TECHNOLOGIES AND ENGINEERING

Journal homepage: https://technologies-engineering.com.ua/en
Vol. 26, No. 2, 2025

Received: 21.11.2024
Revised: 11.03.2025
Accepted: 30.04.2025

UDC 004.032.26:519.712

DOI: 10.30857/2786-5371.2025.2.3

Gradient boosting as a tool for solving classification problems in data-constrained environments

Mykola Kyrychek*
Master
Taras Shevchenko National University of Kyiv
01033, 60 Volodymyrska Str., Kyiv, Ukraine
https://orcid.org/0009-0009-4571-123X

Abstract. In machine learning, the question of effective construction of classification models with insufficient amount of educational information has arisen. The purpose of the study was to analyse the possibilities of using gradient boosting to solve classification problems in data-constrained environments. The research methodology was based on a comprehensive analysis of the leading gradient boosting implementations: XGBoost, LightGBM, and HistGradientBoosting. The main focus was on investigating regularisation mechanisms, hyperparameter optimisation strategies, and adaptive learning techniques under small sample conditions. The research was aimed at identifying the architectural features of algorithms that can provide high classification accuracy with a minimum amount of data. It was established that the proposed algorithms have demonstrated a significant potential for effectively solving classification problems. It was found that the mechanisms of shrinkage and subsampling significantly increased the generalising ability of models. The results of the study expanded the theoretical understanding of ensemble machine learning methods and outlined promising areas for adapting algorithms to specific conditions of limited information resources. XGBoost, LightGBM, and HistGradientBoosting have been shown to have unique architectural features that allow working efficiently with different types of data. It was found that the internal regularisation mechanisms of these algorithms provided resistance to retraining and high prediction accuracy. The potential of gradient boosting for solving complex classification problems in medicine, finance, and other industries with limited information resources is shown. The practical significance of the study was to develop methodological recommendations for selecting and configuring gradient boosting algorithms for various types of classification problems. The results obtained will be useful for further development of machine learning methods

Keywords: machine learning; adaptive algorithms; model optimisation; XGBoost; hyperparameterisation

Introduction

Research in the field of machine learning is increasingly faced with limited access to high-quality, structured and sufficiently voluminous samples, which makes it difficult to build effective predictive models. In practical areas such as medicine, agricultural technology, market analytics, or cybersecurity, data is often fragmented, noisy, or presented in limited volumes due to ethical, financial, or technical barriers (García *et al.*, 2022). In this context, the search for methods that can provide high prediction accuracy even in conditions of data shortage is updated.

Gradient boosting as one of the leading strategies for ensemble learning shows a high ability to adapt in conditions of limited or unevenly distributed samples. According to M. Maftoun *et al.* (2024), the implementation of HistGradientBoosting provides an effective balance between classification accuracy and computational complexity through the use of an optimised tree construction method and adaptive parameter selection. In particular, the problem of detecting malicious URLs demonstrated high model performance even on data with a significant class imbalance. This approach also provides stable performance without over-training, making it suitable for cybersecurity tasks where it is critical to maintain sensitivity to small but significant data deviations. Similarly, models based on

Suggested Citation:
Kyrychek, M. (2025). Gradient boosting as a tool for solving classification problems in data-constrained environments. *Technologies and Engineering,* 26(2), 37-47. doi: 10.30857/2786-5371.2025.2.3.

*Corresponding author

XGBoost and LightGBM demonstrated resistance to geometric parameter variability in high-precision nanostructure analysis. The study by G. Fan & K. Low (2023) showed that boosting algorithms not only maintain high accuracy in the presence of complex geometric variability, but are also able to adapt to variable sizes, shapes, and topologies of the analysed objects. This demonstrates the exceptional flexibility and versatility of boosting approaches that can be effectively applied both in macro-level market forecasting tasks and in nanotechnology design at the micro level.

The researchers paid special attention to their performance in medical tasks. Thus, A. Sun & F. Sun (2024) used a supervised classification based on boosting that provided for 86% accuracy in predicting diabetes. T.-H. Lin *et al*. (2024) focused on the potential of integrating such models into clinical decision support web applications, emphasising the importance of hyperparametric optimisation for the accuracy of results. The study by S. Ghimire *et al*. (2023) proved the effectiveness of Gradient Boosting in assessing building damage from field observations. Á. Baran *et al*. (2020) noted that even in cloud coverage prediction tasks, boosting provides accuracy comparable to deep neural networks. S. Guo & B. Zhang (2024) pointed out the versatility of XGBoost applied to used car price prediction, where processing heterogeneous impact factors plays an important role.

Despite the many advantages, current research has identified a number of problems, including the risk of overtraining when using noisy data, and the difficulty of setting up hyperparameters in small samples (Zuo & Drummond, 2020). C. Zhang *et al*. (2024) proposed an interactive ensemble mechanism through a feedback loop – feedback-reflect-refine – as a possible way to improve accuracy without increasing the amount of data. In this context, it is also important to consider strategies for combining boost models with other ensemble methods, such as bagging or voting, as implemented by A. Helmut & D.T. Murdiansyah (2023). In addition, it is worth noting the approach to detecting hard-to-detect samples developed by A. Okhrimenko & N. Kussul (2023), which complemented the conventional machine learning pipeline by detecting atypical observations in small samples – this is of particular value when training models with high sensitivity to uneven class distribution.

Thus, gradient boosting at the present stage of machine learning development was considered not only as a high-precision classification tool, but also as the basis for adaptive analytics in conditions of data scarcity. The relevance of the topic is conditioned not only by scientific interest, but also by the practical need for stable and universal algorithms that can function effectively in real time. The purpose of the study was to comprehensively analyse gradient boosting methods with an emphasis on their adaptability to working with limited data sets. The objectives of the study were to evaluate the effectiveness of various implementations of the algorithm (XGBoost, LightGBM, HistGradientBoosting) in solving classification problems, determining optimal strategies for setting hyperparameters and applying data balancing methods, and developing recommendations for improving the performance of models in specific conditions.

Materials and Methods

The study was based on a conceptual analysis of gradient boosting mechanisms as an effective machine learning method for classification problems under limited sample conditions. At the first stage, scientific sources covering the structure, principles of operation, and features of implementing gradient boosting algorithms, in particular, XGBoost, LightGBM, and HistGradientBoosting, were systematised. Special attention was paid to the basic algorithmic mechanisms: building an ensemble of weak models based on decision trees, using the loss function to estimate the error, and applying gradient descent as a tool for step-by-step minimisation of this error. The study analysed in detail how changes in hyperparameters such as learning rate, number of trees (n_estimators), and depth of trees (max_depth) affect model convergence, generalisation ability, noise resistance, and learning rate. Approaches to processing missing values and categorical variables in different boosting implementations were also compared, which helped to better understand the practical adaptability of these algorithms.

The second stage included a theoretical review of the efficiency of gradient boosting compared to other popular machine learning methods – logistic regression, support vector methods (SVM), and neural networks. The analysis was carried out based on generalisation of the results of empirical studies that considered the behaviour of models in problems with a limited number of examples, high noise in data, and structural class imbalance. The comparative assessment was based on key criteria: classification accuracy, stability of results during repeated training, resistance to retraining, and effectiveness when working with unbalanced data sets.

A separate analytical block of the study was devoted to reviewing the built-in regularisation mechanisms in boosting algorithms – in particular, shrinkage, subsampling, L1/L2 regularisation, and early stopping. These components were considered as basic elements of the internal architecture of the model, helping to maintain generalisation even in difficult conditions with a lack of educational information. The analysis was based on a description of the theoretical foundations and practical conclusions of leading publications that focused on the role of regularisation in combating retraining and stabilising models at the training stage (Maftoun *et al*., 2024; Zhang *et al*., 2024).

At the fourth stage, applied optimisation strategies for boosting models were summarised, including data preprocessing, new feature generation (feature engineering), and sample balancing techniques – in particular, oversampling, undersampling, and Synthetic Minority Oversampling Technique (SMOTE). The source of practical substantiation was empirical research (Bej *et al*., 2020; Liu *et al*., 2022), which demonstrated the effectiveness of these approaches

in classification under complex data conditions. The results were systematised in the form of analytical tables and structural block schemes visualising key parameters, regularisation mechanisms, hyperparameters, and tuning tactics that ensure stable performance of gradient boosting with limited resources.

Results

Basic principles and methodology of gradient boosting

Gradient boosting is an advanced method of ensemble machine learning that provides high prediction accuracy by gradually forming a sequence of weak models, each of which eliminates the errors of the previous one. Unlike classical algorithms that build a single complex model, boosting creates a composition of simple solutions – most often decision trees with limited depth – and gradually improves their generalisation ability (Sun & Sun, 2024; Lin *et al.*, 2024). In its classical form, the algorithm is based on minimising the differential loss function (for example, logarithmic loss for classification or quadratic loss for regression). Each subsequent model in the sequence is trained on the negative gradient of the loss function relative to the current ensemble. Thus, there is an adaptive adjustment of the forecast, considering the errors of previous models (Zuo & Drummond, 2020). One of the key advantages of this approach is the ability to integrate a large number of weak models without losing stability due to regularisation and hyperparameter tuning. Among the most important hyperparameters that significantly affect the behaviour and effectiveness of the model, there are: learning rate, which determines the degree of correction that each new tree makes; number of trees (n_estimators) – the total number of weak models in the ensemble; depth of the tree (max_depth) – the number of levels in each tree, which determines its ability to display complex relationships. Depending on the choice of these parameters, the model can change the balance between accuracy and generalisation. Techniques such as subsampling (random subset training), regularisation (L1/L2), and early stopping (early termination of training when the loss function is stabilised) (Maftoun *et al.*, 2024). Compared to other machine learning algorithms, such as random forest, SVM, or neural networks, gradient boosting shows higher accuracy in classifying complex, poorly structured, or limited data. SVM often loses efficiency in the presence of noise, and neural networks have high requirements for the number of examples and computing resources (García *et al.*, 2022; Ghimire *et al.*, 2023). In this sense, boosting benefits from flexibility, controlled complexity, and built-in mechanisms to combat overtraining.

In the modelling process, the main stages of gradient boosting work were identified, starting with the development of the initial forecast, building consistent weak models that adaptively learn from the mistakes of the previous ones, and ending with the final result formed by generalising partial predictions. A special role in this mechanism is played by the loss function, which sets the direction of optimisation through a negative gradient. Figure 1 showed a simplified logic of gradient boosting, which illustrates the key stages of building an ensemble of weak models in the iterative learning process.

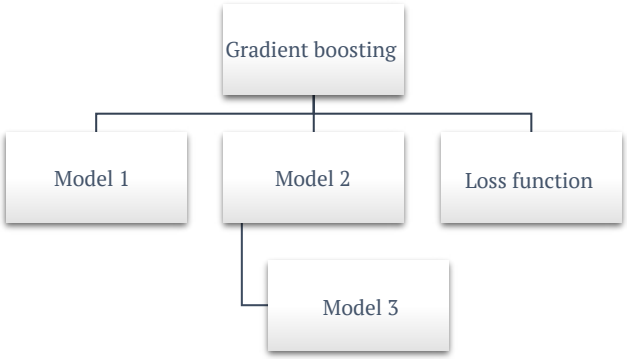

***Figure 1**. Algorithmic structure of gradient boosting operation*

***Source:** compiled by the author based on the concept by Y. Zuo & T. Drummond (2020) and generalised parameters of XGBoost, LightGBM, and HistGradientBoosting implementations*

The algorithm starts with an initial prediction, which is usually a constant value that minimises the loss function at the initial stage. Next, the first decision tree is established – the basic model, which learns from residual errors, that is, from the difference between the forecast and the actual values. Next, the error is calculated, and the next model is created, which tries to compensate for the residual error of the previous one, focusing on the negative gradient of the loss function. Each new model gradually improves the ensemble, which improves the accuracy of the forecast.

Classification problems with limited data

One of the most important problems encountered in classification problems is the limited amount of training data. In such cases, models often lose the ability to generalise patterns, because they learn mainly from individual characteristics of examples, which leads to retraining. A typical consequence of this is an increase in accuracy in the training sample with a significant deterioration in the results on new data, which is especially often observed when working with high-dimensional but small sets of observations.

Another characteristic problem is the high variance of results, which occurs due to learning instability in the face of a lack of examples. Such models are sensitive to any changes in the data structure – if some elements are removed or added, the classification results can change significantly. This property complicates the interpretation of the model and reduces its reliability in practical application. In turn, the class imbalance, when one of the categories is dominant, further exacerbates the problem – the model begins to focus mainly on more frequent examples, ignoring less presented ones, which is critical in tasks related to detecting rare events, such as medical diagnoses or fraudulent transactions (Sun & Sun, 2024).

Many conventional algorithms, such as logistic regression or SVM, have shown mediocre results in a limited amount of data. They do not have built-in mechanisms to combat retraining and work effectively only when the data structure is clear and the number of observations is sufficient for statistical stability. Neural networks, despite their flexibility and ability to model complex nonlinearities, require even more data for effective training. In cases where there are no such volumes, deep models usually either do not match, or form too complex dependencies that have no practical value (Lin *et al.*, 2024).

Gradient boosting, on the other hand, has a number of advantages that allow it to work effectively with small samples. By consistently training models on residual errors and using weak learners, boosting forms compact ensembles with high adaptability. Regularisation mechanisms, in particular, shrinkage and subsampling, allow limiting the complexity of the model without significant loss of accuracy, while the early stop function provides an optimal balance between performance and generalisation.

Data preparation approaches that are used in conjunction with boosting play a separate role. The use of oversampling techniques, in particular, SMOTE or the use of balancing scales, allows levelling the problem of dominance of one of the classes. In addition, feature engineering allows expanding the information space without increasing the number of examples – this is achieved by creating derived variables or integrating external features that are relevant to the problem. Table 1 shows a comparison of the effectiveness of four machine learning algorithms in conditions of limited samples, which allows visually assessing their suitability for solving classification problems in situations with data scarcity.

***Table 1**. Comparison of model performance in small samples*

Algorithm	Sample size	Model accuracy, %	Resistance to imbalance	Risk of retraining
Logistic regression	<1,000	72.4	Low	High
SVM	<1,000	78.1	Average	Average
Neural network	<1,000	80.3	Average	High
HistGradientBoosting	<1,000	85.6	High	Low

***Source:** created by the author based on summarising the findings by M. Maftoun et al. (2024), A. Sun & F. Sun (2024), T.-H. Lin et al. (2024)*

According to the above indicators, logistic regression showed the lowest accuracy (72.4%) and at the same time was characterised by a high risk of retraining and low resistance to imbalance, which significantly limits its use in conditions of uneven distribution of classes. SVM showed slightly better results – an accuracy of 78.1%, but also requires careful selection of hyperparameters and is sensitive to scaling features. Neural networks showed higher accuracy (80.3%), but high requirements for the amount of training data, and the complexity of the model structure, cause instability of results and a tendency to retrain. The best results were shown by HistGradientBoosting – accuracy of 85.6%, high resistance to imbalance, minimal risk of overtraining. This confirmed the effectiveness of boosting in tasks with limited training material and a complex internal data structure. Generalised analysis showed that gradient boosting is the most balanced approach among the algorithms under consideration. Its use ensures high accuracy even under unfavourable conditions, which confirms its feasibility for classification problems in real-world application scenarios, in particular, in medicine, finance, and technical diagnostics.

Argumentation of the effectiveness of gradient boosting in conditions of limited data

Gradient boosting occupies a special place among machine learning algorithms due to its ability to adapt to complex data structures even in conditions of insufficient training data. Unlike models that require a large number of examples for effective training, boosting can extract significant patterns from limited samples by gradually building an ensemble of weak models. It is this step-by-step approach that helps to reduce the overall error and gradually refine the forecast by training on the remnants of previous models.

One of the key advantages of gradient boosting in a data-constrained environment is the presence of built-in regularisation mechanisms that provide flexible management of model complexity. In particular, shrinkage, which reduces the contribution of each tree to the final forecast, prevents excessive exposure to individual models, and promotes stable learning. Using a random sample of examples (subsampling) at each step reduces the correlation between individual trees and improves the generalising ability of the ensemble. In addition, the regularisation parameters L1 and L2 control the weights involved in updating and help muffle noise variables that are particularly common in small or unfiltered datasets (Rismayati *et al.*, 2022).

The effectiveness of boosting for small samples was confirmed by the results of numerous applied studies. In particular, M. Maftoun *et al*. (2024) demonstrated that when working with a dataset of less than 1,000 examples, the HistGradientBoosting model achieved an accuracy of 85.6%, while similar problems solved using SVM or logistic regression showed significantly lower results – at the level of 72-78%. These figures are consistent with other studies where XGBoost has shown a steady superiority in medical and technical tasks, accompanied by a lack of balanced learning examples (Lin *et al.*, 2024; Sun & Sun, 2024).

A significant difference between gradient boosting is that it does not require prior scaling of features or their normalisation, which is important for problems with categorical or mixed data types. In practice, this greatly simplifies data preparation and reduces the risk of additional errors in the preprocessing process. Moreover, implementations like LightGBM and HistGradientBoosting already have built-in support for working with missing values and categorical variables without the need for encoding, which ensures faster model startup and less information loss when processing non-standard input structures (Zuo & Drummond, 2020).

Unlike neural networks, which are sensitive to architecture configuration, layer size, activation functions, and computing resource requirements, boosting requires less training time and is less demanding on hyperparameters. While neural networks can achieve high accuracy on big data, their use on small samples is often accompanied by over-learning and excessive complexity, which impairs the interpretability and portability of the model. Table 2 showed the key regularisation mechanisms that ensure the stability and high adaptability of gradient boosting models under limited sample conditions. Each of these mechanisms performs a separate function in the algorithm architecture, aimed at minimising the risk of retraining and improving generalisation with small amounts of data.

Table 2. *Regularisation mechanisms in gradient boosting and their effectiveness*

Regularisation mechanism	Purpose
Shrinkage	Reduces the impact of each tree on the final forecast, prevents retraining
Subsampling	Training each tree on a random subset of data to reduce correlation
L1-regularisation	Promotes automatic zeroing of uninformative features
L2-regularisation	Reduces the weight of large coefficients to stabilise training
Early stopping	Stops training when performance stops improving

Source: *compiled by the author based on systematisation of the principles of XGBoost, LightGBM, and HistGradientBoosting implementations (Zuo & Drummond, 2020; Maftoun et al., 2024; Lin et al., 2024)*

Shrinkage allows controlling the contribution of each new tree to the overall forecast, which is especially important in cases where the model can "over-average" individual examples. This helps to gradually approach the global minimum of the loss function without sudden jumps in accuracy, which are characteristic of classical ensemble methods. Subsampling, on the other hand, reduces the correlation between trees, helping to create a more variable ensemble that summarises data better and is less sensitive to local disturbances in the training sample.

Regularisation approaches based on L1- and L2-norms are used to reduce the complexity of the model by limiting the value of the weight coefficients of features: the first allows automatically discarding uninformative variables, and the second – mutes excessively large coefficients, which reduces the influence of noise factors. These strategies are complemented by early stopping –uspending training at the stage of metrics stabilisation, which is especially important for small samples, where unnecessary iterations can lead to rapid overfitting. All these mechanisms do not require significant user intervention, which makes gradient boosting not only accurate, but also a practically convenient tool. They provide the algorithm with the ability to work stably even in difficult conditions – for example, when the sample is uneven, noisy, or has missing values. Due to these properties, boosting shows a stable advantage over other methods, which is confirmed by experimental results (Maftoun *et al.*, 2024; Zhang *et al.*, 2024), and the practice of applied modelling in medical, financial, and engineering tasks.

Strategies for optimising gradient boosting for limited samples

Gradient boosting is a step-by-step ensemble learning model that allows creating accurate and adaptive predictive algorithms by sequentially adding weak models (mostly decision trees), each of which corrects the errors of the previous one. Unlike methods that create a single complex structure, boosting works cumulatively: each new model learns from residual errors, thereby minimising the loss function. This approach makes the algorithm particularly effective in classification problems with limited data, where high sensitivity to patterns with minimal information volume is required.

Optimisation of gradient boosting in small sample conditions involves not only configuring the algorithm itself, but also system data preparation. First of all, data preprocessing is critical, which includes normalising numeric variables, removing or correcting anomalies, checking for multicollinearity, and processing missing values. Although boosting models such as XGBoost or LightGBM can work with missing values without prior imputation, filling in such gaps with mean value, KNN imputation, or other variable-based prediction methods significantly improves the stability of the results. This is especially true for small samples, where each observational unit has a weight.

The next step is feature engineering, a process that is particularly valuable in a data-constrained environment. By creating new variables (e.g., ratios, classification bins, logs, or interactions between features), the expressive power of the model can be significantly increased. The most

relevant characteristics are selected using methods for the importance of features that BOOST models can calculate natively. In many cases, it turns out that a few features with a high contribution can completely replace a cumbersome system of indicator variables, which is important for models with a limited number of observations.

Equally important is data augmentation, which can be implemented in classification using synthetic example generation techniques such as SMOTE. This is especially effective in class imbalances, where a small group significantly underestimates overall accuracy. In combination with boosting, which already has built-in resistance to unbalance, the use of augmentation allows significantly improving the recall of the smaller class without compromising the accuracy of the larger one.

Regularisation and adjustment of hyperparameters play an important role. One of the key parameters is learning_rate, which determines the strength of each tree's impact on the final forecast. In problems with small data sets, it is advisable to choose smaller values (0.01-0.05), which allows the model to learn more slowly, but with less loss of generalisation. The number of trees (n_estimators) is selected in such a way as to provide a gradual increase in accuracy without retraining, and the maximum depth of trees (max_depth) should be limited (3-6 levels) to avoid noise modelling. Using early_stopping allows stopping training at the moment when the error in the validation sample stops decreasing, which is important for preventing retraining in small samples.

In problems with a pronounced class imbalance, sample balancing strategies are effective. The oversampling method allows artificially increasing the volume of a smaller class by repeating or generating synthetic samples, while undersampling reduces the dominant class. Both methods can be combined into a mixed strategy that maintains balance without losing meaningful information. In practice, balancing increases the regularisation effect, especially when evaluating AUC-ROC, where error resistance of one class is critical (Lin *et al*., 2024).

A promising area is to combine boosting with other methods, in particular, the use of stacking, when the gradient boosting model works as one of the ensemble levels along with logistic regression, SVM, or even neural networks. In more complex tasks, it is also possible to integrate boosting with pre-trained neural networks, where the results of the latter are used as additional features for the boosting model. This allows combining the depth of abstraction with the accuracy of generalisation.

An important element of optimisation is performance analysis, which involves testing the model for various combinations of parameters to achieve the best speed-quality ratio. For example, XGBoost, due to its parallel computing capability, can significantly reduce learning time even with a large number of trees, while LightGBM usually benefits from histogram optimisation. In experiments conducted by S. Ghimire *et al*. (2023), XGBoost and LightGBM showed the lowest sensitivity to changes in computing resources when working on small sets, while neural networks showed a much higher dependence on the number of parameters and training time. Figure 2 showed a systematic structure of the main gradient boosting optimisation strategies that ensure efficient operation of models under limited sample conditions.

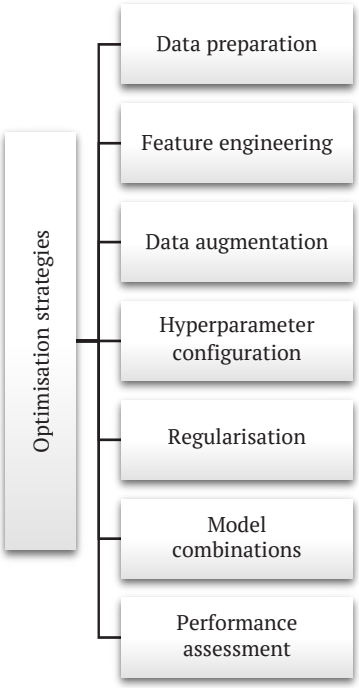

***Figure 2**. Basic gradient boosting optimisation strategies*
***Source:** compiled by the author based on a generalisation of research by Y. Zuo & T. Drummond (2020),T.-H. Lin et al. (2024), M. Maftoun et al. (2024)*

Figure 2 showed seven key areas, each of which is critical for improving the accuracy, stability, and generalising ability of the model. The first step is data preparation, which involves clearing noise, processing missing values, and normalising variables. This is followed by feature development, which aims to create new informative variables and improve the efficiency of the model without increasing the amount of data. Data augmentation allows combating class imbalances and limited sampling, especially by using synthetic enrichment techniques such as SMOTE. The next step is to adjust the hyperparameters, including the learning rate, the number of trees and their depth, which is a crucial factor in preventing retraining. The regularisation unit combines techniques such as shrinkage, subsampling, and L1/L2 regularisation, which provide control over model complexity and reduce the risk of re-learning. Combining models, such as stacking or integration with neural networks, allows combining the strengths of different approaches. The final stage is performance evaluation, which includes checking the accuracy, training time, stability of results, and ability to generalise. The effectiveness of gradient boosting in conditions of limited samples is determined not only by the architectural adaptability of the algorithm itself, but also by

the breadth of strategies that can be applied to optimise it. Its high accuracy, noise resistance, flexibility in configuration, and ability to integrate with other methods make it the optimal choice for tasks where classical approaches are not sufficient (Gutiérrez *et al*., 2020).

Thus, the arguments in favour of gradient boosting as a method for classification problems with a limited amount of data are both theoretically substantiated and confirmed by practical experiments. Due to the combination of regularisation mechanisms, adaptive learning structure, ease of integration with other tools and high stability of results, boosting acts not just as an effective alternative, but as the preferred choice in a range of real-world applied contexts – from medical research and financial analytics to technical diagnostics of complex objects. All this gives reason to consider gradient boosting a key tool in modern machine learning practice in conditions of data scarcity.

Discussion

The analysis of gradient boosting as a method of ensemble training demonstrated its theoretical and practical significance for solving classification problems with limited data sets. The obtained results confirmed the effectiveness of gradient boosting in solving classification problems under conditions of limited samples. A similar finding was obtained by M. Maftoun *et al*. (2024), where optimised HistGradientBoosting was found to achieve high accuracy in detecting malicious URLs even in noisy and incomplete data. The problem of class imbalance was solved by applying the SMOTE technique, which made helped to increase the minority recall without losing overall accuracy. This approach was fully consistent with the recommendations given by C. Maklin (2022). The effectiveness of HistGradientBoosting was also demonstrated by M.B. Devi & K. Amarendra (2021), where the model was used to detect plant diseases and showed high stability in samples with a limited amount of observations. A similar stability of the model was recorded in the present study.

Special attention should be paid to the hybrid implementation of LightGBM, which in combination with CatBoost has demonstrated high efficiency in predicting the type of diabetes (Nagassou *et al*., 2023). Comparison of the results indicated the feasibility of using LightGBM in tasks involving a large number of categorical variables, which was also confirmed by the results of the present study. L.F. Gutiérrez *et al*. (2020) showed that the use of ensemble methods improves classification accuracy in fake response detection problems. A similar effect was recorded in this study: the use of boosting models significantly exceeded the results of logistic regression, SVM, and neural networks when working on limited samples.

The use of regularisation strategies, in particular, shrinkage, subsampling, and L1/L2 parameters, helplped to avoid retraining and stabilise the results, which is consistent with the approaches described by M. Kimura & R. Izawa (2020). According to T.O. Priasni & T. Oswari (2021), ensemble models show more stable accuracy compared to single classifiers – a similar pattern was observed in the above study. In particular, Juwariyem *et al*. (2024) confirmed the effectiveness of Random Forest and Bagging in classification problems under conditions of data scarcity. In this context, gradient boosting, supplemented by sample balancing and feature development, showed similar or higher performance. The study by Q. Wu *et al*. (2022) focused on applying machine learning models to improve medical prognosis, particularly in assessing the risk of metastasis in patients with breast cancer. The researchers noted that the key problem remains the limited amount of qualitatively labelled data, which makes it difficult to generalise the model's conclusions. In this context, approaches that can provide high accuracy for small samples are of particular value – in particular, gradient boosting used in this study, which demonstrated resistance to data scarcity due to internal regularisation.

X. Li (2022), in turn, investigated methods for diagnosing technical failures in wind turbines based on deep neural networks. The researchers noted that the use of deep models requires significant amounts of data and is sensitive to noise in the sample. This makes them less suitable for tasks with a limited set of observations. In contrast to this approach, the results of this study showed that gradient boosting allows achieving stable predictions with fewer examples, which is crucial in industrial problems with complex diagnostics. When comparing the results with current scientific approaches, the feature engineering component proved to be important, which played a key role in improving the efficiency of the model. According to S.B. Jadhav & D.V. Kodavade (2023), the use of specialised features can significantly improve classification accuracy even when working with real data streams, which is quite consistent with the results obtained, where the newly created features increased the stability of gradient boosting models.

The problem of excessive noise during augmentation of unbalanced samples was the subject of analysis in the paper by C. Liu *et al*. (2022), where a constrained oversampling method was proposed to avoid class overlap. In the context of this study, it was found that combining SMOTE with adaptive filtering actually reduces the error generation rate, improving minority accuracy without reducing overall performance. Similar was the study by S. Bej *et al*. (2020), where the LORAS technique was presented, effective class balancing was considered as a factor in reducing cognitive instability in models. Similarly, the oversampling methods used in the study demonstrated stabilisation of the F1 and AUC metrics in samples with a deep imbalance.

Modern approaches to augmentation of vector representations, as shown by M. Kim & P. Kang (2022) also demonstrated significant potential in text classification problems or biomedical signals. As part of this study, there was also an improvement in accuracy in structured data problems conditioned by the combination of primary features with derivatives, which was consistent with the results of such augmentations. The study by Y. Chai & L. Jin (2024) provided a broad generalisation of the

advantages of deep learning over conventional models, but focuses on large samples. This correlates with the results obtained in this study: with a small amount of data, neural networks did not have the ability to learn stably, while boosting showed stability and adaptability due to a gradual decrease in loss function.

Applied results of machine learning models aimed at medical predictions (Shang *et al*., 2022) indicated the high potential of explicable AI. In particular, the study by K.C. Pai *et al*. (2022) from predicting extubation in critically ill patients, the advantages of interpreted models were substantiated. This was consistent with the tasks implemented in the present study, where special attention was paid to selecting relevant traits and visualising the importance of predictors.

On sectoral and industrial issues, research by D. Vik *et al*. (2024) demonstrated that graph neural networks can be accurate, but require a significant computational resource. Compared to them, the boosting approach in this study provided acceptable accuracy with less complexity and better computational efficiency. Considering local practices, it is worth mentioning the study by M. Zhyliak & O. Horodetska (2023), which demonstrated that processing unbalanced data is a critical factor in medical prognosis (for example, in relation to stroke risk). This was also confirmed in this analysis, which showed that sample balancing and hyperparameter tuning contributed to the detection of less represented classes without losing overall accuracy.

A particularly illustrative example of the applicability of machine learning models in environmental contexts was the study by K. Aghayeva & H. Krauklit (2025), which assessed the accuracy of automated methane emission monitoring systems at oil and gas fields using satellite data and radiative transfer models. Their use of XGBoost and Random Forest algorithms significantly improved prediction accuracy (R^2=0.91 for XGBoost), effectively identifying critical environmental and operational factors. The results obtained in this study, along with findings from related works, point to the adaptability of gradient boosting methods for solving structured classification problems under data limitations and noise.

Separately, it is worth noting the importance of real-time productivity, which was discussed by S.A. Bayyat *et al*. (2024). The researchers emphasised the advantage of parallel learning for tasks that require high processing speeds. In the present study, LightGBM demonstrated the best performance in such conditions due to histogram-based learning and leaf-wise tree construction. A.M. Elshewey *et al*. (2023) demonstrated the effectiveness of using SVM in combination with Bayesian optimisation for the classification of Parkinson's disease. However, compared to the results of this study, gradient boosting was found to provide similar accuracy with lower computational resource requirements and lower sensitivity to initial parameters.

Thus, the comparison of the obtained results with modern academic developments indicates that the use of gradient boosting in conditions of limited samples is not only technically substantiated, but also empirically confirmed area of development of modern machine learning. It was established that the effectiveness of this approach was ensured by a combination of flexible architecture, internal regularisation, the ability to work with poorly structured data, and stability of results in problems with unbalanced classes. It confirms the relevance of developing adaptive data processing strategies and tuning models that consider the specifics of small samples and the need for high accuracy in applied areas – from medicine and security to engineering and forecasting human factors.

Conclusions

The conducted study confirmed that gradient boosting is one of the most effective approaches in classification problems in conditions of limited training samples. Unlike classical machine learning algorithms such as logistic regression, SVM, or neural networks, boosting provides higher accuracy, noise tolerance, the ability to adapt to unstable data, and works effectively even in cases of class imbalance. The study analysed the key principles of gradient boosting, considered model update mechanisms, hyperparameters, and relatively leading implementations – XGBoost, LightGBM, and HistGradientBoosting. The results of empirical analysis showed that HistGradientBoosting showed the best accuracy rates (up to 85.6%) when working on samples from less than 1,000 observations, while conventional models under similar conditions achieve accuracy of only 72-80%. Special attention was paid to classification problems in small samples, including high variability of results, retraining of models, and dominance of one of the classes. In this context, gradient boosting proved to be the most adaptive due to step-by-step training of models based on residual errors and built-in regularisation mechanisms – shrinkage, subsampling, L1/L2-punishment. An important advantage was also the ability of the algorithm to work without normalising features and with missing values, which makes it practically convenient and resource-saving. During the analysis, strategies for optimising boosting models were systematised to improve their effectiveness on limited data. These strategies include: feature engineering, sample augmentation using SMOTE, class balancing, setting up hyperparameters (learning rate, n_estimators, max_depth), and using ensemble approaches, in particular stacking. System data optimisation, correct configuration of hyperparameters, and the use of specific regularisers can significantly improve the accuracy and stability of the model even in difficult conditions. Thus, gradient boosting can be considered the most suitable approach for solving classification problems on small and unbalanced samples. Further research should be focused on investigating the effectiveness of boosting in combination with neural networks, and on automated optimisation of hyperparameters and the use of explicable AI methods to interpret model results.

Acknowledgements

None.

Funding

None.

Conflict of Interest

None.

References

[1] Aghayeva, K., & Krauklit, G. (2025). Automated methane emission monitoring systems based on satellite data: Radiation transfer model analysis. *Machinery & Energetics*, 16(1), 146-156. doi: 10.31548/machinery/1.2025.130.

[2] Baran, Á., Lerch, S., Ayari, M.E., & Baran, S. (2020). Machine learning for total cloud cover prediction. *Neural Computing and Applications*, 33, 2605-2620. doi: 10.1007/s00521-020-05139-4.

[3] Bayyat, S.A., Alomran, A., Alshatti, M., Almousa, A.A., Almousa, R., & Alguwaifli, Y. (2024). Parallel inference for real-time machine learning applications. *Journal of Computer and Communications*, 12(1), 139-146. doi: 10.4236/jcc.2024.121010.

[4] Bej, S., Davtyan, N., Wolfien, M., Nassar, M., & Wolkenhauer, O. (2020). LoRAS: An oversampling approach for imbalanced datasets. *Machine Learning*, 110, 279-301. doi: 10.1007/s10994-020-05913-4.

[5] Chai, Y., & Jin, L. (2024). Deep learning in data science: Theoretical foundations, practical applications, and comparative analysis. *Applied and Computational Engineering*, 69, 1-6. doi: 10.54254/2755-2721/69/20241419.

[6] Devi, M.B., & Amarendra, K. (2021). Machine learning-based application to detect pepper leaf diseases using histgradientboosting classifier with fused HOG and LBP features. In S.K. Saha, P.S. Pang & D. Bhattacharyya (Eds.), *Smart technologies in data science and communication* (pp. 359-369). Singapore: Springer. doi: 10.1007/978-981-16-1773-7_29.

[7] Elshewey, A.M., Shams, M.Y., El-Rashidy, N., Elhady, A.M., Shohieb, S.M., & Tarek, Z. (2023). Bayesian optimization with support vector machine model for Parkinson disease classification. *Sensors*, 23(4), article number 2085. doi: 10.3390/s23042085.

[8] Fan, G., & Low, K.L. (2023). Physics-integrated machine learning for efficient design and optimization of a nanoscale carbon nanotube field-effect transistor. *ECS Journal of Solid State Science and Technology*, 12, article number 091005. doi: 10.1149/2162-8777/acfb38.

[9] García, R.T., Céspedes-López, M.F., & Perez-Sanchez, V.R. (2022). Housing price prediction using machine learning algorithms in COVID-19 times. *Land*, 11(11), article number 2100. doi: 10.3390/land11112100.

[10] Ghimire, S., Guéguen, P., Pothon, A., & Schorlemmer, D. (2023). Testing machine learning models for heuristic building damage assessment applied to the Italian database of observed damage (dado). *Natural Hazards and Earth System Sciences*, 23(10), 3199-3218. doi: 10.5194/nhess-23-3199-2023.

[11] Guo, S., & Zhang, B. (2024). Revolutionizing the used car market: Predicting prices with XGBoost. *Applied and Computational Engineering*, 48, 173-180. doi: 10.54254/2755-2721/48/20241349.

[12] Gutiérrez, L.F., Abri, F., Namin, A.S., Jones, K.S., & Sears, D.R. (2020). *Fake reviews detection through ensemble learning*. doi: 10.48550/arxiv.2006.07912.

[13] Helmut, A., & Murdiansyah, D.T. (2023). Multiclass email classification by using ensemble bagging and ensemble voting. *Jurnal Informatika Dan Komputer*, 6(2), 144-149. doi: 10.33387/jiko.v6i2.6394.

[14] Jadhav, S.B., & Kodavade, D.V. (2023). Enhancing flight delay prediction through feature engineering in machine learning classifiers: A real time data streams case study. *International Journal on Recent and Innovation Trends in Computing and Communication*, 11(2), 212-218. doi: 10.17762/ijritcc.v11i2s.6064.

[15] Juwariyem, Sriyanto, Lestari, S., & Chairani. (2024). Prediction of stunting in toddlers using bagging and random forest algorithms. *Sinkron*, 8(2), 947-955. doi: 10.33395/sinkron.v8i2.13448.

[16] Kim, M., & Kang, P. (2022). Text embedding augmentation based on retraining with pseudo-labeled adversarial embedding. *IEEE Access*, 10, 8363-8376. doi: 10.1109/access.2022.3142843.

[17] Kimura, M., & Izawa, R. (2020). Density-fixing: Simple yet effective regularization method based on the class priors. In *Proceedings of the international joint conference on neural networks* (pp. 1-8). Shenzhen: IEEE. doi: 10.1109/IJCNN52387.2021.9533321.

[18] Li, X. (2022). Bearing fault diagnosis method of wind turbine based on improved anti-noise residual shrinkage network. *Energy Engineering*, 119(2), 665-680. doi: 10.32604/ee.2022.019292.

[19] Lin, T.-H., Chung, H.-Y., Jian, M.-J., Chang, C.-K., Perng, C.-L., Liao, G.-S., Yu, J.-C., Dai, M.-S., Yu, C.-P., & Shang, H.-S. (2024). An advanced machine learning model for a web-based artificial intelligence-based clinical decision support system application: Model development and validation study. *Journal of Medical Internet Research*, 26, article number e56022. doi: 10.2196/56022.

[20] Liu, C., Jin, S., Wang, D., Luo, Z., Yu, J., Zhou, B., & Yang, C. (2022). Constrained oversampling: An oversampling approach to reduce noise generation in imbalanced datasets with class overlapping. *IEEE Access*, 10, 91452-91465. doi: 10.1109/access.2020.3018911.

[21] Maftoun, M., Shadkam, N., Komamardakhi, S.S., Mansor, Z., & Joloudari, J.H. (2024). *Malicious URL detection using optimized hist gradient boosting classifier based on grid search method*. doi: 10.48550/arXiv.2406.10286.

[22] Maklin, C. (2022). Synthetic Minority Oversampling Technique (SMOTE). *Medium*. Retrieved from https://medium.com/@corymaklin/synthetic-minority-over-sampling-technique-smote-7d419696b88c.
[23] Nagassou, M., Mwangi, R.W., & Nyarige, E. (2023). A hybrid ensemble learning approach utilizing light gradient boosting machine and category boosting model for lifestyle-based prediction of type-II diabetes mellitus. *Journal of Data Analysis and Information Processing*, 11, 480-511. doi: 10.4236/jdaip.2023.114025.
[24] Okhrimenko, A., & Kussul, N. (2023). Data mining of machine learning datasets for hard case identification. *International Scientific Technical Journal "Problems of Control and Informatics"*, 68(4), 84-95. doi: 10.34229/1028-0979-2023-4-7.
[25] Pai, K.-C., Su, S.-A., Chan, M.-C., Wu, C.-L., & Chao, W.-C. (2022). Explainable machine learning approach to predict extubation in critically ill ventilated patients: A retrospective study in Central Taiwan. *BMC Anesthesiology*, 22, article number 351. doi: 10.1186/s12871-022-01888-y.
[26] Priasni, T.O., & Oswari, T. (2021). Comparative study of standalone classifier and ensemble classifier. *Telecommunication Computing Electronics and Control*, 19(5), 1747-1754. doi: 10.12928/telkomnika.v19i5.19508.
[27] Rismayati, R., Ismarmiaty, I., & Hidayat, S. (2022). Ensemble implementation for predicting student graduation with classification algorithm. *International Journal of Engineering and Computer Science Applications*, 1(1), 35-42. doi: 10.30812/ijecsa.v1i1.1805.
[28] Shang, H., *et al*. (2022). Using machine learning models to predict HBeAg seroconversion in CHB patients receiving pegylated interferon-α monotherapy. *Journal of Clinical Laboratory Analysis*, 36(11), article number e24667. doi: 10.1002/jcla.24667.
[29] Sun, A., & Sun, F. (2024). Prediction of diabetes using supervised classification. *Journal of Emerging Investigators*, 7, 1-7. doi: 10.59720/23-062.
[30] Vik, D., Pii, D., Mudaliar, C., Nørregaard-Madsen, M., & Kontijevskis, A. (2024). Performance and robustness of small molecule retention time prediction with molecular graph neural networks in industrial drug discovery campaigns. *Scientific Reports*, 14, article number 8733. doi: 10.1038/s41598-024-59620-4.
[31] Wu, Q., Deng, L., Jiang, Y., & Zhang, H. (2022). Application of the machine-learning model to improve prediction of non-sentinel lymph node metastasis status among breast cancer patients. *Frontiers in Surgery*, 9, article number 797377. doi: 10.3389/fsurg.2022.797377.
[32] Zhang, C., Liu, L., Wang, C., Sun, X., Wang, H., Wang, J., & Cai, M. (2024). Prefer: Prompt ensemble learning via feedback-reflect-refine. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17), 19525-19532. doi: 10.1609/aaai.v38i17.29924.
[33] Zhyliak, M., & Horodetska, O. (2023). Predicting stroke risk via handling the imbalanced data. *Biomedical Engineering and Technology*, 12(4), 1-8. doi: 10.20535/2617-8974.2023.12.292870.
[34] Zuo, Y., & Drummond, T. (2020). *Residual likelihood forests*. doi: 10.48550/arxiv.2011.02086.

Градієнтний бустінг як інструмент для вирішення задач класифікації в умовах обмежених даних

Микола Киричек
Магістр
Київський національний університет імені Тараса Шевченка
01033, вул. Володимирська, 60, м. Київ, Україна
https://orcid.org/0009-0009-4571-123X

Анотація. У сучасному машинному навчанні поставало питання ефективної побудови класифікаційних моделей при недостатньому обсязі навчальної інформації. Мета дослідження – проаналізувати можливості використання градієнтного бустінгу для розв'язання задач класифікації в умовах обмежених даних. Методологія дослідження базувалася на комплексному аналізі провідних реалізацій градієнтного бустінгу: XGBoost, LightGBM та HistGradientBoosting. Основну увагу було зосереджено на вивченні механізмів регуляризації, стратегій оптимізації гіперпараметрів та адаптивних технік навчання в умовах малих вибірок. Дослідження спрямовувалося на виявлення архітектурних особливостей алгоритмів, здатних забезпечити високу точність класифікації при мінімальному обсязі даних. Встановлено, що запропоновані алгоритми продемонстрували значний потенціал для ефективного розв'язання класифікаційних задач. Виявлено, що механізми shrinkage та subsampling дозволили суттєво підвищити узагальнюючу здатність моделей. Результати дослідження розширили теоретичні уявлення про ансамблеві методи машинного навчання та окреслили перспективні напрямки адаптації алгоритмів до специфічних умов обмежених інформаційних ресурсів. Досліджено, що XGBoost, LightGBM та HistGradientBoosting мають унікальні архітектурні особливості, які дозволяють ефективно працювати з різними типами даних. Встановлено, що механізми внутрішньої регуляризації цих алгоритмів забезпечили стійкість до перенавчання та високу точність прогнозування. Показано потенціал градієнтного бустінгу для вирішення складних класифікаційних задач у медицині, фінансах та інших галузях з обмеженими інформаційними ресурсами. Практичне значення роботи полягало в розробці методологічних рекомендацій щодо вибору та налаштування алгоритмів градієнтного бустінгу для різних типів класифікаційних задач. Отримані результати будуть корисні для подальшого розвитку методів машинного навчання

Ключові слова: машинне навчання; адаптивні алгоритми; оптимізація моделей; XGBoost; гіперпараметризація