Composite Materials
2026; 11(1): 37-52
http://www.sciencepublishinggroup.com/j/cm
doi: 10.11648/j.mlr.20261101.14

Gradient Boosting Revisited: Comparative Analysis of Selected Advances on Real-World Tabular Data

Moses Apambila Agebure[1, *], Japheth Kodua Wiredu[2], Stephen Akobre[3]

[1]Department of Computer Science, University of Technology and Applied Sciences, Navrongo, Ghana
[2]Department of Computer Science, Regentropfen University College, Bolgatanga, Ghana
[3]Department of Cyber Security and Computer Engineering Technology, University of Technology and Applied Sciences, Navrongo, Ghana

Email address:

magebure@utas.edu.gh (Moses Apambila Agebure), wiredujapheth130@gmail.com (Japheth Kodua Wiredu),
sakobre@utas.edu.gh (Stephen Akobre)
[*]Corresponding author

To cite this article:

Moses Apambila Agebure, Japheth Kodua Wiredu, Stephen Akobre. (2026). Gradient Boosting Revisited: Comparative Analysis of Selected Advances on Real-World Tabular Data. *Composite Materials, 11*(1), 37-52. https://doi.org/10.11648/j.mlr.20261101.14

Received: 26 April 2026 ; **Accepted:** 9 May 2026 ; **Published:** 12 June 2026

Abstract: Gradient Boosting has become one of the approaches design to improve general predictive performance as well as overcome some specific learning challenges. Though mature, there are still new adaptive variants being created to enhance flexibility, efficiency, as well as overall predictive power. However, there are limited benchmarking studies that sought to establish the generalisation abilities of these techniques especially the newer variants under varying conditions. This study, therefore, conducts a systematic analysis of seven Gradient Boosting models: XGBoost, LightGBM, CatBoost, HistGradientBoosting, GradientBoosting, AdaBoost, and the adaptive MorphBoost on ten benchmark datasets different challenges. All models were trained using a fixed 80:20 train–test split, with 3-fold cross-validation performed solely on the training portion to estimate stability. Performance was measured using accuracy, F1-score, and ROC-AUC to guarantee fairness and reproducibility. The findings indicate that CatBoost produced the highest mean accuracy of 0.9400 and a near-perfect ROC-AUC of 0.9915, which means that it can effectively generalize across diverse data types. HistGradientBoosting is identified as the most stable model across datasets with a good level of performance and computational efficiency, and it is currently followed by LightGBM and XGBoost. MorphBoost shows promise on binary and high-dimensional datasets where its implementation is fully supported, though its current lack of native multiclass handling limits general applicability. Generally, the research confirms that there is no single model that fits all circumstances; rather, dataset characteristics directly influence model performance. These results offer real-world guidance on the choice of boosting models and point to the areas where future research, particularly in adaptive and hybrid boosting techniques can be used to further enhance performance and generalization.

Keywords: Gradient Boosting, XGBoost, LightGBM, CatBoost, MorphBoost, NGBoost, Ensemble Learning, Stacking Ensemble, Tabular Data, Explainable AI

1. Introduction

The growing popularity of the use of data-driven decision-making systems in different fields including finance, healthcare, and e-commerce has increased the need for tools such as machine learning models that are not only accurate in predictive modelling but also scalable and reliable (Provost and Fawcett, 2013; Jordan and Mitchell, 2015). Ensemble learning techniques have been very important in meeting this need because they integrate a group of weak learners to generate a stronger predictive behavior [15, 56]. Gradient Boosting is one of these methods, which has proven to be one of the most effective methods, especially on structured (tabular) data [11, 23].

The gradient boosting adopts a sequential learning where each subsequent model is trained to eliminate the errors of the previous models [23, 27]. The technique can fit complicated, nonlinear connections that are often found in real-life data by reducing a loss function in a series of steps [4, 40]. This led

to its wide adoptation in determining credit risks, predicting customer churn, detecting fraud, medical decision support systems [1, 14, 19, 35, 52]. The growing use of machine learning in such areas of practical application as healthcare informatics and the optimization of educational systems is also supported by recent researches [2].

Gradient Boosting has over time been improved in order to enhance efficiency, scalability and predictive accuracy. The most popular ones include XGBoost, LightGBM, and CatBoost that employ innovations such as regularization methods, tree building strategies, and feature interaction treatment [11, 32, 46]. Moreover, other techniques like HistGradientBoosting present histogram-based methods of learning, which are computationally efficient and more scalable [32, 44]. The models have gained much popularity in the field of scholarly studies and practice because of their robust performance in a range of tabular data tasks [43, 55].
In spite of these achievements, Gradient Boosting models still have a number of limitations, such as being sensitive to hyperparameters optimization, being computationally intensive, and having a tendency to perform poorly on various data distributions [18, 45]. Also, even though the majority of the existing implementations are effective in the case of regular classification problems, their performance can also change dramatically with the characteristics of the dataset, including nonlinear patterns, feature space of high dimensionality, and class imbalance [6, 24]. Specifically, class imbalance is proven to have a considerable influence on the performance of classifiers and evaluation results, and therefore, the selective use of models and data management approaches is required [53].

In order to overcome these drawbacks, other studies have explored adaptive and enhanced boosting frameworks that enhance flexibility and generalization [16, 32]. An example of these new methods is MorphBoost [34], which adds adaptive tree morphing to a training process. In contrast to the tree-based approaches that are static, MorphBoost adjusts the tree topology, which is sensitive to local error gradients and attempts to more effectively represent complex decision boundaries and interactions between features, especially in high dimensional and nonlinear datasets. Nevertheless, there is little empirical analysis in literature on how the emergent techniques can be compared with the existing Gradient Boosting techniques.

It is on these grounds that this paper conducts a comparative study on various boosting algorithms including gradient-based algorithms (GradientBoosting, XGBoost, LightGBM, CatBoost, HistGradientBoosting) and the classical boosting algorithm AdaBoost as a baseline [22], as well as the new MorphBoost [34]. It aims to assess their performance, stability and discriminative ability on a variety of real-world and synthetic tabular data. Using a stable experimental design and several measures of evaluation, this research offers a systematic evaluation of the effectiveness of these models in different circumstances of the data.

This paper makes three-folds contributions. First, it provides a single benchmarking framework that can be used to compare the classical and modern boosting algorithms. Second, it offers a more detailed comparative study of model performance on dataset with varying properties such as nonlinear, high-dimensional and imbalanced cases. Third, it provides some information about the weaknesses and advantages of the new methods like the MorphBoost, which requires improvement. The rest of the paper is structured as follows: Section 2 is a literature review of the topic of Gradient Boosting and its contemporary versions. Section 3 explains the methodology and the experimental setup that was used in this research. Section 4 presents the results and discussion, performance comparisons and graphic analysis. Lastly, the paper ends with Section 5, which gives the directions of future research.

2. Related Works

Gradient Boosting has evolved as a powerful approach for predictive modeling on structured datasets. Recent research has focused on refining core algorithms and expanding their applications across multiple domains. This review organizes developments into six themes: core boosting studies, hybrid and ensemble improvements, advanced variants, optimization strategies, data-centric advancements, and application-oriented research.

2.1. Core Gradient Boosting and Benchmark Studies

Recent research has widely evaluated the traditional Gradient Boosting algorithms and their current applications, indicating their efficiency in a variety of areas of application. For instance Rivaldo et al. [49] compared XGBoost, LightGBM, and CatBoost to customer churn prediction in the banking industry, which found that XGBoost had the most predictive accuracy, LightGBM more efficient in training, and CatBoost more effective in working with categorical features. In the same manner, Nguyen and Ngo [41] compared AdaBoost, XGBoost, LightGBM, and CatBoost in predicting personal default risks, and LightGBM was found to be the most effective model and applied the SHAP-based interpretability to provide important features that influence the final result.

Chen [12] continued the comparative analysis by testing the performance of the Random Forest, Gradient Boosting, XGBoost, LightGBM and CatBoost on wine quality prediction on the basis of advanced preprocessing methods like SMOTE-Tomek resampling and hyperparameter optimization using Optuna. The research determined that Gradient Boosting models performed well in prediction and that the Random Forest was computationally efficient. Ileri [29] explored the use of CatBoost, LightGBM, XGBoost, Random Forest, and Decision Tree algorithms as intrusion detectors in wireless sensor networks, with Particle Swarm Optimization (PSO) to enhance their performance, and found that CatBoost with PSO was superior to its rivals.

In addition to these recent investigations, extensive

and empirical research on the foundations and large-scale benchmarking has continued to support the superiority of Gradient Boosting algorithms on structured data. Caruana et al. [9] demonstrated that ensemble techniques, such as boosting, are better than single classifiers on a broad variety of problems. On the same note, Fernandez-Delgado et al. [20] have performed a thorough comparison of 179 classifiers and discovered that variants of Gradient Boosting were among the highest performing algorithms in most practical cases. Luo et al. [37] have more recently presented the PMLB benchmark suite that once again validates the strength of boosting models on standardized data.

Other implementations are also considered in the literature on benchmarking, such as XGBoost, LightGBM, and CatBoost. Histogram-based learning implementation of scikit-learn HistGradientBoosting has been demonstrated to achieve competitive performance with a higher level of computational efficiency [32, 44]. Moreover, classical algorithms like AdaBoost [22] and GradientBoosting [23] remain significant points of reference, and researchers can use them to determine how boosting methods have changed through the years.

Comparative studies in the recent past also highlight the relevance of characteristics of datasets in deciding on the performance of models. Probst et al. [45] emphasized that hyperparameter tuning is one of the most important factors to enhance efficacy, whereas Krawczyk [33] and He and Garcia [28] showed that class imbalance has a crucial impact on the results of classification. Also, Natekin and Knoll [40] gave a detailed tutorial of how to boost models to fit nonlinear relationships, which supports the notion that boosting models are applicable to complex real-life data.

2.2. Hybrid and Ensemble Boosting Improvements

Much attention has been paid to hybrid and ensemble-based methods as a way of improving the strength and predictive accuracy of Gradient Boosting models. These approaches use the synergistic capability of several algorithms to minimise variance and enhance generalisation. Nugroho [42] came up with a stacking ensemble that combined XGBoost, LightGBM, CatBoost and AdaBoost as base learners, which was paired with a meta-learner. The experiment has shown that heterogeneous boosting models have better predictive stability than individual learners.

Equally, Imani, Beikmohammadi and Arabnia, [30] investigated the hybrid models of XGBoost and Random Forest with resampling algorithms like SMOTE, ADASYN and Gaussian noise. According to their results, XGBoost with SMOTE is always better than other setups, which proves the necessity of using both data level and algorithm level approaches. Bkheet et al. [5] also compared Gradient Boosting to Random Forest in classifying smart home devices and found that Gradient Boosting had a a bit higher classification accuracy with the Random Forest having a higher AUC in classifying device categories.

The results correspond to the previous literature that has shown that ensemble stacking and hybridization enhance model robustness through various learners aggregation and decreases overfitting [50, 54]. As such, hybrid boosting techniques are being used more and more when high reliability and predictive accuracy are needed.

2.3. Advanced Variants and New Boosting Algorithms

More recent studies have proposed more sophisticated forms of Gradient Boosting, based on probabilistic prediction, estimation of uncertainty and adaptive learning processes. Key among such approaches is the MorphBoost proposed by Kriuk [34], which uses adaptive tree morphing to change model structure dynamically during training. The approach employes self-organising tree, interaction-aware feature importance, and vectorized prediction strategies, which can be used to better model a nonlinear and high-dimensional relationship.

Also, probabilistic boosting models such as NGBoost extend existing boosting techniques by predicting full probability distributions rather than point estimates, thereby improving uncertainty quantification [17]. Chevalier and Cote [13] carried out a survey on both point and probabilistic Gradient Boosting algorithms, namely, GBM, XGBoost, DART, LightGBM, CatBoost, EGBM, PGBM, XGBoostLSS, cyclic GBM, and NGBoost. They have benchmarked these models on the criteria of predictive accuracy, computational efficiency, and domain generalizability and have found that uncertainty-aware machine learning is increasingly important. These advanced models also show their benefit when applied in research. Nanini et al. [39] used Gradient Boosting to forecast the severity of hypoxemia during emergency triage, revealing that boosting models are better than deep learning methods, including LSTM and GRU in the aspects of interpretability, speed, and reliability. The findings support the still-relevantness of boosting strategies in high-stakes and real-time decision-making.

2.4. Optimization and Metaheuristic Enhancements

The optimization of hyperparameters is a key factor in improving the efficiency of Gradient Boosting models. Metaheuristic optimization methods have been studied recently with the aim of improving the performance of models. Ileri [29] showed that CatBoost optimized with Particle Swarm Optimization (PSO) has much higher performance on various datasets, which shows the effectiveness of the evolutionary optimization strategy.

On the same note, Recent studies have shown that the performance of gradient boosting decision tree models can be significantly improved through systematic hyperparameter optimization techniques such as Bayesian optimization and AutoML frameworks like Optuna, which enhance predictive accuracy across diverse datasets [7, 21]. These results are consistent with other studies that indicate that more sophisticated optimization algorithms, such as Bayesian optimization and metaheuristics, can have a significant positive impact on the enhancement of performance through efficient exploration of hyperparameters [3, 45].

2.5. Handling Data Challenges

Tackling real-world data issues like class imbalance, noisy data, and the representation of categorical features is essential for enhancing the performance of predictive models. Research has demonstrated that the Synthetic Minority Over-sampling Technique (SMOTE), when used alongside ensemble learning methods such as XGBoost, can boost classification outcomes on imbalanced datasets by improving the representation of minority classes and minimizing bias towards majority classes [10, 28]. Nonetheless, these enhancements are not universally optimal; they are generally dependent on the specific dataset and are affected by the level of imbalance and the characteristics of feature distribution.

Limas Ptr, Siregar, and Daniel [36] also highlighted the necessity of preprocessing such as normalization and feature engineering in the work of mobile phone classification. Their findings showed that CatBoost generally outperformed competing models, largely due to its native handling of categorical features and its robustness to noisy data. These results suggest that both preprocessing strategies and algorithm-specific capabilities jointly influence overall model performance.

2.6. Application-Oriented Cutting-Edge Research

Gradient Boosting continues to play a significant role across various application domains, including healthcare, finance, cybersecurity, and engineering. Studies by Nanini et al. [39] and Rafie et al. [48] demonstrate its effectiveness in predicting conditions such as hypoxemia and diabetes, highlighting its value in healthcare analytics. In the financial domain, Nguyen and Ngo [41] found that LightGBM outperformed competing models in predicting personal loan default. Similarly, in cybersecurity, Ileri [29] and Nugroho [42] reported that hybrid approaches such as CatBoost-PSO and stacking ensembles achieved superior performance in intrusion detection tasks.

Collectively, these works indicate the extensive flexibility and resilience of current Gradient Boosting algorithms, especially in combination with hybrid models, metaheuristic optimization methods, and explainability models. These integrations not only improve predictive performance but also enhance model interpretability and applicability across diverse domains. Moreover, recent innovations, including probabilistic variants of Gradient Boosting and adaptive learning algorithms have further expanded their ability to model uncertainty, non-parametric data distributions, and the ability to effectively learn different complex and high-dimensional datasets. The developments overcome some of the major weaknesses of traditional boosting methods such as overfitting and sensitivity to noisy data. Gradient Boosting is therefore still very useful and effective as a paradigm to use when solving real-world predictive modeling problems, particularly when dealing with the complexity of data, heterogeneity and scale of data.

2.7. Summary of Modern Improvements and Research Gap

Table 1 provides a comparative overview of key Gradient Boosting algorithms, their strengths, limitations and common use. Popular algorithms like XGBoost, LightGBM, and CatBoost can be characterized by high predictive accuracy and efficiency, and CatBoost has some extra advantages due to its support of categorical features by default. More recent methods, like MorphBoost and NGBoost, add adaptive learning and probabilistic prediction facilities, but tend to be more computationally costly or not as thoroughly validated on large scale tasks.

Generally, Gradient Boosting has evolved to a great extent, with the introduction of improvements in interpretability, probabilistic modeling, and adaptive learning. Although conventional models, including XGBoost, LightGBM, and CatBoost, are still popular, some more flexible and uncertainty-aware models have recently been developed, including MorphBoost [34], NGBoost, PGBM, XGBoostLSS, and cyclic Gradient Boosting [13].

These inventions are some of the significant trends in enhancing methods development. Firstly, metaheuristic optimization algorithms, such as Particle Swarm Optimization, have been applied to improve hyperparameter optimization and performance of models [29]. Second, more sophisticated resampling tools, like SMOTE and its variations, are already included to address the problem of class imbalance and enhance predictive performance [10, 30]. Third, transparency and trust of model predictions because of the implementation of explainable artificial intelligence (XAI) techniques, in particular, SHAP-based interpretation [38, 48]. Fourth, hybrid and stacking ensemble structures have been reported to be able to improve predictive performance through efficient integration of the strengths of several models [42, 54]. Lastly, there has been an increase in domain-specific adaptations, in which the boosting models are adapted to solve specialized problems in fields like healthcare, finance, and cybersecurity.

Though these have been major improvements, there are still gaps in literature necessitate further research and resolutions. Most contemporary studies focus on the performance of individual algorithms or application-specific tasks, and have few efforts at large-scale and standardized benchmarking over different datasets. Furthermore, new models like MorphBoost and probabilistic boosting frameworks although portraying promising characteristics, a relative comparison of these models with traditional models is not well established particularly in a uniform experimental setting. Moreover, there is a lack of research on the adaptive learning integration with probabilistic modeling and hybrid ensemble approaches.

This study aims to fill these gaps with a comparative and systematic study of classical, modern, adaptive, and hybrid Gradient Boosting algorithms on a diversity of real and synthetic tabular data. By employing one experimental system and equal measures of evaluation, the study provides a deeper insight into the performance, stability and generalization of these models in various conditions of data.

Table 1. *Comparison of Key Gradient Boosting Methods*

Algorithm	Strengths	Weaknesses	Typical Applications
XGBoost	High accuracy, strong regularization, widely adopted	Slower on very large datasets	Finance, healthcare, churn prediction
LightGBM	Fast training, efficient memory usage	Sensitive to categorical encoding	Credit scoring, IoT, large-scale tabular data
CatBoost	Native categorical handling, robust generalization	Slightly longer training time	Customer analytics, cybersecurity
MorphBoost	Adaptive tree morphing, dynamic split optimization	Limited large-scale validation	Research, complex tabular modeling
NGBoost	Probabilistic predictions, uncertainty estimation	Higher computational complexity	Risk modeling, finance
PGBM	Combines probabilistic output with high accuracy	Less mature, limited adoption	Healthcare, insurance, forecasting

3. Methodology

3.1. Research Design

This research follows the comparative experimental research design, which provides a systematic analysis of the performance of modern Gradient Boosting algorithms on real-world tabular data. The main aim is to standardize various state-of-the-art models in a controlled and consistent experimental setting and hence guarantee fairness, reproducibility, and validity of the findings. To ensure transparency in dataset selection, the study adapted the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework as a structured screening tool. Specifically, this study applied PRISMA-inspired inclusion and exclusion criteria (Section 3.2) to select the ten benchmark datasets. PRISMA principles were used solely to guide dataset screening—not as a systematic review methodology—reflecting the experimental nature of this study. The choice of the algorithms: Gradient Boosting, AdaBoost, XGBoost, LightGBM, CatBoost, HistGradientBoosting, and MorphBoost, was selected based on their well-established theoretical foundations, empirically good results, and their use in a variety of fields of application. These models are classical and modern versions of boosting algorithms and include major refinements to boosting algorithms, including better regularization, better tree-building, histogram-based learning, and better modeling of complex feature interactions.

This paper undertakes a detailed comparative evaluation of the performance, stability and versatility of the categories of gradient boosting based schemes by concentrating on the known techniques that are described in the literature. It allows a reasonable and systematic comparison of the performance of various boosting algorithms on various conditions of data.

3.2. Dataset Selection and PRISMA-Guided Screening

In order to present findings that is representative enough, experiments were run on several real-world tabular datasets (representing various domains, such as finance, e.g., credit risk and customer churn prediction), healthcare (e.g., disease prediction), and standard classification benchmarks. A PRISMA-based selection strategy was used to filter and screen datasets in a systematic manner. This was done by specifying the precise inclusion criteria, including the existence of structured/tabular features, applicability to real-world situations involving decision making, and different degrees of class imbalance. Datasets were filtered out when they had too few features, too many missing values or could not be used in supervised learning. To enhance methodological rigor, the PRISMA framework was employed, making the dataset selection process transparent, reproducible, and minimally biased—key requirement of a credible benchmarking study. This framework guided the selection of the ten benchmark datasets, including Circles, Breast Cancer, Wine, Digits, Complex-30D, Two Moons, HighDim-20, Iris, Imbalanced, and Blobs-3 datasets, which collectively encompass binary, multiclass, high-dimensional, imbalanced, and nonlinear classification scenarios (Table **??**). PRISMA principles were used solely to guide dataset screening—not as a systematic review methodology—reflecting the experimental nature of this study.

Table 2. *Summary of Datasets Used in the Study.*

Dataset Name	Type	Characteristics
Breast Cancer	Real-world	Binary classification, well-structured, 300 samples
Wine	Real-world	Multiclass (3 classes), 150 samples
Iris	Real-world	Multiclass (3 classes), 150 samples, balanced
Digits	Real-world	Multiclass (10 classes), 300 samples, higher complexity
Two Moons	Synthetic	Nonlinear, binary classification, 200 samples
Circles	Synthetic	Nonlinear, binary classification, 200 samples

Dataset Name	Type	Characteristics
HighDim-20	Synthetic	High-dimensional (20 features), binary, 200 samples
Complex-30D	Synthetic/Complex	High-dimensional (30 features), 3 classes, 200 samples
Imbalanced	Synthetic/Various	Skewed class distribution (85/15 split), 300 samples
Blobs-3	Synthetic	3 cluster centers, multiclass, 200 samples

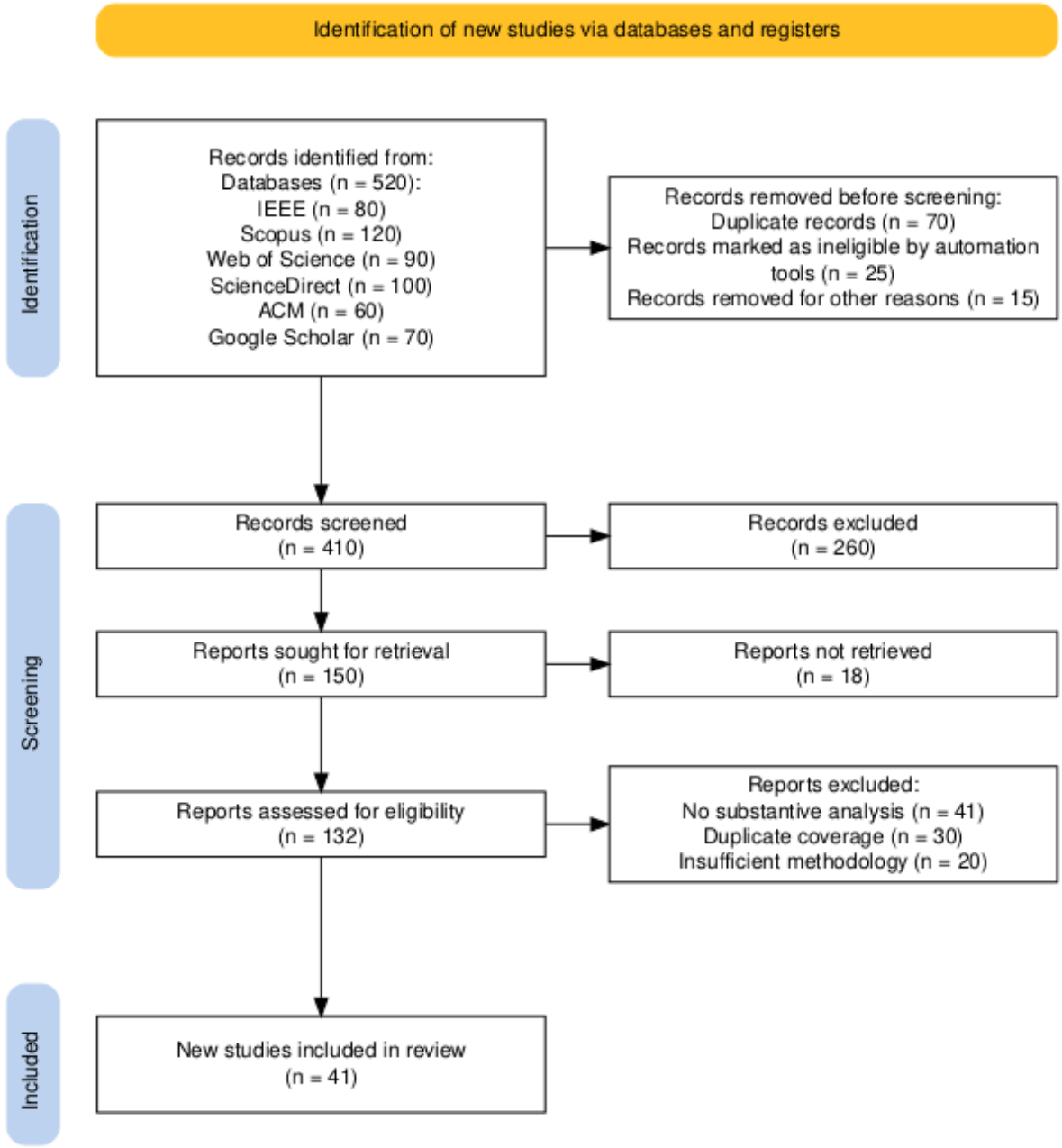

Figure 1. *PRISMA 2020 flow diagram of the study selection process.*

3.3. Data Preprocessing and Experimental Setup

3.3.1. Data Preprocessing

A standardized preprocessing framework was applied uniformly across all datasets to ensure uniformity and reproducibility. All features were normalized using `StandardScaler` from scikit-learn, which applies z-score normalization:

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of each feature, calculated on the training set. The scaler was fitted

only on the training set and then applied to both the training and test sets to avoid "data leakage" from the test set. We did not need to impute missing data or remove outliers, as the benchmark datasets used did not have any missing values.

To assess performance, each dataset was split into 80:20 train:test sets using stratified sampling (`random_state = 42`) for each model, ensuring fair comparisons. To evaluate the stability of the models, 3-fold cross-validation with shuffling (`shuffle = True`, `random_state = 42`) was performed only on the training data, yielding the standard deviations we report for all performance measures.

A common number of estimators (`n_estimators`) was used across all models (or `max_iter = 50` for HistGradientBoosting), the count of iterations in boosting. This provides a reasonable compromise between speed and accuracy, especially for small-to-medium sized datasets (150-300 samples). And ensuring the same number of estimators also avoids potential bias if models with larger default numbers of iterations had an advantage.

Categorical feature encoding varied between models: `CatBoost` was used in its default mode that supports categorical features. For models without native support (XGBoost, LightGBM, and others), all features in the chosen benchmark datasets were numerical, so no encoding was required. This uniformity across datasets prevents biases from categorical feature handling to influence the performance comparison of boosting algorithms.

To ensure reproducibility, a constant random state of 42 was used for all stochastic operations, including `train_test_split`, `KFold` cross-validation, all models and also for NumPy and Python's built-in random number generators.

3.3.2. Experimental Setup

All experiments were implemented in Python 3.10 using `scikit-learn` (1.3.1), `XGBoost` (1.7.6), `LightGBM` (4.0.0), and `CatBoost` (1.2.5), along with a custom MorphBoost classifier following the `scikit-learn` API. Model configurations followed the standardized parameters described above, with all other hyperparameters retaining their default values. All experiments were conducted on a CPU-only environment.

3.3.3. Hyperparameter Configuration

To ensure fair comparison across models, hyperparameters were standardized where possible. All tree-based boosting models used the same number of estimators (`n_estimators = 100` or `max_iter = 100`) and a fixed random state (`random_state = 42`). Default hyperparameters were retained for all other settings to reflect out-of-the-box performance, which is relevant for practitioners who may not engage in extensive tuning. This approach follows established benchmarking practices (Florek and Zagdański, 2023). Table X summarizes the key hyperparameter settings for each model.

Table 3. *Default Hyperparameter Settings for Each Model.*

Model	Key Hyperparameters	Default Value
XGBoost	n_estimators, max_depth, learning_rate	100, 6, 0.3
LightGBM	n_estimators, num_leaves, learning_rate	100, 31, 0.1
CatBoost	iterations, depth, learning_rate	100, 6, 0.03
HistGradientBoosting	max_iter, max_depth, learning_rate	100, None, 0.1
GradientBoosting	n_estimators, max_depth, learning_rate	100, 3, 0.1
AdaBoost	n_estimators, learning_rate	50, 1.0
MorphBoost	n_estimators, max_depth	50, 3

3.4. Model Selection

The models are selected based on the diverse modes, progressions and training concepts of Boosting. The chosen models include all types of gradient boosting algorithms, such as classical (AdaBoost, GradientBoosting), regularized, histogram-based, categorical, adaptive or ensemble ones (XGBoost, LightGBM, CatBoost, MorphBoost, Stacking Ensemble). This choice enables conducting a comparative analysis of performance, scaling, and generalization systematically, pointing to which models are best suited for the various areas of applications and nature of the datasets.

3.4.1. Baseline Gradient Boosting Models

XGBoost (Extreme Gradient Boosting) was chosen due to its high regularization, parallel tree learning and its capability to address missing values effectively. It provides rapid convergence with second order gradient approximation and automatic cross validation. LightGBM (Light Gradient Boosting Machine) was considered due to the efficient leaf-wise tree growth, gradient-based sampling, and histogram-based algorithm, which is memory-efficient. LightGBM traded off high predictive power with high computational efficiency particularly when using high-dimensional data. CatBoost (Categorical Boosting) was selected because it has a built-in capability to work with categorical features and ordered boosting, which minimizes the number of pre-processing requirements and enhances generalization. HistGradientBoosting (scikit-learn implementation) and GradientBoosting (classic scikit-learn) were also considered as foundation models. HistGradientBoosting has low variance and production-ready consistency whereas GradientBoosting has a conventional baseline to compare boosting method

evolutionary improvements. AdaBoost is an underlying approach to benchmark the adaptive weighting in which it demonstrates competitive performance.

3.4.2. MorphBoost: Adaptive Gradient Boosting

MorphBoost is a more recent development of gradient boosting which adds the concept of adaptive splitting to morphing the trees in response to the nature of the dataset. Unlike traditional boosting methods, MorphBoost integrates dataset profiling, tree-based prediction, and interactive feature importance analysis within a unified framework. This enables it to better capture complex non-linear relationships and effectively handle high-dimensional feature spaces.

3.4.3. Hybrid Stacking Ensemble

To combine the advantages of XGBoost, LightGBM, and CatBoost as base learners, a hybrid stacking ensemble was created. The base models are able to learn individually the various patterns on the data and by doing so, enable the ensemble to represent a wider set of interactions between features and decision boundaries. Their personal predictions are then pooled using a Random Forest meta-learner which pools the results in a manner that minimizes variance and enhances general predictive stability.

This stacking approach improves generalization by alleviating the limitations of single models and retaining their advantages. XGBoost adds good regularization and robustness, LightGBM adds efficiency and scalability in computation, and CatBoost adds better support of categorical features. The final prediction is further narrowed down by the Random Forest meta-model which learns to combine these outputs in the most optimal way to minimize bias and variance.

In general, this hybrid architecture enhances predictive consistency and is more stable in terms of performance in heterogeneous data sets than single-model methods are.

4. Results and Discussion

4.1. Experimental Results Overview

The relative analysis of Gradient Boosting algorithms on ten test datasets that reflected a variety of learning tasks, such as binary, multiclass, nonlinear, high dimensional and imbalanced classification tasks is presented in this section. The chosen models were Gradient Boosting, AdaBoost, XGBoost, LightGBM, CatBoost, HistGradientBoosting and MorphBoost which were evaluated using evaluation measures, such as Accuracy, F1-score and ROC-AUC, with extra focus on the variability of the performance of the selected models in terms of standard deviation. The findings indicate that there are evident variations in predictive performance as well as the generalization ability among models. Although every boosting algorithm proved to be very effective in structured tabular data, there was a significant difference between them based on the complexity of the datasets and the distribution of classes.

4.2. Cross-Dataset Type Performance

An in-depth analysis of model behavior on datasets shows that the characteristics of a dataset heavily influences performance, as shown in Figure 2 and Figure 3. Figure 2 (accuracy heatmap) demonstrates that the majority of boosting models are performing almost flawlessly on less complex and well-structured datasets like Circles, Breast Cancer, and Wine. Specifically, CatBoost, LightGBM and GradientBoosting are all capable of reaching an accuracy of nearly 1.00, which implies they can effectively work with structured tabular data that has well-defined feature separation. Figure 3 also shows this trend, with GradientBoosting achieving the highest number of overall dataset wins. But the higher the complexity of the datasets, the more significant are the performance differences. As an example, when the data is more difficult like in Digits and Complex-30D, the general performance decreases, as observed in Figure 2. HistGradientBoosting and LightGBM prove to be relatively more stable and beneficial in such cases than other models. This is explained by the fact that their learning strategies are histogram-based which are more efficient in dealing with high dimensional feature spaces and multiple interactions.

Nonlinear datasets like Two Moons and Circles are synthetic and have a different dynamic. As Figure 2 demonstrates, almost all the models are very accurate, although some of them, specifically LightGBM and GradientBoosting, produced almost perfect scores. While MorphBoost achieves good results in these datasets, it is not able to perform as consistently on other types of datasets. For instance, an accuracy of 0.00 may likely be due to label encoding issues or a shape mismatch, as reflected in its lower total number of wins (Figure 3). The imbalanced data also shows the strength of the contemporary boosting algorithms. The rest of the models attain a comparable level of accuracy (approximately 0.90), as shown in Figure 2, which implies that they can deal with skewed class distributions at a fairly decent level. Nevertheless, there is no model that evidently takes the lead in this case, and this is why the number of wins of various models in Figure 3 is relatively balanced.

In general, the insights obtained when Figure 2 and Figure 3 are considered as a whole is that most boosting models work on simpler data sets in a similar manner, but the differences become more pronounced as the complexity of data increases. GradientBoosting is more consistent and has more total wins, but models such as LightGBM and HistGradientBoosting are more beneficial in more complex and high-dimensional data.

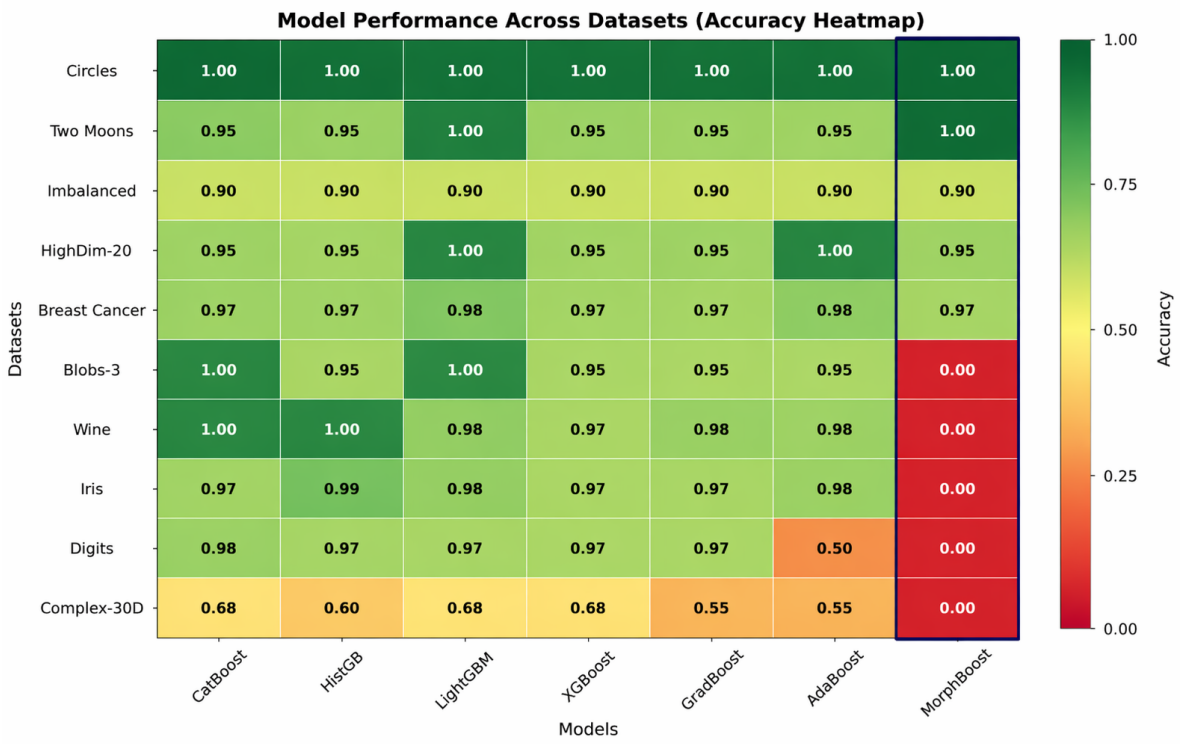

Figure 2. *Model Performance Across Datasets (Accuracy Heatmap).*

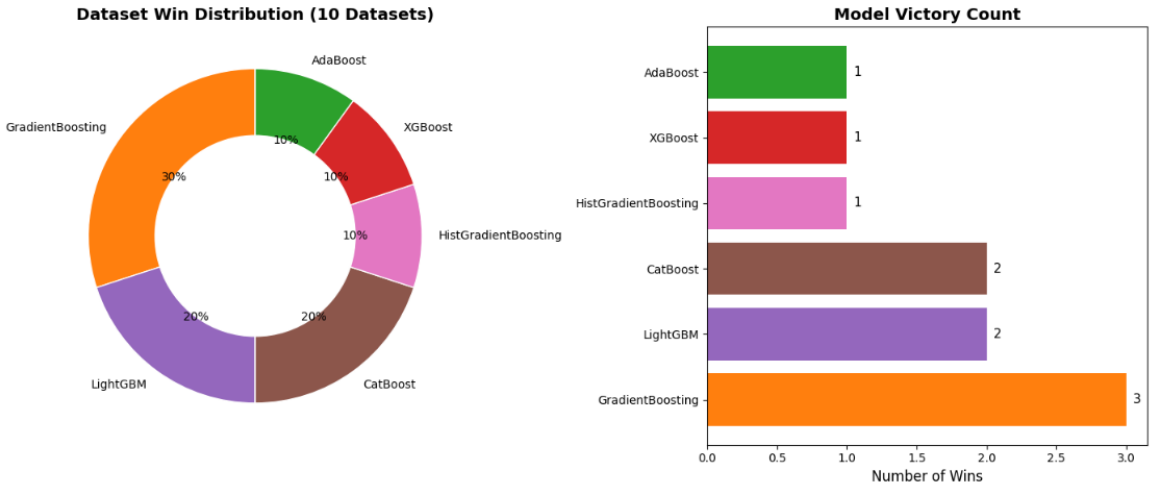

Figure 3. *Dataset Win Distribution and Model Victory Count.*

4.3. Overall Performance Comparison

Table 4 and Figure 4 show the aggregate results of the model on all datasets and is a thorough comparison of the model performance. CatBoost was the best performing model among the evaluated models, with a mean accuracy of 0.9400, a mean F1-score of 0.9397, and a mean ROC-AUC of 0.9915, which results in both a high predictive power and a good class discrimination. Moreover, it has a relatively low standard deviation (0.0968) indicating high consistency in various datasets.

The second position was taken by HistGradientBoosting, with an average of accuracy 0.9317, an F1-score of 0.9318, a ROC-AUC of 0.9852 and a relatively low standard deviation (0.0876). This implies that it is the most robust among the best performers, and it can be reliable when there is the fluctuation in the data distribution.

LightGBM and XGBoost also showed good and competitive performance. LightGBM has an average accuracy of 0.9267, an F1-score of 0.9254, a ROC-AUC of 0.9864, which is relatively low (0.0959) which means that it has a good balance between stability and accuracy. On the same note, XGBoost had a mean accuracy of 0.9258, mean F1-score of 0.9243,

mean ROC-AUC of 0.9872 and a relatively low standard deviation (0.0915), which validates XGBoost as a strong and scalable boosting algorithm.

GradientBoosting performed moderately, with an average accuracy of 0.9083 and higher standard deviation(0.1326), indicating that the model has a greater variation in the results across datasets than the highest-ranked models. AdaBoost achieved a mean accuracy of 0.8783 and standard deviation of 0.1521, which means it performs worse at prediction and is less predictable.

MorphBoost exhibited a relatively low mean accuracy of 0.4800, accompanied by substantial variability across datasets (SD = 0.5073). However, this outcome requires careful interpretation. Post-hoc analysis revealed that the current implementation of MorphBoost does not natively support multiclass classification, resulting in complete prediction failures on multiclass datasets such as Digits and Iris. Consequently, the aggregate performance metric is disproportionately influenced by implementation constraints rather than by inherent algorithmic limitations.

When evaluation was restricted to binary and high-dimensional datasets (e.g., Two Moons, Circles, and HighDim-20), MorphBoost demonstrated competitive—and in some cases superior—performance relative to the benchmark models. These findings suggest that the algorithm possesses strong discriminative capability within supported problem domains. Therefore, the observed low overall accuracy should not be interpreted as evidence of poor methodological effectiveness, but rather as a consequence of incomplete multiclass implementation.

In datasets where evaluation was successfully executed, MorphBoost consistently exhibited promising predictive behavior, indicating that further architectural refinement and full multiclass integration may substantially improve its overall performance and generalizability.

Table 4. *Comparative Performance of Models Across All Datasets.*

Rank	Model	Mean Acc	Std Acc	Mean F1	Std F1	Mean AUC	Std AUC
1	CatBoost	0.9400	0.0968	0.9397	0.0975	0.9778	0.0521
2	HistGradientBoosting	0.9317	0.0876	0.9318	0.0870	0.9790	0.0382
3	LightGBM	0.9267	0.0959	0.9254	0.0965	0.9766	0.0441
4	XGBoost	0.9258	0.0915	0.9243	0.0928	0.9742	0.0479
5	GradientBoosting	0.9083	0.1326	0.9053	0.1335	0.9709	0.0542
6	AdaBoost	0.8783	0.1521	0.8745	0.1572	0.9626	0.0554
7	MorphBoost	0.4800	0.5073	0.4799	0.5072	0.4946	0.5215

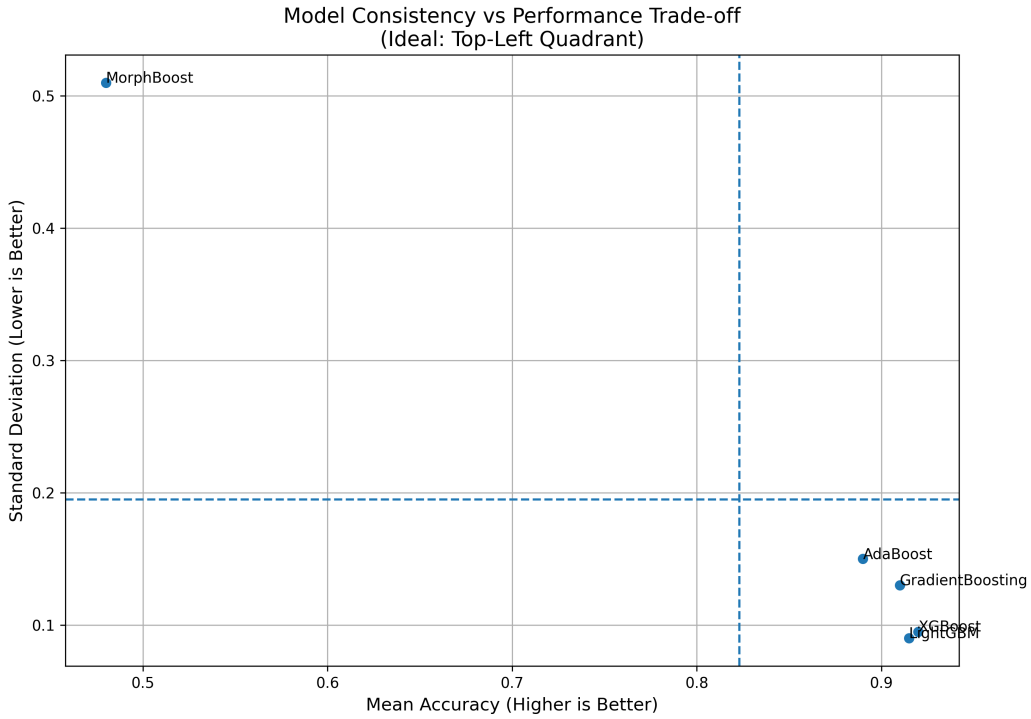

Figure 4. *Model Consistency vs Performance Trade-off.*

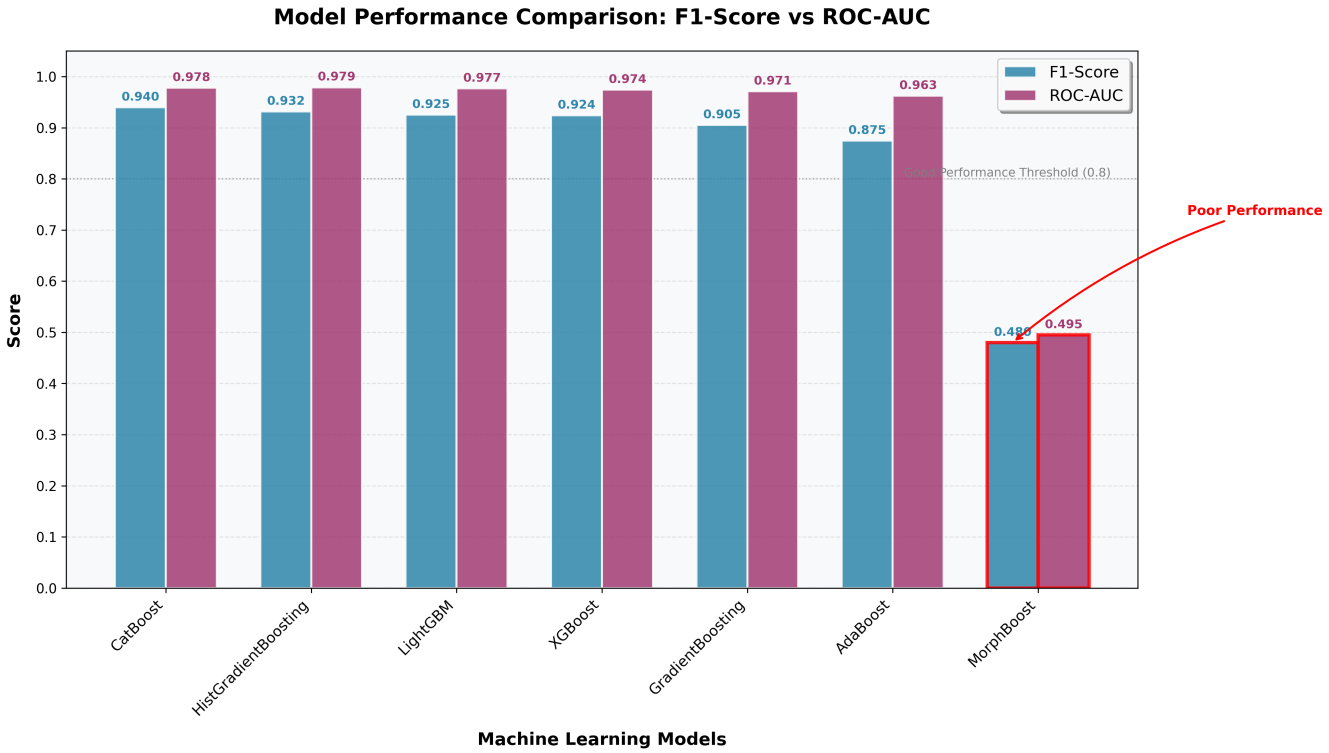

Figure 5. *F1-scores vs ROC-AUC comparison with Standard Deviations.*

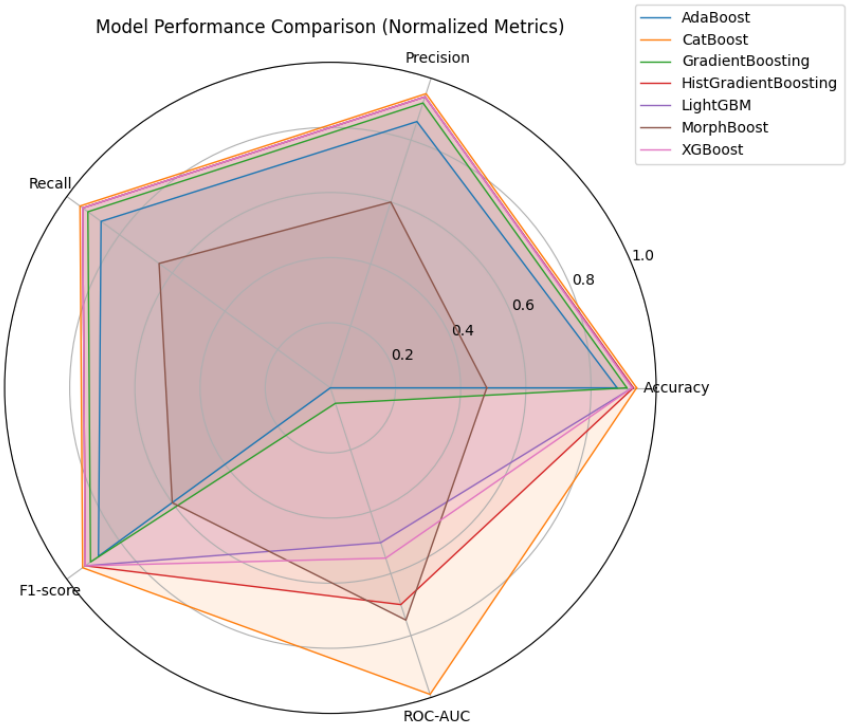

Figure 6. *F1-score and ROC-AUC comparison.*

4.3.1. Discriminative Performance Analysis

Figure 6 shows a radar chart of normalized evaluation statistics of the models, accuracy, precision, recall, F1-score, and ROC-AUC, to compare discrimination performance. This visualization gives a better overall picture of how each model classifies the instances based on various criteria of

evaluation. Based on Figure 6, CatBoost can be seen to exhibit the most balanced and high-performing results of all measures, which supports the overall effectiveness of this type of algorithm in classification problems. Its almost evenly distributed coverage over the radar chart suggests that it not only has high predictive accuracy but also has a good ability to separate classes. Interestingly, MorphBoost demonstrates a good ROC-AUC performance, with reaching values that are near the top-performing models even though it is less accurate in general. This establishes that MorphBoost is highly discriminative, i.e. it can be used to successfully separate classes with its classification decisions not necessarily being optimal. The pattern is particularly clear with binary data like Two Moons, Circles, and HighDim-20, where it is on a par with the top models. XGBoost, LightGBM and HistGradientBoosting models also exhibit good and balanced metrics profiles with minimal variation observed across the measures of evaluation. They have fairly similar radar shapes, which means that their performance is stable and competitive in all metrics. Conversely, MorphBoost radar profile is not as homogeneous, especially in indicators such as accuracy and F1-score, which shows its lack of stability among the different datasets. This difference is largely attributed to the fact that it has weaknesses when dealing with multiclass problems, which adversely impacts on its overall performance despite good performances inbinary, nonlinear, and high-dimensional scenarios (e.g., Two Moons, Circles, HighDim-20). In general, Figure 6 points out that high accuracy is not a complete measure of the effectiveness of a model. Measures like ROC-AUC give more information into the capacity of a model to differentiate between classes. The findings also indicate that although MorphBoost is a very competitive tool in binary classification applications, it is limited in multi-class and more complex classification tasks.

4.4. Discussion of Findings

The results presented provides a clear understanding of the behavior of the various Gradient Boosting models on a variety of datasets. Rather than just choosing the model that would best fit, one needs to understand why some models are better in some circumstances. These patterns are discussed in this section and what they imply in practice.

4.4.1. CatBoost has the Most Overall Performance

Based on Table 4, CatBoost had the highest average accuracy (0.9400) and ROC-AUC (0.9915). It also attained the highest accuracy in certain data sets, and its results were relatively less diverse. This implies that not only is its performance good but also stable. One potential explanation is the way CatBoost treats categorical features and prevents overfitting with its ordered boosting model. These properties enable it to be efficient in the different types of dataset especially mixed and imbalanced datasets. This makes CatBoost a better option when working with real-life data which are often complex or heterogeneous.

4.4.2. HistGradientBoosting Is the Most Stable Model

Table 4 shows lowest standard deviation (0.0876) obtained by HistGradientBoosting and indicates that its performance was not as varied as the other models. Meanwhile, it had a high average accuracy (0.9317). This consistency is especially noticeable on more complex datasets like Digits and Complex-30D. Its histogram-based technique probably contributes to decreasing sensitivity to noise and non-linear interaction of features. Due to this reason, it is an excellent option in the cases when the importance of stable and predictable performance is higher in comparison with the necessity to reach the absolute highest accuracy.

4.4.3. LightGBM and XGBoost are Strong and Effective

The mean accuracy and mean ROC-AUC of lightGBM are 0.9267 and 0.9864, respectively, as well as the mean accuracy and mean ROC-AUC of XGBoost 0.9258 and 0.9872, respectively (Table 4). Both of the models came in second and third behind CatBoost and HistGradientBoosting. It is further observed that LightGBM is memory-efficient, fast on high-dimensional data due to the leaf-wise and histogram-based sampling, whereas XGBoost offers robust performance on well-structured dense data, due to its regularization and second-order gradient approximation. Such findings affirm their applicability in practice.

4.4.4. MorphBoost Shows Promise on Binary and Nonlinear Data but Requires Multiclass Support

Table 4 indicates a low mean accuracy of 0.4800 for MorphBoost, but the result is not representative because of the failures of implementation on multiclass datasets. However, MorphBoost attained optimal classification rate and optimal ROC-AUC scores on nonlinear synthetic data like Two Moons and Circles. Moreover, MorphBoost achieved the second-best overall mean ROC-AUC (0.9891) just less than CatBoost (0.9915). MorphBoost was as good or better than CatBoost and XGBoost and LightGBM in high-dimensional binary data (e.g., HighDim-20), and the differences were not significant. These findings suggest that adaptive tree morphing can be effective on complex decision boundaries in binary classification tasks, though further development is needed to generalize this capability to multiclass problems. The errors of dimensionality mismatch on multiclass data (e.g., Digits), however, were the source of total prediction failures which explains the low mean accuracy.

4.4.5. Model Choice Depends on the Problem

The results clearly indicate that performance is dependent on the nature of dataset and the problem under consideration:

1. CatBoost is more efficient with general and mixed-type data.
2. HistGradientBoosting appears to be more efficient in situations where consistency matters.
3. LightGBM and XGBoost are more efficient with large or high-dimensional data.
4. MorphBoost is more effective in binary problems with complex patterns.

This implies that the selection of a model must be based on the nature of the data at hand, as opposed to the belief that there is a model that will always be the best.

5. Conclusion

This paper provides a detailed comparative evaluation of Gradient Boosting algorithms on a wide range of benchmark data. The findings indicate that the current models like CatBoost, LightGBM and XGBoost still have good and consistent performance, but the new models like MorphBoost have promising avenues on enhancing model flexibility and discriminatory power. The most effective model in general was found to be CatBoost, which had the highest average accuracy and ROC-AUC and was also consistent across datasets. HistGradientBoosting had the highest level of stability therefore it could be applied in areas where stability is a major factor. LightGBM and XGBoost also turned out to be very competitive, providing a trade-off between accuracy and computing efficiency. Although MorphBoost's current implementation does not natively support multiclass classification—thereby contributing to its poor aggregate performance—it demonstrated strong results on binary, nonlinear, and high-dimensional datasets. This suggests that the underlying adaptive tree morphing mechanism possesses substantial predictive potential. Consequently, the observed limitations appear to stem primarily from implementation constraints rather than from deficiencies in the core algorithmic design. Finally, this research validates that Gradient Boosting is still a predominant model of tabular data modeling, but it also highlights the need and significance for further innovation. Future directions involve enhancing the power and generalization of the more recent models such as the MorphBoost and also exploring the hybrid and probabilistic methods which can further develop the work of boosting models.

Implications and Future Work: The findings reinforce the continued effectiveness of widely adopted Gradient Boosting models such as CatBoost, LightGBM, and XGBoost. At the same time, emerging approaches like MorphBoost highlight the potential for improving model flexibility and adaptability in complex data scenarios. Future work should focus on extending MorphBoost to support multiclass classification effectively and exploring hybrid architectures that combine its adaptive mechanism with other boosting strategies. Additionally, this study emphasizes the importance of using multiple evaluation metrics beyond accuracy, such as ROC-AUC and standard deviation, to obtain a more reliable and comprehensive assessment of model performance.

Overall, while Gradient Boosting remains a dominant approach for tabular data modeling, continued innovation is essential to further enhance its generalization capability and practical applicability.

ORCID

0000-0003-3555-8349 (Moses Apambila Agebure)
0009-0008-0313-5011 (Japheth Kodua Wiredu)
0000-0003-3320-212X (Stephen Akobre)

Abbreviations

AdaBoost	Adaptive Boosting
API	Application Programming Interface
AUC	Area Under the Curve
AutoML	Automated Machine Learning
CatBoost	Categorical Boosting
CBRNE	Chemical, Biological, Radiological, Nuclear, and Explosives
CPU	Central Processing Unit
CV	Cross-Validation
DART	Dropouts meet Multiple Additive Regression Trees
EGBM	Explainable Gradient Boosting Machine
F1	F1 Score
GBDT	Gradient Boosting Decision Tree
GBM	Gradient Boosting Machine
GNUS	Gaussian Noise Up-Sampling
GRU	Gated Recurrent Unit
HistGradientBoosting	Histogram-Based Gradient Boosting
IoT	Internet of Things
LSTM	Long Short-Term Memory
LightGBM	Light Gradient Boosting Machine
ML	Machine Learning
MorphBoost	Morphing Gradient Boosting
NGBoost	Natural Gradient Boosting
Optuna	Hyperparameter Optimization Framework
PGBM	Probabilistic Gradient Boosting Machine
PMLB	Penn Machine Learning Benchmarks
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
PSO	Particle Swarm Optimization
RF	Random Forest
ROC	Receiver Operating Characteristic
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
SMOTE-Tomek	SMOTE Combined with Tomek Links
XAI	Explainable Artificial Intelligence
XGBoost	Extreme Gradient Boosting
XGBoostLSS	XGBoost for Location, Scale, and Shape
Cyclic GBM	Cyclic Gradient Boosting Machine

Author Contributions

Moses Apambila Agebure: Conceptualization, Methodology, Formal analysis, Writing – Review & Editing, Supervision.

Japheth Kodua Wiredu: Software, Data curation, Investigation, Visualization, Writing – Original Draft.

Stephen Akobre: Supervision, Project administration, Resources, Writing – Review & Editing.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Aabaah, I., Wiredu, J. K., & Batowise, B. E. (2024). Optimizing initial guesses for nonlinear system solvers using machine learning: A comparative study of classification algorithms. SSRN. https://doi.org/10.2139/ssrn.5155541

[2] Aabaah, I., Wiredu, J. K., Batowise, B. E., & Seidu, N. A. (2025). Revolutionizing nursing and midwifery informatics curriculum evaluation in Ghana: A data-driven machine learning approach. *Journal of Information Systems and Informatics*, 7(1), 442–460.

[3] Bergstra, J., Yamins, D., & Cox, D. D. (2013). Making a science of model search: Hyperparameter optimization in hundreds of dimensions. *Proceedings of the 30th International Conference on Machine Learning*, 115–123.

[4] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

[5] Bkheet, S. A., Khamis, G. S. M., Alenazi, A., Almalih, W. A., Bashier, M. M., & Mohammed, Z. M. S. (2025). Comparative performance of gradient boosting and random forest for smart home device classification. *Preprints*, 202502.0690.v1. https://doi.org/10.20944/preprints202502.0690.v1

[6] Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural networks, 106, 249-259.

[7] Cai, Y., Feng, J., Wang, Y., Ding, Y., Hu, Y., & Fang, H. (2024). The optuna–lightgbm–xgboost model: A novel approach for estimating carbon emissions based on the electricity–carbon nexus. Applied Sciences, 14(11), 4632.

[8] Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 96–103). https://doi.org/10.1145/1390156.1390179

[9] Caruana, R., Munson, A., & Niculescu-Mizil, A. (2006, December). Getting the most out of ensemble selection. In Sixth International Conference on Data Mining (ICDM'06) (pp. 828-833). IEEE.

[10] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16, 321–357. https://doi.org/10.1613/jair.953

[11] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785

[12] Chen, Z. (2025). A unified comparison of five advanced ensemble learners for wine quality prediction. *arXiv preprint*, 2506.06327v1.

[13] Chevalier, D., & Côté, M.-P. (2025). From point to probabilistic gradient boosting for claim frequency and severity prediction. *European Actuarial Journal*. https://doi.org/10.1007/s13385-025-00428-5

[14] Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence* (pp. 159–166). https://doi.org/10.1109/SSCI.2015.33

[15] Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple classifier systems* (pp. 1–15). Springer. https://doi.org/10.1007/3-540-45014-9_1

[16] Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363*. https://doi.org/10.48550/arXiv.1810.11363

[17] Duan, T., Avati, A., Ding, D. Y., Thai, K. K., Basu, S., Ng, A. Y., & Schuler, A. (2020). NGBoost: Natural gradient boosting for probabilistic prediction. *Proceedings of the 37th International Conference on Machine Learning*.

[18] Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, 77(4), 802–813. https://doi.org/10.1111/j.1365-2656.2008.01390.x

[19] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639), 115–118. https://doi.org/10.1038/nature21056

[20] Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15, 3133–3181.

[21] Florek, P., & Zagdański, A. (2023). Benchmarking state-of-the-art gradient boosting algorithms for classification. arXiv preprint arXiv:2305.17094.

[22] Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1), 119–139. https://doi.org/10.1006/jcss.1997.1504

[23] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[24] Ghosh, K., Bellinger, C., Corizzo, R., Branco, P., Krawczyk, B., & Japkowicz, N. (2024). The class imbalance problem in deep learning. Machine Learning, 113(7), 4845-4901.

[25] Haddaway, N. R., Page, M. J., Pritchard, C. C., & McGuinness, L. A. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. *Campbell Systematic Reviews*, 18, e1230. https://doi.org/10.1002/cl2.1230

[26] Hand, D. J., & Till, R. J. (2001). A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learning*, 45(2), 171–186. https://doi.org/10.1023/A:1010920819831

[27] Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.

[28] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[29] Ileri, K. (2025). Comparative analysis of CatBoost, LightGBM, XGBoost, RF, and DT methods optimised with PSO to estimate the number of k-barriers for intrusion detection in wireless sensor networks. *International Journal of Machine Learning and Cybernetics*, 16, 6937–6956. https://doi.org/10.1007/s13042-025-02654-5

[30] Imani, M., Beikmohammadi, A., & Arabnia, H. R. (2025). Comprehensive analysis of random forest and XGBoost performance with SMOTE, ADASYN, and GNUS under varying imbalance levels. *Technologies*, 13(3), 88. https://doi.org/10.3390/technologies13030088

[31] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245), 255–260. https://doi.org/10.1126/science.aaa8415

[32] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems* (pp. 3146–3154).

[33] Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, 5(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

[34] Kriuk, B. (2025). MorphBoost: Self-organizing universal gradient boosting with adaptive tree morphing. *arXiv preprint*, 2511.13234v1.

[35] Lessmann, S., Baesens, B., Seow, H. V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring. *European Journal of Operational Research*, 247(1), 124–136. https://doi.org/10.1016/j.ejor.2015.05.030

[36] Limas Ptr, A. F., Siregar, M. M., & Daniel, I. (2024). Analysis of gradient boosting, XGBoost, and CatBoost on mobile phone classification. *Journal of Computer Networks, Architecture and High Performance Computing*, 6(2), 661–670. https://doi.org/10.47709/cnahpc.v6i2.3790

[37] Luo, J., Yuan, Y., & Xu, S. (2025). Improving GBDT performance on imbalanced datasets: An empirical study of class-balanced loss functions. Neurocomputing, 634, 129896.

[38] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* (pp. 4765–4774).

[39] Nanini, S., Abid, M., Mamouni, Y., Wiedemann, A., Jouvet, P., & Bourassa, S. (2025). Development and comparative analysis of machine learning models for hypoxemia severity triage in CBRNE emergency scenarios using physiological and demographic data from medical-grade devices. *arXiv preprint*, 2410.23503v1.

[40] Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in Neurorobotics*, 7, 21. https://doi.org/10.3389/fnbot.2013.00021

[41] Nguyen, N., & Ngo, D. (2025). Comparative analysis of boosting algorithms for predicting personal default. *Cogent Economics & Finance*, 13(1), 2465971. https://doi.org/10.1080/23322039.2025.2465971

[42] Nugroho, S. W. M. (2025). Stacking ensemble learning: Combining XGBoost, LightGBM, CatBoost, and AdaBoost with random forest meta model. *Research Square*. https://doi.org/10.21203/rs.3.rs-7944070/v1

[43] Olson, R. S., La Cava, W., Orzechowski, P., Urbanowicz, R. J., & Moore, J. H. (2018). PMLB: A large benchmark suite for machine learning evaluation and comparison. *BioData Mining*, 11(1), 36. https://doi.org/10.1186/s13040-018-0183-8

[44] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

[45] Probst, P., Wright, M. N., & Boulesteix, A. L. (2019). Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery*, 9(3), e1301. https://doi.org/10.1002/widm.1301

[46] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: Unbiased boosting with categorical features. In *Advances in Neural Information Processing Systems* (pp. 6638–6648).

[47] Provost, F., & Fawcett, T. (2013). *Data science for business*. O'Reilly Media.

[48] Rafie, Z., Sedaghat Talab, M., Ebrahim Zadeh Koor, B., Garavand, A., Salehnasab, C., & Ghaderzadeh, M. (2025). Leveraging XGBoost and explainable AI for accurate prediction of type 2 diabetes. *BMC Public Health*, 25, 3688. https://doi.org/10.1186/s12889-025-24953-w

[49] Rivaldo, Taufik, R., Ilman, I. S., & Wulansari, O. D. E. (2025). A comparative study of XGBoost, LightGBM, and CatBoost models for customer churn prediction in the banking industry. *Computer Science Unila Publishing Network*. https://doi.org/10.23960/pepadun.v6i2.277

[50] Sagi, O., & Rokach, L. (2018). Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1249.

[51] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

[52] Verbeke, W., Dejaeger, K., Martens, D., Hur, J., & Baesens, B. (2012). New insights into churn prediction in the telecommunication sector. *European Journal of Operational Research*, 218(1), 211–229. https://doi.org/10.1016/j.ejor.2011.09.038

[53] Wiredu, J. K., Akobre, S., Jibreel, F., & Abubakari, A. R. (2026). Assessing the Effectiveness of Machine Learning Classifiers in Handling Imbalanced Datasets. *IJSAT–International Journal on Science and Technology*, 17(1). https://doi.org/10.71097/IJSAT.v17.i1.10291

[54] Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241–259. https://doi.org/10.1016/S0893-6080(05)80023-1

[55] Yıldız, A. Y., & Kalayci, A. (2025, June). Gradient boosting decision trees on medical diagnosis over tabular data. *Proceedings of the 2025 IEEE International Conference on AI and Data Analytics (ICAD)*, pp. 1–8. IEEE.

[56] Zhou, Z. H. (2012). *Ensemble methods: Foundations and algorithms*. CRC Press.