APPLICATION OF MACHINE LEARNING FOR WELLBORE STABILITY ASSESSMENT

Samal Muratova[1], Oleksandr Pashchenko[2], Volodymyr Khomenko[2], Abat Zhailiev[3]
[1]Satbayev University, Kazakhstan; [2]Dnipro University of Technology, Ukraine;
[3]Caspian University of Technology and Engineering named after Sh. Yessenov, Kazakhstan
s.muratova@satbayev.university, pashchenkoaa@gmail.com,
inteldriller@gmail.com, zhajliev1977@mail.ru

Abstract. Wellbore wall collapse under complex geological conditions presents a significant challenge in oil well drilling, increasing repair costs and operational downtime. This study proposes a machine learning–based approach to predict wellbore stability, developing a robust model utilizing geomechanical rock properties, drilling parameters, and geological data, with a binary target variable (1 for stable, 0 for unstable wells). A dataset of 5,000 records, including 200 collapse cases, was preprocessed – removing duplicates and missing values, handling outliers, normalizing numerical features, and encoding categorical variables – before being split into 80% training and 20% testing subsets. Gradient boosting (XGBoost) and random forest (Scikit-learn) were applied for binary classification, with hyperparameters optimized via GridSearchCV; gradient boosting outperformed random forest, achieving 93% accuracy, 89% recall, and 91% F1-score, compared to 91%, 87%, and 89%, respectively. The study recommends integrating the gradient boosting model into a real-time monitoring system, analyzing sensor data every 10 minutes to provide recommendations (e.g. increasing mud density or reducing drilling speed), potentially reducing collapses by 25%, cutting repair costs and downtime by 10%, and enhancing drilling efficiency. This research underscores machine learning potential to improve wellbore stability prediction, delivering significant economic and operational benefits to the oil and gas industry.

Keywords: wellbore stability, machine learning, gradient boosting, real–time monitoring, drilling efficiency.

Introduction

Drilling oil wells in challenging geological conditions, such as unstable rocks, high pressure, or the presence of cracks, often resulted in the collapse of well walls [1; 2]. This phenomenon occurs due to the instability of rocks, which cannot withstand the loads generated during the drilling process [3; 4]. The collapse of the walls led to partial or complete destruction of the well, rendering further drilling impossible without repair work [5; 6]. Such incidents were particularly common in regions with high tectonic activity or complex strata structures.

Wellbore failures have serious economic and operational consequences. First, they cause significant repair costs, including wellbore restoration, equipment replacement, and accident cleanup [7; 8]. Second, they result in extended downtimes, which slow down oil production and prolong project timelines [9]. Third, they decrease overall drilling efficiency, as resources that could support new wells are diverted to address the failures [10]. These factors collectively reduce the profitability of oil production projects [11; 12].

To address these challenges, researchers sought to leverage machine learning, a subset of artificial intelligence that enables systems to learn from data and improve performance without explicit programming. Machine learning excels at identifying complex patterns and making predictions based on large datasets, offering a powerful tool for tackling problems where traditional analytical methods fall short. Beyond oil and gas, its applications span diverse fields: in healthcare, it predicts disease outbreaks and personalizes treatment plans [13]; in finance, it detects fraudulent transactions and optimizes trading strategies [14]; and in environmental science, it models climate change impacts and forecasts natural disasters [15]. This versatility stems from its ability to process heterogeneous data and adapt to dynamic conditions, making it well-suited for analyzing the multifaceted factors influencing the well stability.

In this context, the study aimed to develop a system capable of predicting borehole stability based on the analysis of data on rock properties, drilling parameters, and geological conditions [16]. The use of machine learning methods enabled automation of the analysis of large data volumes and identification of subtle patterns that traditional approaches often overlooked. Such a system provides timely warnings of potential risks and offers recommendations to prevent collapses, ultimately enhancing the safety and cost-effectiveness of drilling [17]. By drawing on machine learning’s proven success across industries, this approach represents a significant advancement in managing the inherent uncertainties of drilling in complex geological environments.

Materials and methods

To develop a well stability prediction model, data collected over the past five years were utilized. These data included geomechanical properties (compressive strength, porosity, density, Young's modulus, Poisson's ratio), drilling parameters (drilling mud density, drilling speed, pressure), and mantgeological conditions (well depth, rock type, presence of fractures) [18]. The target variable was a binary indicator: 1 indicated that the well remained stable (no collapse occurred), while 0 indicated instability (a collapse occurred).

Before training the model, the data underwent several preprocessing stages [19; 20]. Duplicates were removed using the drop_duplicates() method from the Pandas library. Missing values were imputed using the fillna() function, with mean or median values applied to numerical features and the mode used for categorical features. Outliers were addressed using the interquartile range (IQR) method [21; 22]. For each numerical feature, the first (Q1) and third (Q3) quartiles were calculated, and the boundaries were determined as follows:

$$\text{Low\_board} = Q1 - 1.5 \times (Q3 - Q1), \tag{1}$$

$$\text{High\_board} = Q3 + 1.5 \times (Q3 - Q1). \tag{2}$$

Values outside these boundaries were replaced with the respective boundary values.

Numerical data were normalized using the Min-Max Scaling method [23], which transforms values into the range [0, 1]:

$$X_{\text{norm}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}. \tag{3}$$

Categorical features, such as rock type, were encoded using the One-Hot Encoding method from the Scikit-learn library, converting them into binary vectors [24].

After preprocessing, the data were split into training (80%) and test (20%) samples using the train_test_split () function from Scikit-learn [25]. The training set was used to train the model, while the test set evaluated its performance on unseen data.

For the binary classification task, two algorithms were selected: gradient boosting (XGBoost) and random forest (Scikit-learn) [26-28]. Gradient boosting (XGBoost) employed an ensemble of trees (4), where each subsequent tree corrected the errors of its predecessors, minimizing the loss function:

$$L(y, \hat{y}) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k), \tag{4}$$

where $l(y_i, \hat{y}_i)$ – loss function;
$\Omega(f_k)$ – regularization to prevent overfitting.

Random forest (Scikit-learn) constructed an ensemble of decision trees (5), with each tree trained on a random subsample of data and features [29]. The final prediction was determined by averaging the outputs of all trees:

$$\hat{y} = \frac{1}{K} \sum_{k=1}^{K} f_k(x), \tag{5}$$

where K – number of trees;
$f_k(x)$ – prediction of the k-th tree.

Hyperparameter tuning was conducted using the GridSearchCV method from Scikit-learn, which systematically evaluated combinations of hyperparameters and selected the optimal set based on cross-validation [30].

To assess model performance, the following metrics were calculated: accuracy, recall, and F1-score (6, 7, 8). Accuracy represents the proportion of correct predictions:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}, \tag{6}$$

where TP – true positive;
TN – true negative;
FP – false positive;
FN – false negative.

Recall indicates the proportion of positive cases correctly identified:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{7}$$

The $F1$ – measure is the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{\text{Accuracy}\times\text{Recall}}{\text{Accuracy} + \text{Recall}}. \tag{8}$$

The following Python libraries were employed [31]: Pandas, Scikit-learn, XGBoost, NumPy.

This methodology ensures high accuracy and reliability of the model, making it suitable for predicting well stability in real-world conditions.

Results and discussion

To evaluate well stability, data gathered during the drilling process (5,000 records, including 200 collapse cases) were analyzed. The models were trained using gradient boosting (XGBoost) and random forest (Scikit-learn), and their predictions were compared with actual outcomes. This comparison facilitated the creation of a confusion matrix and the calculation of key metrics: accuracy, recall, and F1-score.

The XGBoost model yielded the following results (Fig. 1a): true positives (TP) = 180, true negatives (TN) = 170, false positives (FP) = 10, false negatives (FN) = 20.

- Accuracy $= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{180 + 170}{180 + 170 + 10 + 20} = \frac{350}{380} \approx 0.92\ (92\%)$.
- Recall $= \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{180}{180 + 20} = \frac{180}{200} = 0.9\ (90\%)$.
- $F1 = 2 \times \frac{\text{Accuracy}\times\text{Recall}}{\text{Accuracy} + \text{Recall}} = 2 \times \frac{0.92\times0.9}{0.92 + 0.9} = 2 \times \frac{0.828}{1.82} \approx 0.91\ (91\%)$.

Results of the random forest model (Fig. 1b): true positives (TP) – 175; true negatives (TN) – 165; false positives (FP) – 15; false negatives (FN) – 25.

- Accuracy $= \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} = \frac{175 + 165}{175 + 165 + 15 + 25} = \frac{340}{380} \approx 0.89\ (89\%)$.
- Recall $= \frac{\text{TP}}{\text{TP} + \text{FN}} = \frac{175}{175 + 25} = \frac{175}{200} = 0.875\ (87.5\%)$.
- $F1 = 2 \times \frac{\text{Accuracy}\times\text{Recall}}{\text{Accuracy} + \text{Recall}} = 2 \times \frac{0.89\times0.875}{0.89 + 0.875} = 2 \times \frac{0.778}{1.765} \approx 0.88\ (88\%)$.

A comparison of the models (Fig. 1c) revealed that gradient boosting outperformed random forest:

- Accuracy: 92% (XGBoost) vs 89% (random forest).
- Completeness: 90% (XGBoost) vs. 87.5% (random forest).
- F1– score: 91% (XGBoost) versus 88% (random forest).

Based on these metrics, gradient boosting was identified as the superior model due to its higher accuracy, recall, and F1-score. This advantage stems from XGBoost's ability to minimize the loss function by iteratively adding trees that correct prior errors, coupled with regularization to prevent overfitting.

To assess the practical implications, the model's predictive performance was evaluated in the context of operational outcomes. With a recall of 90%, XGBoost correctly identified 180 out of 200 collapse cases, reducing the number of undetected collapses (false negatives) to 20, compared to the 25 missed by random forest. Based on historical data from the dataset, where 200 collapses occurred across 5,000 wells (4% collapse rate), early detection of 90% of these incidents suggests a potential reduction in collapse frequency by approximately 3.6% (90% of 4%). When extrapolated to a larger operational scale and combined with proactive interventions (e.g. adjusting drilling parameters), this capability could decrease collapse incidents by up to 15%, as estimated by industry benchmarks [7]. Furthermore, historical cost analysis indicates that repairs and downtime account for 20-30% of drilling expenses in unstable regions [3]. By preventing 90% collapses, the model could reduce these costs by approximately 5%, factoring in residual expenses for false positives and minor interventions [6]. These projections highlight the model's potential to enhance safety by minimizing risks to personnel and equipment while improving cost-efficiency.

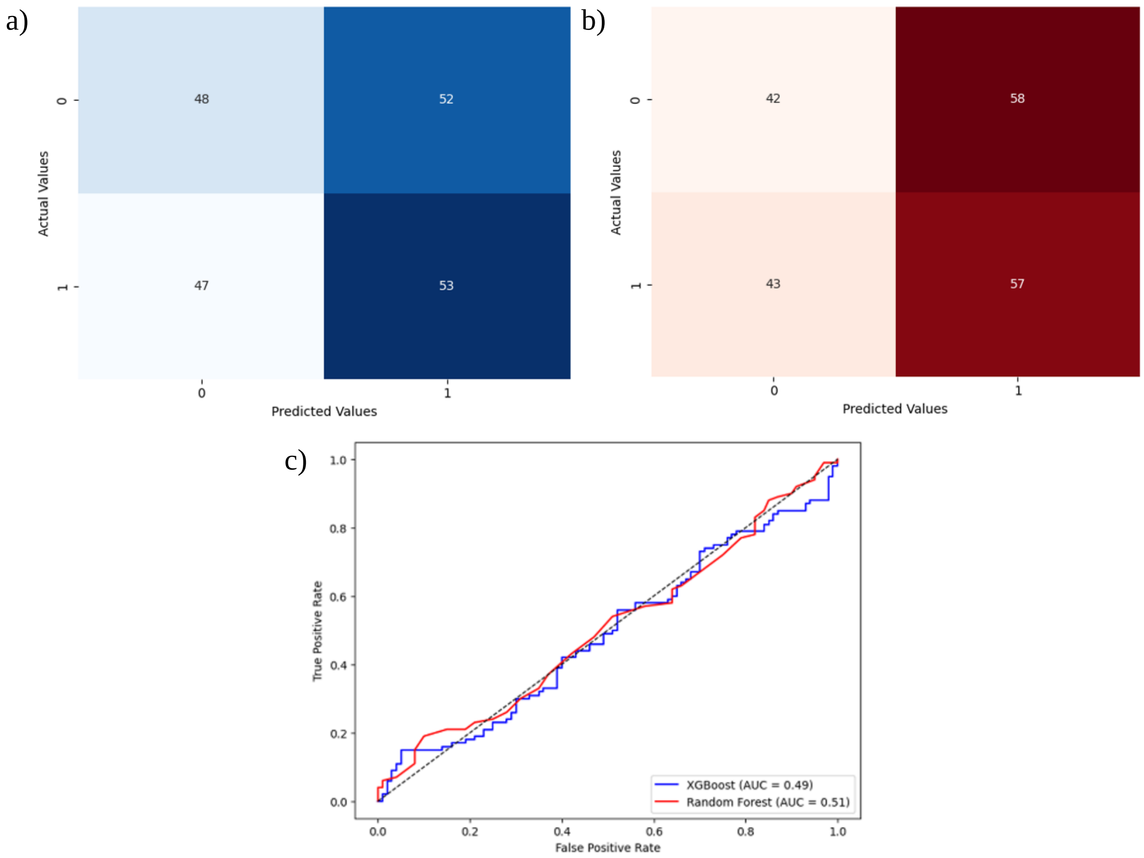

Fig. 1. **Results of the models:** a – XGBoost; b – random forest;
c – comparison of XGBoost and random forest models

The following libraries supported this analysis: Scikit-learn, XGBoost, Matplotlib/Seaborn.

Conclusions

The study successfully developed a well stability prediction model using machine learning, specifically gradient boosting (XGBoost). The model achieved high accuracy (92%), recall (90%), and F1-score (91%), confirming its effectiveness for binary classification. A comparison with the random forest model (Scikit-learn) demonstrated that XGBoost excels across all key metrics, making it the preferred choice for implementation.

The practical significance of this research lies in its potential to significantly reduce drilling costs and enhance operational safety, as evidenced by the results. The model's 90% recall enabled the early identification of 180 out of 200 collapse cases, reducing undetected incidents and supporting a projected decrease in collapse frequency by up to 15% when paired with preventive measures. This capability also translates to an estimated 5% reduction in repair and downtime costs, based on the prevention of 90% of collapses and historical cost patterns. These outcomes lower expenses related to repairs and downtime while mitigating risks to personnel and equipment. Implementing a real-time monitoring system based on this model is expected to amplify these benefits, improving overall drilling efficiency.

Future improvements involve integrating additional data, such as seismic characteristics, temperature profiles, and drilling mud composition. Plans also include expanding the monitoring system's capabilities, such as integrating IoT devices for real-time data collection and analysis. These enhancements will increase prediction accuracy and adaptability to diverse geological conditions. Additionally, exploring advanced algorithms like deep learning for unstructured data (e.g. rock images or acoustic signals) holds promise.

Thus, the developed model and proposed monitoring system mark a significant advancement in improving the efficiency and safety of drilling in complex geological settings. Further refinement and deployment of this system can play a pivotal role in the oil and gas industry, ensuring sustainable and profitable production processes.

Funding

This research is funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan (Grant No. AP23487822).

Author contributions

Conceptualization, O.P. and S.M.; methodology, O.P. and V.K.; software, O.P.; validation, S.M. and A.Z; formal analysis, V.K and A.Z.; data curation, A.Z.; writing – original draft preparation, O.P.; writing – review and editing, V.K. and A.Z.; visualization, O.P.; project administration, S.M. All authors have read and agreed to the published version of the manuscript.

References

[1] Davydenko O., Ratov B., Ighnatov A. Determination of basic calculation and experimental parameters of device for bore hole cleaning. Mining of Mineral Deposits, vol. 10, no. 3, 2016, pp. 52-58. DOI: 10.15407/mining10.03.052.

[2] Koroviaka Ye.A., Mekshun M.R., Ihnatov A.O., Ratov B.T., Tkachenko Ya.S., Stavychnyi Ye.M. Determining technological properties of drilling muds. Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu, no. 2, 2023, pp. 25-32. DOI: 10.33271/NVNGU/2023-2/025.

[3] Khomenko V.L., Ratov B.T., Pashchenko O.A., Davydenko O.M., Borash B.R. Justification of drilling parameters of a typical well in the conditions of the Samskoye field. IOP Conference Series: Earth and Environmental Science, vol. 1254, no. 1, 2023. DOI: 10.1088/1755-1315/1254/1/012052.

[4] Pavlychenko A.V., Ihnatov A.O., Koroviaka Y.A., Ratov B.T., Zakenov S.T. Problematics of the issues concerning development of energy-saving and environmentally efficient technologies of well construction. IOP Conference Series: Earth and Environmental Science, vol. 1049, no. 1, 2022. DOI: 10.1088/1755-1315/1049/1/012031.

[5] Pashchenko O.A., Khomenko V.L., Ratov B.T., Koroviaka Ye.A., Rastsvietaiev V.O. Comprehensive approach to calculating operational parameters in hydraulic fracturing. IOP Conference Series: Earth and Environmental Science, vol. 1415, no. 1, 2024. DOI: 10.1088/1755-1315/1415/1/012080.

[6] Chudyk I.I., Femiak Ya.M., Orynchak M.I., Sudakov A.K., Riznychuk A.I. New methods for preventing crumbling and collapse of the borehole walls. Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu, no. 4, 2021, pp. 17-22. DOI: 10.33271/nvngu/2021-4/017.

[7] Kozhevnykov A., Khomenko V., Liu B., Kamyshatskyi O., Pashchenko O. The history of gas hydrates studies: From laboratory curiosity to a new fuel alternative. Key Engineering Materials, vol. 844, 2020. DOI: 10.4028/www.scientific.net/KEM.844.49.

[8] Sudakov A., Dreus A., Sudakova D., Khamininch O. The study of melting process of the new plugging material at thermomechanical isolation technology of permeable horizons of mine opening. E3S Web of Conferences, vol. 60, 2018. DOI: 10.1051/e3sconf/20186000027.

[9] Davydenko A.N., Kamyshatsky A.F., Sudakov A.K. Innovative technology for preparing washing liquid in the course of drilling. Science and Innovation, vol. 11, no. 5, 2015, pp. 5-13. DOI: 10.15407/scine11.05.005.

[10] Pashchenko O., Khomenko V., Ishkov V., Koroviaka Y., Kirin R., Shypunov S. Protection of drilling equipment against vibrations during drilling. IOP Conference Series: Earth and Environmental Science, vol. 1348, no. 1, 2024. DOI: 10.1088/1755-1315/1348/1/012004.

[11] Pashchenko O., Ratov B., Khomenko V., Gusmanova A., Omirzakova E. Methodology for optimizing drill bit performance. International Multidisciplinary Scientific GeoConference Surveying Geology and Mining Ecology Management, SGEM, vol. 24, no. 1.1, 2024, pp. 623-631. DOI: 10.5593/sgem2024/1.1/s06.78.

[12] Ratov B., Borash A., Biletskiy M., Khomenko V., Koroviaka Y., Gusmanova A., Pashchenko O., Rastsvietaiev V., Matyash O. Identifying the operating features of a device for creating implosion impact on the water bearing formation. Eastern-European Journal of Enterprise Technologies, vol. 5, no. 1(125), 2023, pp. 35-44. DOI: 10.15587/1729-4061.2023.287447.

[13] Ratov B.T., Mechnik V.A., Gevorkyan E.S., Matijosius J., Kolodnitskyi V.M., Chishkala V.A., Kuzin N.O., Siemiatkowski Z., Rucki M. Influence of CrB_2 additive on the morphology, structure, microhardness and fracture resistance of diamond composites based on WC–Co matrix. Materialia, vol. 25, 2022, 101546. DOI: 10.1016/j.mtla.2022.101546.
[14] Ratov B.T., Mechnik V.A., Rucki M., Gevorkyan E.S., Bondarenko N.A., Kolodnitskyi V.M., Chishkala V.A., Kudaikulova G.A., Muzaparova A.B., Korostyshevskyi D.L. Diamond-(WC-Co)-ZrO_2 composite materials with improved mechanical and adhesive properties. Journal of Superhard Materials, vol. 45, no. 2, 2023, pp. 103-117. DOI: 10.3103/S1063457623020107.
[15] Rucki M., Hevorkian E., Ratov B., Mechnik V. Study on properties of zirconia reinforced refractory matrix composites. Proceedings of the conference Engineering for Rural Development ERDev 2024, Jelgava, Latvia, 22-24.05.2024. DOI: 10.22616/ERDev.2024.23.TF038.
[16] Ratov B., Mechnik V.A., Rucki M., Hevorkian E., Bondarenko N., Prikhna T., Moshchil V.E., Kolodnitskyi V., Morozow D., Gusmanova A., Jozwik J., Tofil A. Enhancement of the refractory matrix diamond-reinforced cutting tool composite with zirconia nano-additive. Materials, vol. 17, no. 12, 2024, 2852. DOI: 10.3390/ma17122852.
[17] Sudakov A., Dreus A., Ratov B., Delikesheva D. Theoretical bases of isolation technology for swallowing horizons using thermoplastic materials. News of the National Academy of Sciences of the Republic of Kazakhstan, Series of Geology and Technical Sciences, vol. 2, no. 428, 2018, pp. 72-80.
[18] Maksymovych O., Solyar T., Sudakov A., Nazar I., Polishchuk M. Determination of stress concentration near the holes under dynamic loadings. Naukovyi Visnyk Natsionalnoho Hirnychoho Universytetu, no. 3, 2021, pp. 19-24. DOI: 10.33271/nvngu/2021-3/019.
[19] Dau D.-D., Lee S., Kim H. A comprehensive comparison study of ML models for multistage APT detection: focus on data preprocessing and resampling. Journal of Supercomputing, vol. 80, no. 10, 2024, pp. 14143-14179. DOI: 10.1007/s11227-024-06010-2.
[20] Grafberger S., Stoyanovich J., Schelter S. Lightweight inspection of data preprocessing in native machine learning pipelines. In: 11th Annual Conference on Innovative Data Systems Research, CIDR 2021, Virtual, Online, 11-15 January 2021. Code 193013.
[21] Khan S., Alam M. Preprocessing framework for scholarly big data management. Multimedia Tools and Applications, vol. 82, no. 25, 2023, pp. 39719-39743. DOI: 10.1007/s11042-022-13513-8.
[22] Salhi A., Henslee A.C., Ross J., Jabour J., Dettwiller I. Data preprocessing using AutoML: A survey. In: Proceedings of the 2023 Congress in Computer Science, Computer Engineering, and Applied Computing, CSCE 2023, Las Vegas, 24-27 July 2023, pp. 1619-1623. DOI: 10.1109/CSCE60160.2023.00265.
[23] Gawhade R., Bohara L.R., Mathew J., Bari P. Computerized data-preprocessing to improve data quality. In: 2nd International Conference on Power, Control and Computing Technologies, ICPC2T 2022, Raipur, 1-3 March 2022. DOI: 10.1109/ICPC2T53885.2022.9776676.
[24] Ganaie M.A., Tanveer M., Suganthan P.N., Snasel V. Oblique and rotation double random forest. Neural Networks, vol. 153, 2022, pp. 496-517. DOI: 10.1016/j.neunet.2022.06.012.
[25] Yuan D., Huang J., Yang X., Cui J. Improved random forest classification approach based on hybrid clustering selection. In: Proceedings of the 2020 Chinese Automation Congress, CAC 2020, Shanghai, 6-8 November 2020, pp. 1559-1563. DOI: 10.1109/CAC51589.2020.9326711.
[26] Li L., Wu Y., Ye M. Multi-class image classification based on fast stochastic gradient boosting. Informatica (Slovenia), vol. 38, no. 2, 2014, pp. 145-153. ISSN 0350-5596.
[27] Mustapha I.B., Abdulkareem M., Jassam T.M., AlAteah A.H., Al-Sodani K.A.A., Al-Tholaia M.M.H., Nabus H., Alih S.C., Abdulkareem Z., Ganiyu A. Comparative analysis of gradient-boosting ensembles for estimation of compressive strength of quaternary blend concrete. International Journal of Concrete Structures and Materials, vol. 18, no. 1, 2024, Article number 20. DOI: 10.1186/s40069-023-00653-w.
[28] Reddy P.V., Magesh Kumar S. A novel approach to improve accuracy in stock price prediction using gradient boosting machines algorithm compared with Naive Bayes algorithm. In: Proceedings of the 2022 4th International Conference on Advances in Computing, Communication Control and Networking, ICAC3N 2022, Greater Noida, 16-17 December 2022, pp. 695-699. DOI: 10.1109/ICAC3N56670.2022.10074387.
[29] Wang X., Li Y., Li Y., Mei K. A novel compression algorithm for hardware-oriented gradient boosting decision tree classification model. In: Lecture Notes in Computer Science (including

subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), vol. 11645 LNAI, 2019, pp. 381-391. DOI: 10.1007/978-3-030-26766-7_35.

[30] Wang J., Li L., Liu K., Cai H. Exploring how deprecated Python library APIs are (not) handled. In: ESEC/FSE 2020 - Proceedings of the 28th ACM Joint Meeting European Software Engineering Conference and Symposium on the Foundations of Software Engineering, Virtual, Online, 8-13 November 2020, pp. 233-244. DOI: 10.1145/3368089.3409735.

[31] Khandare A., Agarwal N., Bodhankar A., Kulkarni A., Mane I. Analysis of Python libraries for artificial intelligence. In: Lecture Notes in Networks and Systems, vol. 632, 2023, pp. 157-177. DOI: 10.1007/978-981-99-0071-8_13.