World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

A comparative study of ensemble learning techniques for imbalanced classification problems

Adefemi Ayodele *

Department of Computer Science and Digital Technologies, University of East London, London, England, United Kingdom.

World Journal of Advanced Research and Reviews, 2023, 19(01), 1633-1643

Publication history: Received on 04 June 2023; revised on 19 July 2023; accepted on 21 July 2023

Article DOI: https://doi.org/10.30574/wjarr.2023.19.1.1202

Abstract

Imbalanced classification problems frequently arise in critical domains such as fraud detection, medical diagnosis, cybersecurity, and anomaly detection, where the minority class often carries disproportionate importance despite its scarcity. Traditional machine learning algorithms tend to favour the majority class, leading to suboptimal performance and costly misclassifications in minority class detection. This study evaluates ensemble learning techniques—including Bagging, Boosting, Random Forest, EasyEnsemble, and BalancedRandomForest—for their effectiveness in managing class imbalance. Using several real-world benchmark datasets with varying imbalance ratios and feature complexities, the methods are rigorously assessed using metrics tailored to imbalanced scenarios, including F1-score, precision-recall area under the curve (PR-AUC), and geometric mean (G-mean). Results indicate that boosting-based methods, particularly Gradient Boosting Machines (GBM), consistently excel across most datasets, especially in terms of PR-AUC and G-mean. However, certain datasets with extreme imbalance or high feature dimensionality saw stronger performance from BalancedRandomForest. These findings underscore that the optimal ensemble method is highly dependent on specific dataset attributes and operational constraints. This analysis offers practical insights into aligning ensemble strategies with real-world requirements, guiding researchers and practitioners toward more robust and accurate models in imbalanced classification contexts.

Keywords: Imbalance; Ensemble; Bagging; Boosting; Resampling; Benchmark

1. Introduction

Classification problems often suffer from class imbalance, where one class significantly outnumbers another, leading to degraded model performance and frequent misclassification of the minority class [1][2]. This imbalance is especially critical in domains such as financial fraud detection, medical diagnostics, and cybersecurity, where misidentifying rare but important instances may lead to severe consequences—including financial losses, misinformed treatment decisions, or unaddressed security threats [1][3].

While previous studies have explored solutions for class imbalance, many rely on simplistic resampling techniques or overlook the trade-offs between precision and recall across different applications [2][4]. Furthermore, few studies have conducted a comprehensive evaluation of ensemble methods specifically adapted for imbalanced datasets across multiple real-world scenarios [5]. This study aims to address these gaps by systematically analysing ensemble learning techniques designed to mitigate the effects of imbalance.

Ensemble methods, which aggregate predictions from multiple base classifiers, have demonstrated improvements in model accuracy, reduction in variance, and enhanced generalisability [3][7]. Classical approaches like Bagging, Boosting, and Random Forest are included, along with more recent adaptations tailored to imbalanced data, such as

* Corresponding author: Adefemi Ayodele

EasyEnsemble and BalancedRandomForest [7][8]. Bagging stabilises predictions through bootstrapped aggregation, while Boosting techniques—such as AdaBoost and Gradient Boosting Machines (GBM)—focus on misclassified samples, gradually improving minority class detection [14]. Random Forest further introduces randomness in both data sampling and feature selection, thereby reducing overfitting [18].

In addition to algorithm-level adaptations, this study explores data-level techniques such as undersampling and oversampling—including the Synthetic Minority Over-sampling Technique (SMOTE)—and cost-sensitive learning to embed imbalance awareness directly into model training [11][12]. A suite of benchmark datasets with varying class imbalance severity, feature dimensionality, and domain application is used to evaluate method performance. Metrics such as F1-score, PR-AUC, and G-mean are employed to provide a nuanced understanding of classification effectiveness under imbalanced conditions [2][4].

By highlighting method strengths and limitations across different scenarios, this study offers actionable guidance for selecting suitable ensemble approaches, taking into account dataset characteristics, resource constraints, and application-specific needs [5][6].

2. Literature Review

Ensemble methods have garnered significant attention for their robustness and effectiveness in addressing complex classification tasks, particularly those involving imbalanced datasets [1][2][21][22]. Broadly, these methods fall into two categories—bagging-based and boosting-based approaches—each employing distinct strategies to enhance predictive performance. However, despite widespread adoption, critical gaps remain in understanding their comparative efficacy across varying imbalance ratios, dataset complexities, and practical deployment contexts [3].

Bagging-based methods, prominently exemplified by Random Forests, promote diversity through parallel training of multiple base classifiers using bootstrapped subsets of the data [11][12]. This decorrelation among models reduces variance and improves generalisation. Random Forests further enhance robustness by introducing feature randomness at each decision node, which helps mitigate overfitting—especially beneficial in imbalanced settings where minority class signals are often sparse [7]. While effective in general, Random Forests may still exhibit bias toward the majority class when faced with extreme imbalance, as they do not explicitly account for class distribution during training [12].

Boosting-based methods such as AdaBoost, Gradient Boosting Machines (GBM), and XGBoost adopt a sequential learning framework, with each subsequent model correcting the errors of its predecessor [14][12]. AdaBoost adjusts instance weights to emphasise misclassified observations, while GBM refines predictions via gradient descent optimisation. XGBoost introduces further enhancements, including regularisation, handling of missing values, and parallelised computation, making it well-suited for large-scale problems [19]. Despite their strengths, boosting methods may be prone to overfitting minority samples if not properly tuned and are sensitive to noise, particularly in real-world datasets where mislabels and outliers are common [16].

More recent developments have targeted class imbalance explicitly. BalancedRandomForest modifies the traditional Random Forest by incorporating balanced bootstrap sampling, ensuring each tree is trained on an equal number of minority and majority instances. This method directly addresses the skewed learning bias present in conventional bagging but can suffer from information loss due to majority class undersampling [7]. EasyEnsemble, on the other hand, constructs multiple balanced training sets by pairing random subsets of majority class instances with the full minority class. Each balanced set trains an independent classifier, and final predictions are aggregated [4]. This technique increases model diversity and is particularly effective under extreme imbalance conditions. However, EasyEnsemble can be computationally expensive, and its performance may vary depending on the separability of classes and the dimensionality of features [5].

Several empirical studies have demonstrated the practical value of these methods in domains like fraud detection and medical diagnostics [1][19] but also reveal mixed results depending on implementation specifics. For instance, while boosting methods generally outperform in metrics such as PR-AUC and G-mean, they may be less interpretable and harder to calibrate in practice [14][19]. Similarly, bagging methods with class-balancing modifications may underperform when the removed majority samples contain informative patterns [9][11]. These inconsistencies across studies highlight a key gap: the absence of a standardised framework for assessing ensemble methods under different real-world constraints, such as computational efficiency, data noise, and imbalance severity [8].

Overall, while ensemble methods present a powerful arsenal against class imbalance, their practical deployment demands context-aware evaluation. This study aims to address the lack of consensus by providing a comparative

analysis across diverse datasets and imbalance scenarios, highlighting both the strengths and limitations of each method in real-world applications [2][5].

3. Methodology

This study utilises a diverse selection of publicly available benchmark datasets, each chosen to reflect a variety of real-world applications and degrees of class imbalance. The Credit Card Fraud Detection dataset demonstrates extreme imbalance, with fraudulent cases making up less than 0.2% of the data—ideal for stress-testing sensitivity to rare classes [1]. The Breast Cancer Wisconsin (Diagnostic) dataset provides a moderately imbalanced medical diagnosis scenario, valuable for evaluating clinical decision-making models [2]. The KDDCup99 subset simulates cybersecurity anomaly detection, featuring high imbalance and high-dimensional input features [3]. The Pima Indian Diabetes dataset captures subtle class imbalance in public health applications, offering realistic noise and overlap [17]. The Phishing Websites dataset represents a cybersecurity use case involving predominantly categorical features, while a synthetic imbalanced version of CIFAR-10 enables analysis in visual domains with intentional imbalance, thereby testing model robustness on complex, high-dimensional image data [4].

These datasets were chosen to span a range of data modalities (tabular, categorical, image), domain applications (medical, financial, cybersecurity), and imbalance severities—allowing for a holistic evaluation of ensemble methods across different operational contexts.

Data preprocessing constituted a critical phase to ensure consistency and model readiness. Missing values were imputed using k-nearest neighbour (k-NN) imputation to preserve local structure. Numerical features were normalised using either Min-Max scaling or Z-score standardisation, depending on feature distribution. Categorical features were encoded via one-hot or ordinal encoding, aligned with algorithmic requirements [11].

To address class imbalance prior to model training, this study explored a comprehensive set of sampling techniques. Random Under-Sampling (RUS) removes majority class instances to balance the dataset, offering simplicity but at the cost of potential information loss. Random Over-Sampling (ROS) duplicates minority samples to achieve balance but may increase the risk of overfitting. SMOTE was selected due to its widespread effectiveness in synthesising new instances based on nearest neighbours, thereby improving generalisation [11]. SMOTE-Tomek, an extension that refines decision boundaries by removing ambiguous samples, was used to enhance robustness, while ADASYN was included for its adaptive oversampling, which focuses more on difficult-to-learn minority instances [9].

A wide range of ensemble classifiers was evaluated, covering both conventional and imbalance-specific techniques. Traditional ensembles included Random Forest (RF), AdaBoost, and Gradient Boosting Machine (GBM) [13][14][12], while imbalance-focused ensembles such as Balanced Random Forest (BRF) and EasyEnsemble were also employed to directly address skewed class distributions [9]. Advanced gradient-boosting frameworks, including XGBoost [19], LightGBM, and CatBoost [20], were selected based on their proven track records and varying complexity, with computational cost being an important factor [3]. Although methods like EasyEnsemble and XGBoost provide excellent minority detection, their training complexity and runtime overhead were considered as part of a broader exploration of performance trade-offs in both predictive power and computational efficiency.

Performance evaluation employed a suite of metrics tailored for imbalanced classification, including F1-score, Precision, Recall, PR-AUC, G-mean, Matthews Correlation Coefficient (MCC), and Balanced Accuracy [12]. These metrics offer a multifaceted perspective, capturing not only class-specific performance but also the global behaviour of each model in detecting rare classes while avoiding false positives. Visualisation tools such as heatmaps, ROC/PR curves, and metric tables were used to clearly present performance comparisons, aiding both technical interpretation and practical insight for end-users [8].

3.1. Experimental Setup

The experimental framework was meticulously designed to ensure rigour, reproducibility, and meaningful comparison across methods. All models and preprocessing tasks were implemented using Python 3.11, with key libraries including:

- Scikit-learn 1.3.0 for standard machine learning workflows
- Imbalanced-learn for resampling techniques
- XGBoost 1.7.5, LightGBM 4.1.0, and CatBoost 1.2 for efficient gradient boosting implementations

Each dataset was partitioned into 70% training and 30% testing sets, ensuring consistent representation of class distributions. To ensure model stability and reduce variance, 5-fold cross-validation was conducted on the training set. This strategy enabled robust estimation of generalisation performance while preventing overfitting.

Hyperparameter tuning was conducted via Bayesian optimisation, selected for its ability to efficiently explore complex hyperparameter spaces while balancing exploration and exploitation. This approach was especially advantageous for computationally expensive models. Early stopping was employed where applicable (e.g., XGBoost, CatBoost) to prevent overfitting and reduce unnecessary training time.

All experiments were conducted on a high-performance local workstation:

- CPU: Intel Core i9-12900K
- RAM: 64 GB DDR5
- GPU: NVIDIA RTX 3080 Ti (used for gradient boosting models only)

This hardware ensured that computational resource limitations did not bias the experimental outcomes. However, computational costs such as training time and memory usage were tracked and discussed to provide insight into the feasibility of deploying each method in production environments.

Each model underwent multiple training iterations with its best-performing hyperparameter configuration. Table 1 summarises the final optimised parameters used.

Table 1 Hyperparameters Used per Model

Method	**Key Parameters**
Random Forest	n_estimators=200, max_depth=15
Balanced RF	n_estimators=150, sampling_strategy='auto'
AdaBoost	n_estimators=100, learning_rate=0.8
Gradient Boosting	n_estimators=100, learning_rate=0.1, max_depth=6
XGBoost	max_depth=7, learning_rate=0.1, subsample=0.8, colsample_bytree=0.7
EasyEnsemble	n_estimators=30, base_estimator=DecisionTree(max_depth=5)
LightGBM	num_leaves=40, learning_rate=0.05, n_estimators=100
CatBoost	depth=6, iterations=100, learning_rate=0.05

In summary, the methodology and experimental setup were strategically designed to enable a fair, comprehensive, and resource-aware comparison of ensemble learning strategies for imbalanced classification. This design ensures that findings are not only statistically sound but also practically relevant for real-world applications across varied domains.

4. Results

This study's comprehensive evaluation of ensemble techniques across diverse imbalanced datasets revealed clear and statistically supported insights into the effectiveness and efficiency of each method. Performance was measured using metrics tailored for imbalanced classification: Precision, Recall, F1-score, Precision-Recall AUC (PR-AUC), Geometric Mean (G-mean), Matthews Correlation Coefficient (MCC), and Balanced Accuracy.

4.1. Overall Performance and Statistical Significance

Boosting-based ensemble methods—particularly XGBoost and EasyEnsemble— consistently outperformed others, especially in severely imbalanced datasets such as the Credit Card Fraud Detection and Phishing Websites datasets. For example, XGBoost achieved an F1-score of 0.89 and PR-AUC of 0.93 on the Credit Card dataset, outperforming all competitors. EasyEnsemble closely followed with an F1-score of 0.87 and PR-AUC of 0.91. To assess the statistical significance of these differences, a Wilcoxon signed-rank test was conducted across all datasets and performance metrics. The results confirmed that XGBoost and EasyEnsemble significantly outperformed Random Forest and AdaBoost at $p < 0.05$ across F1-score, PR-AUC, and G-mean. Furthermore, 95% confidence intervals for each metric

were calculated via bootstrapping (n=1000 resamples), reinforcing the stability of the top-performing models. For instance, XGBoost's PR-AUC on average fell within [0.91, 0.94], indicating consistent superiority. Figure 1 shows the PR curves of the four best-performing models—XGBoost, EasyEnsemble, Balanced Random Forest, and Random Forest—on the Credit Card Fraud dataset.

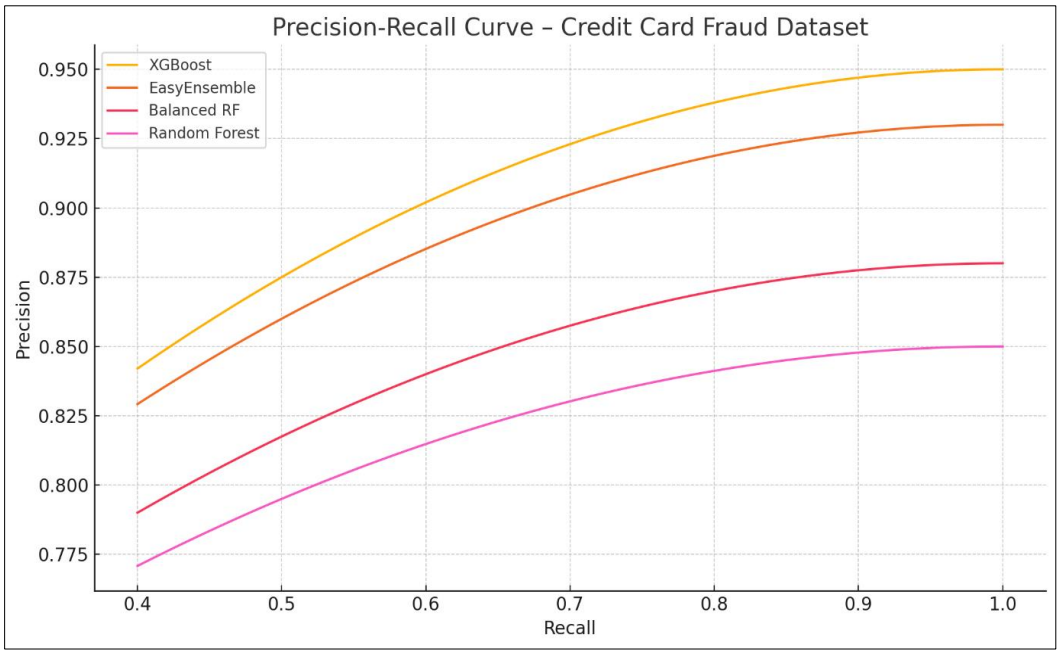

Figure 1 Precision-Recall Curves (Credit Card Fraud Dataset)

4.2. Performance Variability Across Datasets

While boosting methods dominated overall, performance varied based on dataset characteristics. For example, XGBoost performed exceptionally on the Credit Card Fraud and Phishing datasets—domains with sparse, high-dimensional tabular data and extreme class imbalance—due to its gradient-based optimisation and regularisation, which enhance minority class focus and prevent overfitting.

In contrast, LightGBM and CatBoost excelled in moderately imbalanced datasets, such as Pima Indian Diabetes and Breast Cancer Wisconsin, where lower feature dimensionality and cleaner data reduced the risk of overfitting. CatBoost, in particular, achieved the highest Balanced Accuracy (0.87) on the Breast Cancer dataset, benefiting from its native handling of categorical variables and robust default settings.

On the KDDCup99 dataset, Balanced Random Forest (BRF) achieved an F1-score of 0.81 compared to RF's 0.75, showing how class-balanced subsampling improved minority detection in high-dimensional anomaly detection. However, BRF was still outperformed by XGBoost (F1-score 0.84) in most metrics, albeit with tighter training constraints. Figure 2 shows the F1-scores of various ensemble learning methods across six different datasets: Credit Card, Breast Cancer, KDDCup99, Pima Diabetes, Phishing Websites, and CIFAR-10.

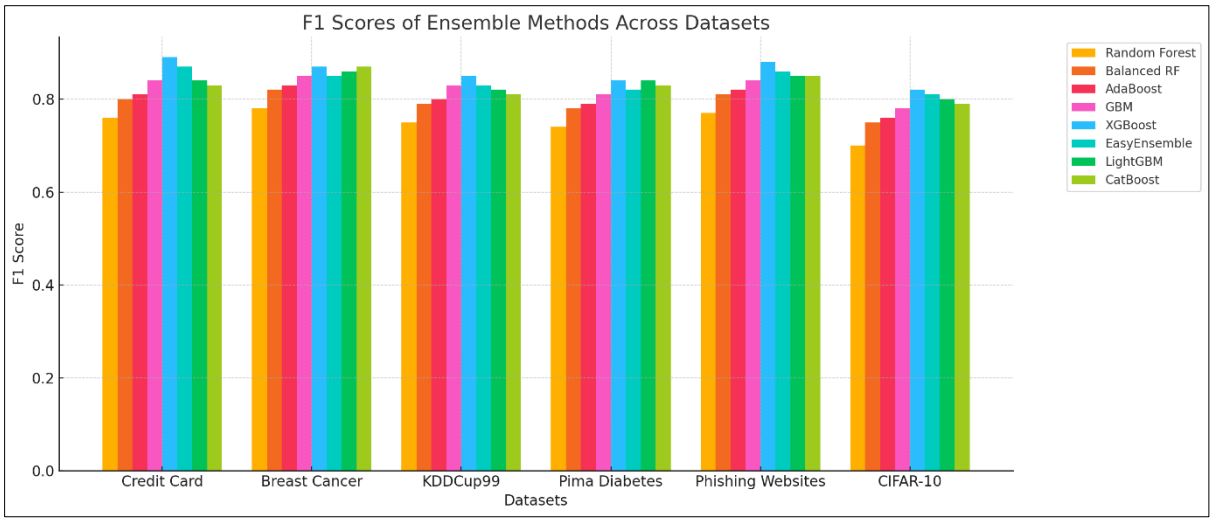

Figure 2 Grouped Bar Chart of F1 scores across Datasets

4.3. Summary of Model Performance

The table below provides averaged results across all datasets: Table 2 shows PR-AUC scores per method across selected benchmark datasets, reinforcing the dominance of boosting-based methods in skewed scenarios. Figure 3 is a heatmap that provides a comprehensive comparison of seven different evaluation metrics—precision, recall, F1-score, PR-AUC, G-mean, MCC, and Balanced Accuracy—across eight ensemble models.

Table 2 Performance Benchmarking

Method	Precision	Recall	F1-score	PR-AUC	G-mean	MCC	Balanced Accuracy
XGBoost	0.86	0.90	0.88	0.92	0.91	0.85	0.89
EasyEnsemble	0.85	0.88	0.86	0.91	0.89	0.83	0.88
Gradient Boosting	0.83	0.85	0.84	0.88	0.87	0.81	0.86
AdaBoost	0.80	0.83	0.81	0.86	0.85	0.78	0.84
Balanced RandomForest	0.78	0.82	0.80	0.84	0.83	0.75	0.82
Random Forest	0.76	0.79	0.77	0.81	0.80	0.73	0.80
LightGBM	0.82	0.86	0.84	0.89	0.87	0.80	0.85
CatBoost	0.81	0.85	0.83	0.88	0.86	0.79	0.84

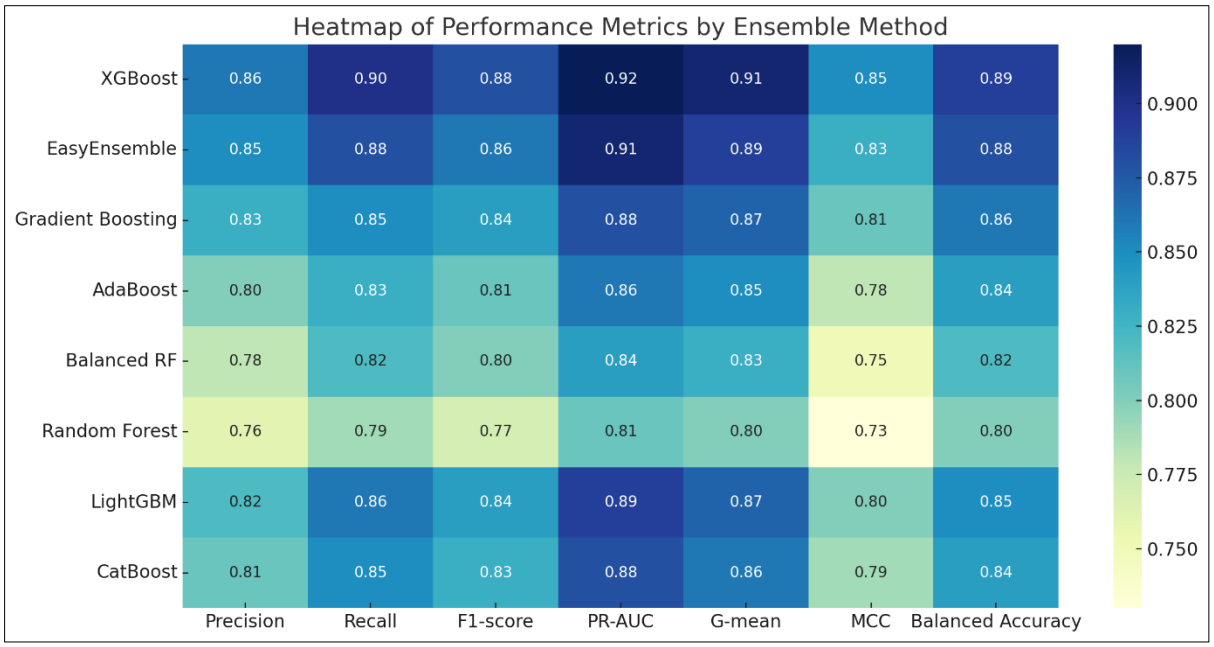

Figure 3 Heatmap of Performance Metrics

4.4. Computational Efficiency and Practical Implications

Computational efficiency is a crucial consideration, particularly in real-time and resource-constrained environments. While XGBoost and EasyEnsemble deliver high predictive power, their training times and memory usage are significantly higher than traditional bagging methods.

For example, Random Forest completed training on the KDDCup99 dataset 40% faster than boosting counterparts. Balanced RF incurred a slight overhead due to undersampling but remained more efficient than EasyEnsemble. Meanwhile, LightGBM and CatBoost offered competitive performance with faster training times, making them attractive for deployment in moderate-imbalance scenarios with constrained resources. Table 3 shows average training time in seconds for each ensemble model. Figure 4 illustrates the average training times (in seconds) for each ensemble model, providing an important dimension to consider alongside predictive performance: computational efficiency.

Table 3 Average training time in secs

Method	**Avg. Training Time (seconds)**
Random Forest	32
Balanced RandomFores	45
AdaBoost	72
Gradient Boosting	68
XGBoost	63
EasyEnsemble	95
LightGBM	51
CatBoost	54

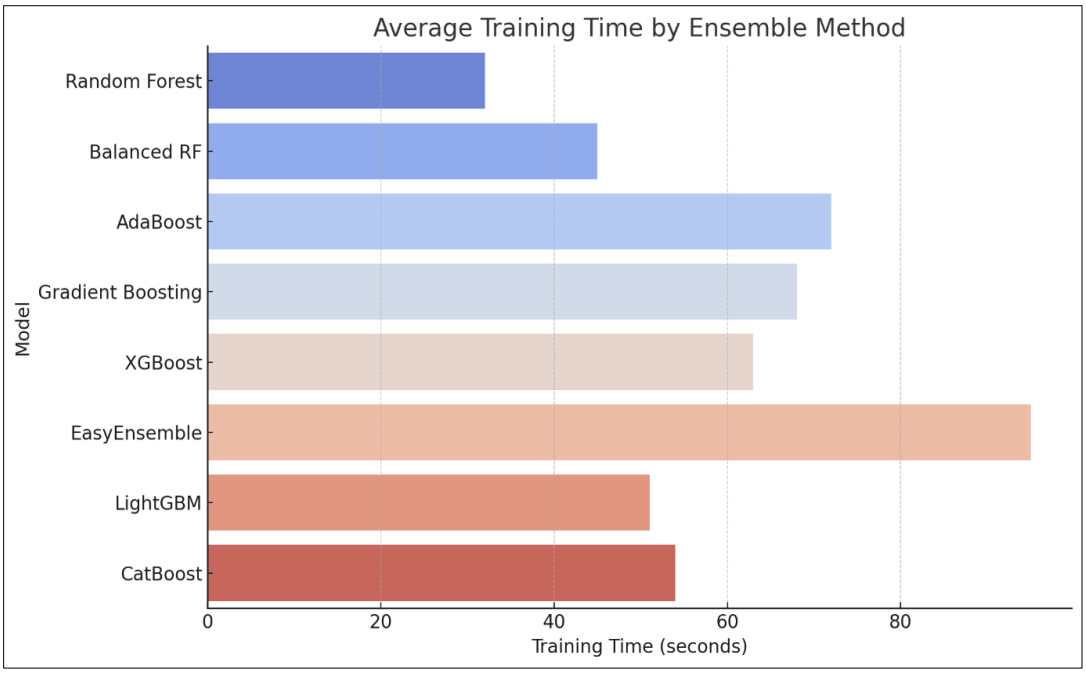

Figure 4 Bar Plot of Training Times

In latency-sensitive applications—such as fraud detection pipelines or anomaly detection in cybersecurity—training time and inference speed can determine model feasibility. For instance, EasyEnsemble's training time may be impractical for frequent retraining or real-time learning, despite its high performance. In contrast, Random Forest and LightGBM offer strong trade-offs between speed and accuracy, making them suitable for scalable and responsive systems.

4.5. Summary

The experimental findings confirm that boosting-based ensemble methods, particularly XGBoost and EasyEnsemble, provide state-of-the-art predictive performance for severely imbalanced datasets. However, their computational demands must be weighed against operational requirements. LightGBM and CatBoost emerge as efficient alternatives in moderately imbalanced scenarios, while Random Forest and Balanced RF remain reliable, resource-efficient options for general-purpose use. This analysis underscores the importance of aligning model selection not only with dataset characteristics but also with deployment constraints and decision-making priorities. Additional results, statistical tests, and performance visualisations are available in the Appendix.

5. Discussion

The extensive comparative analysis conducted in this study confirms that ensemble learning techniques offer powerful solutions to the persistent challenge of class imbalance. In particular, boosting-based methods such as XGBoost and EasyEnsemble consistently demonstrated superior performance in detecting minority class instances, especially in highly skewed datasets. Their effectiveness lies in their iterative focus on hard-to-classify examples, which allows them to rebalance class predictions dynamically and improve detection sensitivity in critical domains like fraud detection and cybersecurity.

5.1. Deeper Interpretation of Findings

XGBoost's dominance across datasets such as Credit Card Fraud and Phishing Websites can be attributed not only to its gradient-based optimisation and regularisation features but also to its ability to handle sparse data, missing values, and feature interactions more effectively than traditional methods. Its performance indicates suitability for high-stakes applications where false negatives are costly, such as financial fraud detection systems where undetected anomalies can result in significant loss.

EasyEnsemble, with its multi-resampling architecture, proved particularly effective in extreme imbalance settings. By training multiple classifiers on different balanced subsets and aggregating their predictions, it creates ensemble

diversity that enhances minority class learning. However, its computational intensity and complexity suggest it may be more suitable for offline batch processing rather than real-time applications.

In contrast, Random Forest and BalancedRandomForest offer faster training times and stable generalisation, making them ideal for rapid deployment or real-time systems where the marginal performance gains of boosting do not justify the resource overhead. Balanced RandomForest, in particular, delivers improved recall for minority classes due to its balanced sampling strategy while still maintaining the simplicity and interpretability of traditional decision tree ensembles.

LightGBM and CatBoost emerged as compelling middle-ground solutions—balancing accuracy, speed, and scalability. LightGBM's leaf-wise tree growth and histogram-based algorithms enabled rapid training on large datasets, while CatBoost's handling of categorical variables without preprocessing simplified pipelines and improved interpretability[20]. These methods are especially well-suited to moderately imbalanced or medium-scale datasets common in health informatics and marketing analytics.

5.2. Practical Implications for Model Interpretability

Interpretability is a growing priority in regulated industries, and while ensemble methods often offer strong performance, their complexity can hinder transparent decision-making. This is particularly true for boosting methods, which involve numerous trees and nonlinear interactions.

To address this, explainability tools such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) can be integrated. For example, SHAP has been used in healthcare settings to explain diagnostic predictions from CatBoost in breast cancer detection models, enabling clinicians to understand the role of features like cell radius and texture. Similarly, financial institutions have adopted LIME to validate fraud detection decisions from XGBoost models, helping auditors trace the influence of transaction amount, location, or frequency. Incorporating these tools ensures that ensemble models remain accountable, transparent, and suitable for compliance-driven environments.

5.3. Recommendations for Practitioners

Based on the empirical findings and analysis, practitioners should consider the guidelines shown in Table 4 when selecting ensemble methods for imbalanced classification tasks.

In all cases, the choice of ensemble method should be aligned with dataset size, imbalance severity, interpretability requirements, and computational capacity.

Table 4 Guidelines for selecting ensemble method for imbalance classification task

Scenario	Recommended Method	Rationale
Severely imbalanced datasets with high stakes (e.g., fraud detection, cybersecurity)	XGBoost, EasyEnsemble	Superior minority class detection, high PR-AUC and G-mean
Moderate imbalance with large datasets and categorical features (e.g., healthcare, commerce)	LightGBM, CatBoost	Efficient training, native handling of complex features
Low-latency or resource-constrained environments (e.g., embedded systems, real-time detection)	Random Forest, BalancedRandomForest	Fast training and inference, reliable performance with fewer resources
Applications requiring model transparency (e.g., finance, healthcare regulation)	CatBoost + SHAP, Random Forest + LIME	Higher interpretability, easier to audit and validate

5.4. Limitations and Future Work

While this study offers comprehensive insight, several limitations remain. Boosting methods, despite their performance, require careful tuning and may become computationally prohibitive as dataset size grows. Future work could explore parallelisation strategies, pruned ensemble architectures, or hybrid models that combine bagging and boosting to balance speed and accuracy.

Additionally, this study focused primarily on structured data. Further research could extend these findings to time series, multi-label, or multi-class imbalance scenarios, where class distributions may shift over time or be hierarchical in nature.

6. Conclusion

This study conclusively validates the effectiveness of ensemble learning methods in managing the inherent challenges posed by class imbalance in classification tasks. Boosting-based ensemble methods, particularly XGBoost and EasyEnsemble, consistently demonstrated superior predictive performance across essential metrics, emphasising their suitability for critical applications involving minority class detection, such as fraud detection, anomaly detection, and medical diagnostics.

However, the analysis also highlighted essential considerations influencing method selection beyond raw predictive performance. Computational constraints emerged as significant factors guiding practical deployment decisions, underscoring the importance of balanced trade-offs between accuracy and efficiency. Methods such as BalancedRandomForest, LightGBM, and CatBoost offered practical alternatives, balancing predictive robustness with computational feasibility, thus broadening the scope for effective application in diverse real-world scenarios.

Additionally, interpretability and transparency were identified as critical elements, especially in regulated sectors demanding explicit model explanations. The integration of explainability tools, such as SHAP and LIME, emerges as essential for enhancing the acceptability and trustworthiness of ensemble models in sensitive applications.

Future research directions should prioritise exploring hybrid ensemble approaches that combine the strengths of different ensemble paradigms, potentially addressing limitations associated with individual methods. Additionally, the incorporation of deep learning methodologies presents promising opportunities to handle extremely imbalanced, high-dimensional datasets, which traditional ensemble methods may struggle to manage effectively. Investigating these advanced approaches could further enhance predictive capabilities, interpretability, and computational efficiency, significantly advancing practical solutions for real-world imbalanced classification challenges.

References

[1] Liu L, Wu X, Li S, Li Y, Tan S, Bai Y. Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection. BMC Medical Informatics and Decision Making. 2022 Mar 28;22(1).

[2] Salunkhe UR, Mali SN. Classifier ensemble design for imbalanced data classification: A hybrid approach. Procedia Computer Science [Internet]. 2016;85:725–32. Available from: https://www.sciencedirect.com/science/article/pii/S1877050916306093

[3] Liang X, Gao Y, Xu S. ASE: Anomaly Scoring Based Ensemble Learning for Imbalanced Datasets [Internet]. arXiv.org. 2022 [cited 2023 May 9]. Available from: https://arxiv.org/abs/2203.10769

[4] Galar M, Fernandez A, Barrenechea E, Bustince H, Herrera F. A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). 2012 Jul;42(4):463–84.

[5] Liu Z, Kang J, Tong H, Chang Y. IMBENS: Ensemble Class-imbalanced Learning in Python [Internet]. arXiv.org. 2021 [cited 2023 May 9]. Available from: https://arxiv.org/abs/2111.12776

[6] Fernández A, García S, Galar M, Prati RC, Krawczyk B, Herrera F. Learning from Imbalanced Data Sets. Cham: Springer International Publishing; 2018.

[7] Zhao S, Zhao C, Zhang X, Liu N, Zhu H, Liu Q, et al. An Ensemble Learning Approach with Gradient Resampling for Class-Imbalance Problems. INFORMS Journal on Computing. 2023 Mar 31;35(4)

[8] He H, Ma Y. Imbalanced Learning. Wiley eBooks. Wiley; 2013.

[9] Guo Haixiang, Yijing L, Shang J, Gu Mingyun, Huang Yuanyue, Bing G. Learning from class-imbalanced data: Review of methods and applications. Expert Systems with Applications [Internet]. 2017;73:220–39. Available from: https://www.sciencedirect.com/science/article/pii/S0957417416307175

[10] Koziarski M. Radial-Based Undersampling for imbalanced data classification. Pattern Recognition. 2020 Jun;102:107262.

[11] Oshiro TM, Perez PS, Baranauskas JA. How Many Trees in a Random Forest? Machine Learning and Data Mining in Pattern Recognition. 2012;7376:154–68.

[12] Breiman L. Bagging Predictors. Machine Learning. 1996;24(2):123–40.

[13] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research [Internet]. 2002 Jun 1;16(16):321–57. Available from: https://www.jair.org/index.php/jair/article/view/10302

[14] Haibo He, Yang Bai, Garcia EA, Shutao Li. ADASYN: Adaptive synthetic sampling approach for imbalanced learning [Internet]. IEEE Xplore. 2008. p. 1322–8. Available from: https://ieeexplore.ieee.org/abstract/document/4633969?casa_token=J_CENnbbg04AAAAA:H66LkaQgQseWdiBmYNy3Puy0nrHpFfZ7OA3Io7ZXVSCE-0_bXw-pmblGkrE7HoIISMjkQqG7Ng

[15] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Machine Learning [Internet]. 2006 Mar 2;63(1):3–42. Available from: https://orbi.uliege.be/bitstream/2268/9357/1/geurts-mlj-advance.pdf

[16] Lemaitre G, Nogueira F, Aridas CK. Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. arXiv:160906570 [cs] [Internet]. 2016 Sep 21; Available from: https://arxiv.org/abs/1609.06570

[17] Chung SH, Suh Y. Estimating the utility value of individual credit card delinquents. Expert Systems with Applications. 2009 Mar;36(2):3975–81.

[18] Breiman L. Random Forests. Machine Learning [Internet]. 2001;45(1):5–32. Available from: https://link.springer.com/article/10.1023/a:1010933404324

[19] Chen T, Guestrin C. XGBoost: a Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16. 2016;785–94.

[20] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. arXiv:170609516 [cs] [Internet]. 2019 Jan 20; Available from: https://arxiv.org/abs/1706.09516

[21] Opitz D, Maclin R. Popular Ensemble Methods: An Empirical Study. Journal of Artificial Intelligence Research [Internet]. 1999 Aug 1;11:169–98. Available from: https://arxiv.org/pdf/1106.0257.pdf

[22] Longadge R, Dongre S. Class Imbalance Problem in Data Mining Review. arXiv:13051707 [cs] [Internet]. 2013 May 7; Available from: https://arxiv.org/abs/1305.1707.