6,497	567.20
Forest_Cover_Type	Multi (7)	6,000	103.13

4.2.Base Classifiers

Three base classifiers were held fixed within each experimental run. Logistic regression, L2-regularized, contributes a single linear decision boundary. The voting classifier is a soft-voting ensemble of logistic regression, a random forest of 300 trees, and an RBF-kernel support vector machine with probability outputs, so it blends a linear model, a bagged tree ensemble, and a margin-based non-linear model. XGBoost [15] uses 300 estimators, a maximum depth of 6, and a learning rate of 0.1. Every mechanism is applied to every classifier through the same per-instance sample-weight interface at fit time, rather than through each library's native class-weight argument, which XGBoost does not expose and the voting wrapper does not accept; this keeps the reweighting semantics identical across classifier families.

4.3.Validation Protocol

Each dataset–mechanism–classifier combination was evaluated with repeated stratified cross-validation, five folds repeated three times for fifteen outer folds in total. The Search-Tuned exponent was chosen by a stratified three-fold inner loop nested inside each outer training fold, so no test information ever reaches the tuning step.

4.4.Performance Metrics

Nine measures were recorded per fold: PR-AUC, macro precision, macro recall, macro F1, balanced accuracy, G-mean, the Matthews correlation coefficient, and minority-class precision and recall. The Matthews correlation coefficient,

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

serves as the primary metric for all statistical comparisons. It uses every cell of the confusion matrix, generalizes cleanly to the multiclass case, and cannot be inflated by majority-class-only prediction, properties that have led to its recommendation over both accuracy-type summaries [16] and the ROC AUC [17].

5. Statistical Methodology

For each classifier separately, the six class-weighting mechanisms were ranked 1 to 6 on every dataset by mean MCC, and the Friedman statistic [18] was computed over the resulting 24 × 6 rank matrix,

$$\chi_F^2 = \frac{12}{N\,K(K+1)} \sum_{j=1}^{K} R_j^2 - 3N(K+1)$$

with N = 24 datasets, K = 6 mechanisms, and R_j the rank sum of mechanism j. Wherever the Friedman test rejected at $\alpha = 0.05$, the Nemenyi critical difference [19]

$$\mathrm{CD} = q_\alpha \sqrt{\frac{K(K+1)}{6N}}$$

determined which pairs of average ranks differ significantly; for this design CD = 1.54. Running the whole procedure once per base classifier, as Demšar's protocol [13] permits, is what makes the classifier-conditional question testable at all: a single pooled analysis would blend three different answers into one.

6. Results

6.1.Overview

Table II states the central result before the per-classifier detail: the statistical significance of mechanism choice collapses as classifier capacity increases. The same critical difference of 1.54 applies to all three rows, so the three Friedman outcomes are directly comparable.

Table 2 Friedman Test Summary Per Base Classifier (N = 24, K = 6)

Base classifier	χ^2F	p-value	CD	Significant
Logistic Regression	49.23	< 0.0001	1.54	Yes

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5174 -5186
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0677

Voting (LR+RF+SVM)	7.61	0.179	1.54	No
XGBoost	1.09	0.955	1.54	No

6.2.Logistic Regression

The critical-difference diagram and mean MCC per dataset and mechanism under logistic regression are presented in Table III and Fig. 1, respectively. Search-Tuned (2.12) and Sqrt-InverseFreq (2.29) are clearly separated from the rest, with the difference between them and Cost-Matrix (4.19) and Balanced (4.31) and Log-Damped (4.90) all greater than the critical difference. The two damped schemes outperform the naive inverse-frequency rule on the overwhelming majority of datasets, with sometimes huge differences, such as on Mammography (0.662 vs. 0.522) and Yeast4 (0.409 vs. 0.298). This replication is also a strengthening of the conclusions of the initial study and has ten more datasets than the original. As shown in Table 3 Mean Mcc Per Dataset and Mechanism, Logistic Regression. Figure 1 Critical-difference diagram under logistic regression (Friedman $p < 0.0001$). Mechanisms joined by a red bar are statistically indistinguishable at $\alpha = 0.05$.

Table 3 Mean Mcc Per Dataset and Mechanism, Logistic Regression

Dataset	Bal	CM	SIF	LD	EN	ST
Abalone9-18	0.461	0.478	0.580	0.332	0.482	0.500
Car_Eval_4	0.820	0.826	0.866	0.796	0.826	0.883
Chess	0.765	0.770	0.827	0.681	0.815	0.819
CreditCard_Mini	0.839	0.838	0.886	0.751	0.886	0.897
Dermatology	0.964	0.962	0.963	0.963	0.964	0.958
DryBean	0.894	0.894	0.900	0.898	0.898	0.901
Ecoli3	0.551	0.567	0.637	0.523	0.574	0.625
Forest_Cover	0.552	0.556	0.570	0.568	0.561	0.571
Glass5	0.644	0.703	0.668	0.508	0.652	0.669
KDD-Cup	0.987	0.987	0.987	0.987	0.987	0.987
Mammography	0.522	0.523	0.662	0.358	0.633	0.671
Oil_Spill	0.504	0.479	0.521	0.409	0.497	0.534
Ozone_Level	0.295	0.290	0.355	0.253	0.332	0.314
Page-Blocks0	0.675	0.677	0.710	0.586	0.706	0.709
Pendigits	0.633	0.633	0.716	0.541	0.701	0.753
PIMA	0.559	0.559	0.514	0.574	0.550	0.568
Poker	-0.025	-0.025	0.000	-0.014	0.006	0.026

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5174 -5186
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0677

Scene	0.162	0.155	0.148	0.155	0.161	0.170
Shuttle-c0-c4	0.996	0.996	0.996	0.996	0.996	0.996
Solar_Flare	0.320	0.316	0.326	0.314	0.317	0.310
USCrime	0.452	0.452	0.488	0.409	0.470	0.477
Vowel0	0.775	0.791	0.815	0.715	0.788	0.833
WineQuality	0.179	0.180	0.286	0.291	0.242	0.284
Yeast4	0.298	0.297	0.409	0.201	0.336	0.384

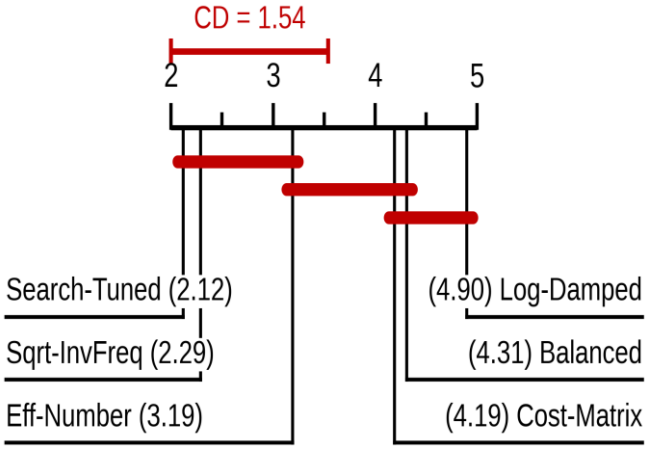

Figure 1 Critical-difference diagram under logistic regression (Friedman p < 0.0001). Mechanisms joined by a red bar are statistically indistinguishable at α = 0.05

6.3.Voting Ensemble

Table IV repeats the comparison under the voting ensemble and Fig. 2 gives the critical-difference diagram. The picture changes completely. Cost-Matrix now holds the best nominal rank (2.98), just ahead of Search-Tuned (3.08), but the whole rank range spans only 2.98 to 4.12, well inside the critical difference, and the Friedman test does not reject (p = 0.179). Every mechanism sits in one connected group, so the apparent ordering carries no statistical weight given this sample of datasets.As shown in Figure 2 Critical-difference diagram under the voting ensemble (Friedman p = 0.179). All six mechanisms fall inside a single connected group.

Table 4 Mean Mcc Per Dataset and Mechanism, Voting (Lr+Rf+Svm)

Dataset	Bal	CM	SIF	LD	EN	ST
Abalone9-18	0.384	0.418	0.469	0.380	0.386	0.471
Car_Eval_4	0.896	0.906	0.911	0.893	0.903	0.913
Chess	0.893	0.894	0.900	0.886	0.902	0.892
CreditCard_Mini	0.900	0.902	0.909	0.892	0.911	0.908
Dermatology	0.965	0.964	0.964	0.965	0.965	0.964
DryBean	0.904	0.906	0.907	0.906	0.905	0.906
Ecoli3	0.617	0.620	0.561	0.640	0.633	0.604

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5174 -5186
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0677

Forest_Cover	0.645	0.656	0.637	0.638	0.640	0.640
Glass5	0.685	0.759	0.598	0.852	0.652	0.765
KDD-Cup	0.997	0.997	0.997	0.997	0.997	0.997
Mammography	0.754	0.752	0.750	0.723	0.745	0.749
Oil_Spill	0.584	0.565	0.637	0.537	0.602	0.626
Ozone_Level	0.318	0.332	0.283	0.361	0.328	0.364
Page-Blocks0	0.819	0.828	0.844	0.811	0.839	0.840
Pendigits	0.980	0.980	0.971	0.980	0.976	0.978
PIMA	0.677	0.687	0.675	0.688	0.680	0.692
Poker	0.614	0.646	0.190	0.715	0.512	0.727
Scene	0.232	0.232	0.207	0.211	0.208	0.218
Shuttle-c0-c4	0.996	0.996	0.996	0.996	0.996	0.996
Solar_Flare	0.173	0.177	0.175	0.178	0.178	0.129
USCrime	0.473	0.483	0.481	0.480	0.487	0.480
Vowel0	0.998	0.996	0.994	0.996	0.998	0.996
WineQuality	0.484	0.487	0.458	0.462	0.469	0.493
Yeast4	0.472	0.445	0.346	0.418	0.419	0.416

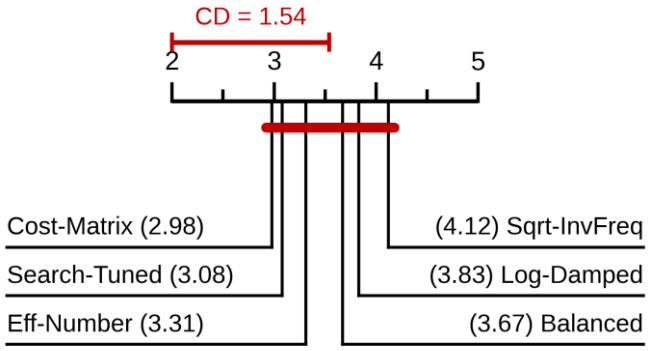

Figure 2 Critical-difference diagram under the voting ensemble (Friedman p = 0.179). All six mechanisms fall inside a single connected group

6.4.XGBoost

Table V and Fig. 3 complete the sweep with XGBoost. The rank spread compresses further, from 3.29 to 3.75, less than a third of the critical difference, and the Friedman p-value of 0.955 means the observed ordering is about as informative as a shuffled one. For this classifier, on this benchmark, mechanism choice is simply not a lever that moves performance.As shown in Figure 3 Critical-difference diagram under XGBoost (Friedman p = 0.955). The rank spread of 0.46 is far below the critical difference of 1.54

Table 5 Mean Mcc Per Dataset and Mechanism, Xgboost

Dataset	Bal	CM	SIF	LD	EN	ST
Abalone9-18	0.390	0.372	0.378	0.403	0.388	0.357
Car_Eval_4	0.966	0.978	0.966	0.895	0.973	0.970
Chess	0.915	0.921	0.922	0.844	0.926	0.911
CreditCard_Mini	0.907	0.908	0.910	0.901	0.910	0.907
Dermatology	0.956	0.960	0.952	0.953	0.958	0.958

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5174 -5186
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0677

DryBean	0.896	0.896	0.897	0.896	0.897	0.897
Ecoli3	0.582	0.576	0.576	0.583	0.577	0.599
Forest_Cover	0.703	0.708	0.703	0.702	0.705	0.702
Glass5	0.719	0.739	0.553	0.557	0.676	0.533
KDD-Cup	0.972	0.972	0.972	0.972	0.972	0.972
Mammography	0.730	0.739	0.746	0.730	0.736	0.743
Oil_Spill	0.579	0.578	0.553	0.581	0.559	0.554
Ozone_Level	0.294	0.292	0.243	0.322	0.282	0.339
Page-Blocks0	0.864	0.864	0.867	0.860	0.865	0.863
Pendigits	0.955	0.958	0.953	0.962	0.952	0.959
PIMA	0.740	0.740	0.737	0.725	0.735	0.742
Poker	0.348	0.439	0.350	0.397	0.366	0.305
Scene	0.235	0.228	0.239	0.258	0.238	0.236
Shuttle-c0-c4	1.000	1.000	1.000	1.000	1.000	1.000
Solar_Flare	0.237	0.214	0.203	0.286	0.207	0.247
USCrime	0.478	0.486	0.484	0.486	0.484	0.489
Vowel0	0.958	0.950	0.972	0.958	0.958	0.966
WineQuality	0.469	0.478	0.489	0.483	0.484	0.488
Yeast4	0.389	0.361	0.358	0.398	0.378	0.382

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5174 -5186
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0677

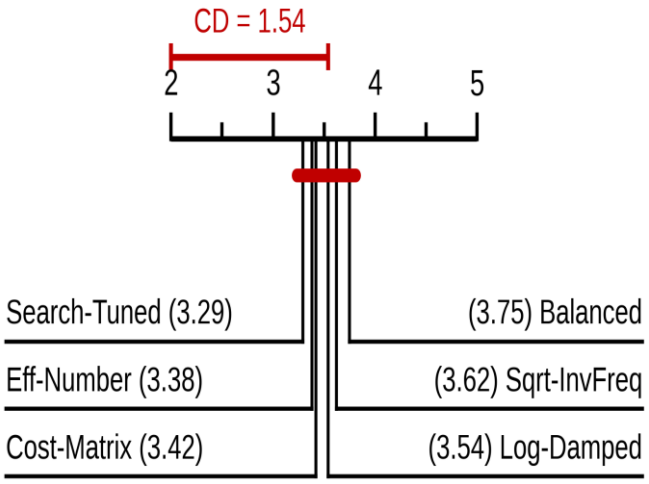

Figure 3 Critical-difference diagram under XGBoost (Friedman p = 0.955). The rank spread of 0.46 is far below the critical difference of 1.54

6.5.Cross-Classifier Comparison

Tables VI and VII place the three classifiers side by side, and Figs. 4 and 5 visualize the same numbers. Two patterns stand out. Across mechanisms, the logistic-regression bars in Fig. 4 swing widely while the voting and XGBoost bars stay nearly level, and the rank curves in Fig. 5 tell the identical story from the ordinal side. Across classifiers, voting and XGBoost sit above logistic regression under every single mechanism, so the choice of classifier dominates the choice of weighting scheme wherever the two conflict.As shown in Table 6 Average Rank of Each Mechanism by Base Classifier, Table 7 Mean Mcc of Each Mechanism by Base Classifier. As shown in Figure 4 Mean MCC of each mechanism under the three base classifiers, averaged over all 24 datasets. Mechanism abbreviations follow Section III

Table 6 Average Rank of Each Mechanism by Base Classifier

Mechanism	Log. Reg.	Voting	XGBoost
Balanced	4.31	3.67	3.75
Cost-Matrix	4.19	2.98	3.42
Sqrt-InverseFreq	2.29	4.12	3.62
Log-Damped	4.90	3.83	3.54
Effective-Number	3.19	3.31	3.38
Search-Tuned	2.12	3.08	3.29

Table 7 Mean Mcc of Each Mechanism by Base Classifier

Mechanism	Log. Reg.	Voting	XGBoost
Balanced	0.5759	0.6858	0.6784
Cost-Matrix	0.5793	0.6928	0.6816
Sqrt-InverseFreq	0.6179	0.6608	0.6677
Log-Damped	0.5331	0.6918	0.6730
Effective-Number	0.5992	0.6804	0.6760
Search-Tuned	0.6184	0.6985	0.6716

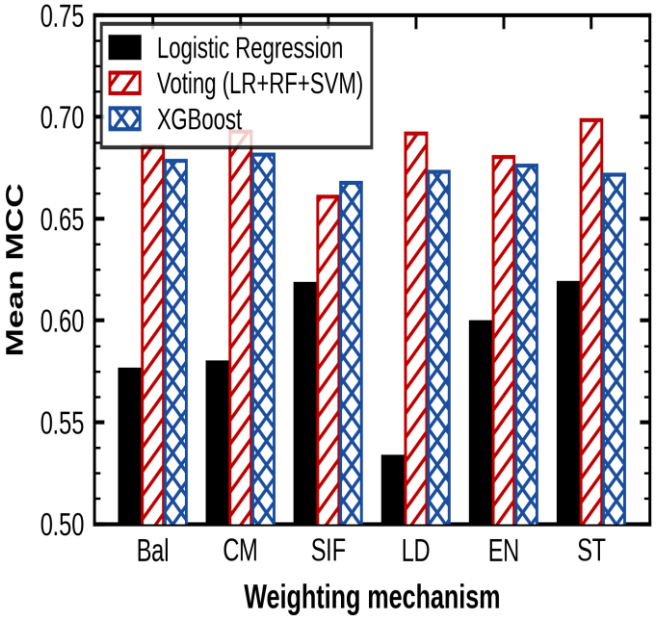

Figure 4 Mean MCC of each mechanism under the three base classifiers, averaged over all 24 datasets. Mechanism abbreviations follow Section III

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5174 -5186
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0677

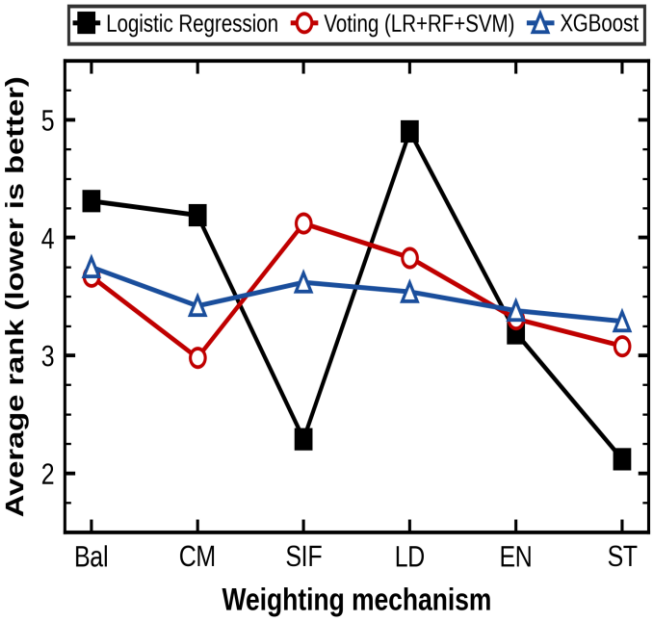

Figure 5 Average Friedman rank of each mechanism under the three base classifiers. The logistic-regression curve swings widely; the two high-capacity curves stay nearly flat

Table VIII condenses the study into three numbers per classifier, and Fig. 6 plots them: from logistic regression to XGBoost the rank spread falls monotonically from 2.77 to 1.15 to 0.46, roughly a six-fold reduction, while the Friedman p-value climbs from below 0.0001 to 0.955. The mean MCC column shows that this collapse in sensitivity costs nothing in absolute performance; both high-capacity classifiers beat the linear model by around nine MCC points on average. As shown in Table 8 Rank Spread, Friedman P-Value, And Overall Mean Mcc by Classifier.As shown in Figure 5 Average Friedman rank of each mechanism under the three base classifiers. The logistic-regression curve swings widely; the two high-capacity curves stay nearly flat. Figure 6 Sensitivity of performance to mechanism choice, by base classifier. As shown in Rank spread and Friedman significance collapse together as capacity grows

Table 8 Rank Spread, Friedman P-Value, And Overall Mean Mcc by Classifier

Base classifier	Rank spread	Friedman p	Mean MCC
Logistic Regression	2.77	< 0.0001	0.5873
Voting (LR+RF+SVM)	1.15	0.179	0.6850
XGBoost	0.46	0.955	0.6747

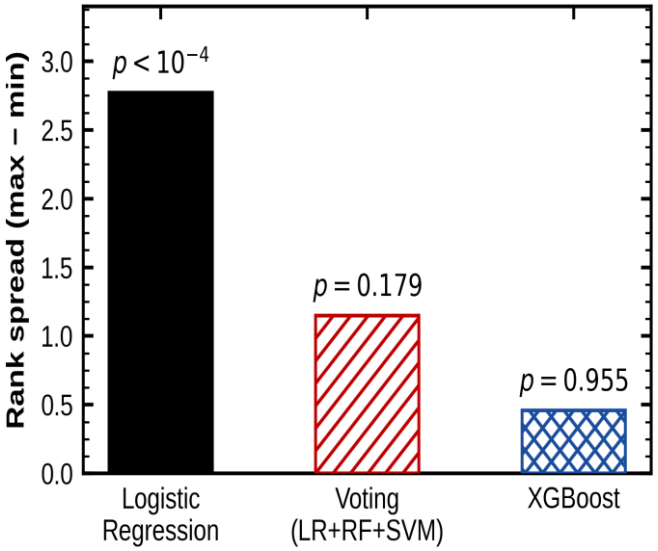

Figure 6 Sensitivity of performance to mechanism choice, by base classifier. Rank spread and Friedman significance collapse together as capacity grows

6.6.Focal Loss

Table IX reports focal loss on the twenty binary datasets. Since it runs as its own logistic model, its numbers are the same whichever base classifier the other six mechanisms use. Focal loss is broadly comparable to the weighting mechanisms under logistic regression but sits below the voting ensemble and XGBoost on most datasets, which is consistent with the reading developed in Section VII: once the learner itself, or in this case the loss, has enough flexibility to handle imbalance structurally, the particular fixed per-class weight matters much less.

Table 9 Mean Mcc of Focal Loss on The Twenty Binary Datasets

Dataset	MCC	Dataset	MCC

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5174 -5186
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0677

Abalone9-18	0.472	Page-Blocks0	0.648
Car_Eval_4	0.757	Pendigits	0.610
Chess	0.734	PIMA	0.540
CreditCard_Mini	0.845	Poker	-0.026
Ecoli3	0.577	Scene	0.167
Glass5	0.712	Shuttle-c0-c4	0.996
KDD-Cup	0.973	Solar_Flare	0.326
Mammography	0.510	USCrime	0.457
Oil_Spill	0.486	Vowel0	0.750
Ozone_Level	0.289	Yeast4	0.296

7. Discussion

There is a coherent reading of the results in Section VI that the more capability of the classifier to represent, the lower is the practical value of using a cost sensitive weighting mechanism carefully. It has a very real impact. A six-fold contraction of the rank spread, paired with a p-value that travels from unambiguous significance to 0.955, is a clean, monotonic collapse of a signal that is strong at one end of the capacity range and absent at the other. A feasible mechanism is not too elusive. The decision boundary for logistic regression is one hyperplane and directly reacts to the movement of the hyperplane by the change in the class weights and is visible. Random forests and support vector machines within the voting ensemble, and the boosted trees of XGBoost, on the other hand, construct piecewise decision regions that can separate minority pockets by choosing splits or placing support vectors, and which have little dependence on one global scalar. This is similar to the graceful degradation of random forests observed by Chen et al. [9] and the implicit reweighting effect of boosting, which by design emphasizes hard examples, which is why XGBoost has the least sensitivity of the three. The comparison also helps to understand which decision should be the one that the practitioner pays attention. Voting (mean MCC 0.685) and XGBoost (0.675) outperform logistic regression (0.587) across all mechanisms considered; the conclusion of this research programme – that classifier ranking is more important than mechanism selection – does not change across all the different mechanisms considered. One last thing to note is worth making as a cautionary tale against reading ranking tables: According to the voting method, Cost-Matrix is the best ranked among the two weakest mechanisms in logistic regression. If only the voting had been conducted and only the result of the vote at the top of the ranking reported, it would have appeared as a real discovery. Not so, according to the Friedman test: the voting and XGBoost ordering are statistically flat: the reversal is noise on a level response surface. Significance testing is not a formality in multi-classifier comparisons of this kind; it is the step that separates findings from artifacts.

8. Classifier-Conditional Selection Policy

The findings translate directly into a two-tier decision

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5174 -5186
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0677

rule that spares practitioners a full grid search over mechanisms and classifiers. For linear or otherwise low-capacity learners, use Search-Tuned when its nested search cost is acceptable, or Sqrt-InverseFreq as a statistically equivalent alternative that requires no tuning at all; the two are separated by only 0.17 rank points, far inside the critical difference. For high-capacity ensemble or boosting learners, use plain Balanced weighting and stop there, because no tested alternative is statistically distinguishable from it and the simplest option wins on cost. In either tier, treat the classifier itself as the primary performance decision. The rule also says where the saved computation should go. The expensive step this study renders unnecessary for strong classifiers, an inner cross-validated search over weighting schemes, can be redirected toward the data-level side of the imbalance toolbox, where SMOTE [20] and its many descendants operate on the training distribution rather than the loss [21]. Whether the capacity-conditional pattern found here for loss-level corrections also governs resampling, and how the two families combine with modern ensembles [22], is an open empirical question and, in our view, the natural next experiment.

9. Limitations

Four caveats bound the conclusions. The study covers cost-sensitive weighting in isolation, so nothing here establishes how resampling interacts with classifier capacity. The Cost-Matrix mechanism was evaluated with the empirical frequency ratio standing in for true misclassification costs, which likely understates its value in domains where validated, non-frequency costs exist; mammography screening is a concrete example. The focal-loss implementation is binary-only, which is why it was scored on twenty datasets and kept out of the primary six-mechanism test. Finally, twenty-four datasets sit at the lower end of the range Demšar [13] recommends for post-hoc power, so the two non-significant results should be read as “no effect detected at this sample size” rather than as proof of exactly zero effect.

Conclusion and Future Work

Across twenty-four imbalanced benchmarks and three base classifiers, the importance of cost-sensitive mechanism selection turned out to be a property not of the imbalance problem but of the classifier applied to it. Under logistic regression, mechanism choice produces large, significant differences, with Search-Tuned and Sqrt-InverseFreq clearly ahead of the naive Balanced and fixed-ratio Cost-Matrix schemes. Under a voting ensemble the differences lose significance, and under XGBoost they vanish into noise, even as both classifiers outperform the linear model under every scheme tested. Representational capacity, it appears, substitutes for careful loss reweighting.

Future work follows three threads. The first is to test whether the same capacity-conditional pattern holds for resampling methods and for classifiers combined with them. The second is to widen the high-capacity roster to LightGBM, CatBoost, and deep tabular networks, to learn whether insensitivity to mechanism choice is a general property of strong learners or a trait of tree ensembles. The third is to make capacity itself measurable in advance, so that the qualitative pattern reported here becomes a quantitative, pre-experiment rule that tells a practitioner, before any training run, whether the weighting scheme is worth tuning at all.

References

[1]. H. He and E. A. Garcia, “Learning from imbalanced data,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[2]. S. Rezvani and X. Wang, “A broad review on class imbalance learning techniques,” Appl. Soft Comput., vol. 143, art. no. 110415, Aug. 2023.
[3]. G. King and L. Zeng, “Logistic regression in rare events data,” Political Anal., vol. 9, no. 2, pp. 137–163, 2001.
[4]. C. Elkan, “The foundations of cost-sensitive learning,” in Proc. 17th Int. Joint Conf. Artif. Intell. (IJCAI), Seattle, WA, USA, 2001, pp. 973–978.
[5]. B. Zadrozny, J. Langford, and N. Abe, “Cost-sensitive learning by cost-proportionate example weighting,” in Proc. 3rd IEEE Int. Conf. Data Mining (ICDM), Melbourne, FL, USA, 2003, pp. 435–442.
[6]. I. Araf, A. Idri, and I. Chairi, “Cost-sensitive learning for imbalanced medical data: a review,” Artif. Intell. Rev., vol. 57, no. 4, art.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 08 August 2026
Page No: 5174 -5186
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0677

no. 80, Mar. 2024.

[7]. Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Long Beach, CA, USA, 2019, pp. 9268–9277.

[8]. T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Venice, Italy, 2017, pp. 2980–2988.

[9]. C. Chen, A. Liaw, and L. Breiman, "Using random forest to learn imbalanced data," Dept. Statistics, Univ. California, Berkeley, CA, USA, Tech. Rep. 666, 2004.

[10]. Y. Sun, M. S. Kamel, A. K. C. Wong, and Y. Wang, "Cost-sensitive boosting for classification of imbalanced data," Pattern Recognit., vol. 40, no. 12, pp. 3358–3378, Dec. 2007.

[11]. L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" in Proc. Adv. Neural Inf. Process. Syst. (NeurIPS), New Orleans, LA, USA, 2022, pp. 507–520.

[12]. R. van den Goorbergh, M. van Smeden, D. Timmerman, and B. Van Calster, "The harm of class imbalance corrections for risk prediction models: illustration and simulation using logistic regression," J. Amer. Med. Inform. Assoc., vol. 29, no. 9, pp. 1525–1534, Sep. 2022.

[13]. J. Demšar, "Statistical comparisons of classifiers over multiple data sets," J. Mach. Learn. Res., vol. 7, pp. 1–30, Jan. 2006.

[14]. J. Alcalá-Fdez et al., "KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework," J. Mult.-Valued Logic Soft Comput., vol. 17, no. 2–3, pp. 255–287, 2011.

[15]. T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining (KDD), San Francisco, CA, USA, 2016, pp. 785–794.

[16]. D. Chicco and G. Jurman, "The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," BMC Genomics, vol. 21, art. no. 6, Jan. 2020.

[17]. D. Chicco and G. Jurman, "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification," BioData Mining, vol. 16, art. no. 4, Feb. 2023.

[18]. M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," J. Amer. Statist. Assoc., vol. 32, no. 200, pp. 675–701, Dec. 1937.

[19]. P. B. Nemenyi, "Distribution-free multiple comparisons," Ph.D. dissertation, Princeton Univ., Princeton, NJ, USA, 1963.

[20]. N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002.

[21]. A. Fernández, S. García, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," J. Artif. Intell. Res., vol. 61, pp. 863–905, Apr. 2018.

[22]. A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: combination, implementation and evaluation," Expert Syst. Appl., vol. 244, art. no. 122778, Jun. 2024.