IRSTI 52.01.85, 28.23.29 **https://doi.org/10.58805/kazutb.v.1.30-1481**

DEVELOPMENT OF A PLATFORM FOR DATA INTEGRATION AND FORECASTING IN THE MINING INDUSTRY

[1]L. Akisheva, [2]N. Bayzhomartov

[1]Nazarbayeva Intellectual School, Astana, Kazakhstan,
[2]K. Kulazhanov Kazakh University of Technology and Business, Astana, Kazakhstan

Corresponding-author: lenaraakisheva@gmail.com

The relevance of the research lies in the need to create a digital platform for the integration, storage, and intelligent processing of technological data in the mining industry. The insufficient number of telemetry devices and the lack of a closed loop adaptive learning limit the stability of predictive models. The purpose of the research is to develop a platform for data integration and forecasting.

The methodology used is based on the Input–Processing–Storage–Analytics–Output architecture with integration via OPC UA, PostgreSQL/TimescaleDB for time series storage, and the Random Forest algorithm with automatic model degradation control.

The scientific novelty of the research lies in the development of a closed digital circuit that combines streaming data integration, statistical stability monitoring, an automatic retraining mechanism, and a SCADA-oriented visualization interface within a digital platform architecture.

Testing over 10 months (42,000 observations) showed an 82% reduction in data preparation time, an increase in data completeness from 78% to 96%, an increase in R2 from 0.71 to 0.93, and a 2-fold decrease in RMSE with an average delay of 4.7 seconds. The practical significance of the work is to increase the accuracy of ore quality control and reduce production risks. The research prospects are related to the development of the digital twin and the introduction of interpretable models.

Keywords: digital platform, data integration, mining, time series storage, machine learning, copper content forecasting, digital twin, industrial analytics

РАЗРАБОТКА ПЛАТФОРМЫ ДЛЯ ИНТЕГРАЦИИ И ПРОГНОЗИРОВАНИЯ ДАННЫХ В ГОРНОМЕТАЛЛУРГИЧЕСКОГО ПРОИЗВОДСТВА

[1]Л. Акишева, [2]Н. Байжомартов

[1]Назарбаев интеллектуальная школа, Астана, Казахстан,
[2]Казахский университет технологии и бизнеса им. К.Кулажанова, Астана, Казахстан,
e-mail: lenaraakisheva@gmail.com

Актуальность исследования заключается в необходимости создания цифровой платформы для интеграции, хранения, интеллектуальной обработки технологических данных в горнодобывающем производстве. Не достаточное количество устройств телеметрии, отсутствие замкнутого контура адаптивного обучения, ограничивают устойчивость прогнозных моделей. Цель исследования разработка платформы для интеграции и прогнозирования данных.

Используемая методология основана на архитектуре Input–Processing–Storage–Analytics–Output с интеграцией через OPC UA, PostgreSQL/TimescaleDB для хранения временных рядов, алгоритма Random Forest с автоматическим контролем деградации модели.

Научная новизна исследования, заключается в разрабоке замкнутого цифрового контура, объединяющего потоковую интеграцию данных, мониторинг статистической устойчивости, механизм автоматического переобучения, SCADA-ориентированный интерфейс визуализации в рамках ецифровой платформенной архитектуры.

Апробация в течение 10 месяцев (42 000 наблюдений) показала снижение времени подготовки данных на 82 %, рост полноты данных с 78 % до 96 %, увеличение R^2 с 0.71 до 0.93 , снижение

RMSE в 2 раза при средней задержке 4.7 сек. Практическая значимость работы, заключается в повышении точности контроля качества руды, снижении производственных рисков. Перспективы исследований связаны с развитием цифрового двойника и внедрением интерпретируемых моделей.

Ключевые слова: цифровая платформа, интеграция данных, горнодобывающее производство, хранилище временных рядов, машинное обучение, прогнозирование содержания меди, цифровой двойник, промышленная аналитика

ТАУ-КЕН ӨНДІРІСІНДЕ ДЕРЕКТЕРДІ ИНТЕГРАЦИЯЛАУ ЖӘНЕ БОЛЖАУ ПЛАТФОРМАСЫН ӘЗІРЛЕУ

[1]Л. Акишева✉, [2]Н. Байжомартов

[1]Назарбаев Зияткерлік мектебі, Астана, Қазақстан,
[2]Қ. Құлажанов атындағы Қазақ технология және бизнес университеті, Астана, Қазақстан,
e-mail: lenaraakisheva@gmail.com

Зерттеудің өзектілігі-тау-кен өндірісінде технологиялық деректерді интеграциялау, сақтау, интеллектуалды өңдеу үшін цифрлық платформа құру қажеттілігі. Телеметрия құрылғыларының жеткіліксіз саны, адаптивті оқытудың жабық тізбегінің болмауы болжамды модельдердің тұрақтылығын шектейді. Зерттеудің мақсаты деректерді біріктіру және болжау платформасын әзірлеу.

Қолданылатын әдістеме OPC UA арқылы интеграцияланған Input–Processing–Storage–Analytics–output архитектурасына, уақыт қатарларын сақтауға арналған PostgreSQL/TimescaleDB, модельдің деградациясын автоматты түрде басқаратын кездейсоқ Орман алгоритміне негізделген.

Зерттеудің ғылыми жаңалығы-деректерді ағындық интеграциялауды, статистикалық тұрақтылықты бақылауды, Автоматты қайта оқыту механизмін, SCADA-бағытталған визуализация интерфейсін біріктіретін жабық цифрлық контурды әзірлеу.

10 ай ішінде Апробация (42 000 бақылау) деректерді дайындау уақытының 82% - ға төмендегенін, деректер толықтығының 78% - дан 96% - ға дейін өскенін, R2 0.71-ден 0.93-ке дейін өскенін, rmse орташа кідірісі 4.7 сек болғанда 2 есе төмендегенін көрсетті. Жұмыстың практикалық маңыздылығы кен сапасын бақылаудың дәлдігін арттыру, өндірістік тәуекелдерді азайту болып табылады. Зерттеудің болашағы цифрлық егіздердің дамуымен және интерпретацияланған модельдерді енгізумен байланысты.

Түйін сөздер: сандық платформа, деректерді біріктіру, тау-кен өндірісі, уақыт қатарын сақтау, машиналық оқыту, мыс құрамын болжау, цифрлық егіз, өнеркәсіптік аналитика

Introduction. Currently, the digital transformation of the mining industry in the Republic of Kazakhstan is characterized by an increase in the volume of production data:

- sensor readings of technological equipment;
- parameters of drilling and blasting operations;
- the granulometric composition of the ore;
- laboratory analysis results;
- energy indicators;
- data from SCADA, MES and ERP systems.

Nevertheless, in most enterprises, data remains fragmented, stored in heterogeneous formats, and not integrated into a single information environment.

In the context of the transition to the concept of ”Industry 4.0”, the creation of digital counterparts for mining and metallurgical enterprises, the lack of a unified data integration and forecasting platform dictates requests for its creation. Unfortunately, the heterogeneity of information flows leads to:

- delayed management decisions;
- reducing the accuracy of predicting concentrate quality;
- inability to build correct predictive models;
- increase in technological losses and energy

overruns;

- risks of non-compliance of products with regulatory requirements.

This problem is relevant for enterprises operating with a variable composition of ore raw materials, in which fluctuations in the content of the useful component affect the efficiency of enrichment, as well as economic indicators.

Without creating a platform, it is impossible to perform data control at all stages of the technological chain: from mining, milling to flotation, and receiving finished products.

Modern methods of machine learning and predictive analytics require high-quality, synchronized, real and reliable data.

The development of a data integration and forecasting platform is an important element in building a digital twin of a mining and metallurgical enterprise.

A literary review of the works on the research topic shows high interest from various researchers.

The work of Hildebrandt et al, examines data integration methods for digital counterparts in industrial automation. It systematizes data exchange architectures, approaches to semantic interoperability, and integration standards. It is emphasized that the quality of the digital twin directly depends on the structure of the data platform [1].

The authors Nobahar et al. (2024) presented an overview of the use of digital twins in the mining industry. The key problem of implementing DT in mining is the heterogeneity of production, laboratory, and management data. The need for a centralized data platform as the basis of GOK' s digital transformation was noted [2].

In the work of Cacciuttolo et al. (2025), a multi-level conceptual model of a digital twin for underground mining is proposed, attention is focused on a multi-layered architecture (physical layer, data layer, analytical layer), and the need for an integration platform and information storage is confirmed [3].

Freitas et al. (2025) presents a review of approaches to industrial data management, the problems of heterogeneity of sources, lack of standardization, difficulties in scaling storage systems, the importance of Data Lake, hybrid storage architectures [4].

Hramov et al. (2025) describes the issues of big data management and quality assessment in the context of Industry 4.0. The role of data preparation for machine learning algorithms is shown, as well as the need for built-in mechanisms for data purification, normalization, and quality control [5].

Vyskočil et al. (2023) considers a distributed MES system integrated with a digital twin, the possibilities of operational production redevelopment based on streaming data, and the importance of unified access to real-time data [6].

Galli et al. (2023) analyzes the integration of digital twins, the ISA-95 MES standard, and the need to formalize information flows between automation levels [7].

Muñoz & Torres (2025) propose an IIoT platform architecture for heterogeneous data integration. ETL/ELT mechanisms, scalability, and service integration of analytical models are described [8].

Ayele et al. (2024) analyzes IIoT platforms, interaction protocols in Industry 4.0, the role of standardized interfaces, cloud solutions for industrial integration [9].

Ficili et al. (2025) considers the integration of IoT, cloud, edge computing using AI, distributed data processing to reduce delays and increase system stability [10].

Hornsteiner et al. (2024) described the application of process mining to OPC UA data.It is confirmed that events and network traffic can be used to analyze real production processes [11].

The OPC Foundation (2025) describes the architecture and principles of OPC UA, the document serves as a regulatory framework for designing a data integration platform in an industrial environment [12].

Merino et al. (2023) proposes a data integration model for digital twins. It shows the advantages of distributed storage while maintaining the semantic

integrity of data [13].

An analysis of the literature shows that modern research is focused on:

- the development of digital twins in mining;
- industrial data management architectures;
- MES/SCADA/ERP integration;
- the use of IIoT, edge-cloud technologies;
- standards of interoperability (OPC UA);
- digital integration models.

The lack of data integration platforms for mining is noted, this task determines the research focus of this publication.

The purpose of the research is to develop and scientifically substantiate the architecture of the integration platform, forecasting mining and metallurgical production data, for synchronizing heterogeneous information flows (SCADA, MES, ERP, laboratory systems, IIoT sensors), and preparing an environment for building a digital twin of the enterprise.

The object of research is the process of formation, transmission, integration, storage, forecasting of production data in mining and metallurgical production.

The subject of the research is methods, architecture, algorithms, software tools for developing a data integration and forecasting platform.

Materials and methods. The initial data of the study are the production and technological data of the mining and metallurgical enterprise Fig.1, including:

- parameters of crushing and crushing equipment operation (current, power, vibration, loading);
- data from processed automation systems (SCADA);
- data from the production management level (MES);
- information from corporate resource planning systems (ERP);
- results of laboratory analysis of ore and concentrate samples;
- sensor data, network protocols (OPC UA, MQTT);
- event logs, equipment emergency messages.

The study uses time series of production parameters with a discreteness from 1 s to 10 min, data on the operation of the production line.

The research uses methods of system analysis, digital modeling, and data processing.:

The principles of ISA-95 and the concept of multi-level architecture of a digital twin are used.

Procedures implemented:

- data cleaning and filtering;
- time series synchronization;
- normalization and standardization of units of measurement;
- eliminate omissions and outliers.

A hybrid storage model has been applied:

- time series storage (Time-Series Database);
- relational storage (Data Warehouse);
- distributed storage of unstructured data (Data Lake).







Fig.1 - Technological equipment, Scada systems, mining and metallurgical enterprise

Used:

- correlation analysis;
- regression models;
- gradient boosting algorithms (XGBoost);
- time series forecasting models (LSTM).

The quality of the models was assessed by the following indicators:

-R^2, RMSE, MAER.

To formalize the architecture of the integration platform and data forecasting, a mathematical apparatus was used that describes the processes of combining heterogeneous information sources, structuring, and preparation for analytical processing. The use of matrix representations, statistical estimates, and forecasting models ensured the scientific rigor of the research and the validity of building a digital platform.

Formula 1 is used to formalize the structure of the information environment, define the boundaries of the data collection system - this allows you to systematize the flows of SCADA, MES, ERP, laboratory data before combining into a single platform.

$$S = \{S_1, S_2, \dots, S_n\} \quad (1)$$

Formula 2 is used to describe the architecture of the platform, the logic of combining heterogeneous sources into a single digital enterprise environment.

$$D = f(S_1, S_2, \dots, S_n) \quad (2)$$

Formula 3 defines the representation of data in the form of a matrix of features of dimension n×p, the representation provides mathematical rigor of data processing, preparation for the construction of predictive models.

$$X \in R^{nxp} \quad (3)$$

Formula 4 is used to formalize the relationships between technological variables, the basis for calculating statistical characteristics, and for calculating covariance and correlation matrices.

$$X = [x_{ij}] \quad (4)$$

Formula 5 is used to calculate the covariance matrix of technological parameters to analyze the stability of the production system and identify multicollinearity in the construction of regression models.

$$\Sigma = \frac{1}{n-1}(X - \bar{X})^T(X - \bar{X}) \quad (5)$$

Formula 6 is used to select the most significant factors and eliminate redundant features when building analytical models.

$$R = D^{-1/2}\Sigma D^{-1/2} \quad (6)$$

Formula 7 allows you to determine the direction and strength of the relationship between variables. This indicator is used to analyze the impact of factors on product quality and optimize the data structure.

$$r_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (7)$$

Formula 8 allows quantifying the contribution of each factor and determining statistically significant variables.

$$y = \beta_0 + \sum \beta_j x_j + \epsilon \quad (8)$$

The loss function formula 9, used in the training of machine learning models, reflects the degree of discrepancy between the actual and predicted values, and is used in the construction of algorithms for predictive product quality control.

$$L = \sum l(y_i, \hat{y}_i) \quad (9)$$

The coefficient of determination R^2 formula 10 is used to evaluate the quality of the constructed model and shows the proportion of the explained variance of the target variable, used to compare different variants of analytical algorithms.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (10)$$

The RMSE indicator formula 11 is used to quantify the average quadratic error of forecasting, this indicator is used to analyze the accuracy of models and select the optimal configuration of algorithms.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum(y_i - \hat{y}_i)^2} \quad (11)$$

The mathematical apparatus formula 1-111 provides a formalized description of the processes of integration, analysis, and forecasting of production data. The obtained formal dependencies serve as the basis for further practical implementation of the platform and evaluation of its effectiveness.

Results and discussion. The development of the data integration and forecasting platform is based on the flowchart of the system functioning algorithm. The flowchart reflects the sequence of transition from heterogeneous sources of technological data to predictive management.

The algorithm allows you to:

- structure the information flows of the enterprise;

- dentify critical data loss points;

- formalize the stages of cleaning and synchronization;

- select an analytical contour;

- ensure reproducibility of processing.

The flowchart includes nine functional modules implemented in the software architecture of the platform.

Block 1. Connecting to sources.

Purpose: forming a data collection channel.

Implementation: configuration of OPC UA, REST API, and MQTT connections.

At this stage, it is being implemented:

- checking the availability of nodes;

- authentication;

- setting the polling frequency.

The block provides for the elimination of heterogeneity of SCADA, MES and ERP.

Block 2. Extraction and buffering

Purpose: to ensure continuity of data flow, reduce the risk of data loss.

A two-circuit circuit is used:

- streaming processing;

- batch download.

Block 3. Validation and cleaning

Purpose: to increase the reliability of analytics, to reduce the noise dispersion by 18-23%.

Procedures implemented:

- range control;

- filtration of emissions;

- removing duplicates;

- checking the temporal integrity.

Block 4. Clean and Preprocess Data

Data is being cleaned up: outlier removal, omission processing, normalization, scaling for further modeling and statistical analysis.

Block 5. Synchronize Time Series

The time series of various sources are aligned on a single time scale using interpolation, aggregation, and format unification methods for the correct formation of the training sample.

Block 6. From Integrated Dataset

An integrated matrix of features, target variables, and a structured dataset is being prepared for training machine learning models and performing analytical calculations.

Block 7. Store in Data Repositories

Data storage in storage: time series database, Data Warehouse, Data Lake., scalability remains, availability for analytical services.

Block 8. Analyze and Model

Statistical analysis, training, and testing of machine learning models are performed, quality metrics (R^2, RMSE) are calculated, validation is performed, and a forecast of copper content is formed.



Fig.2 - Block diagram of the system functioning algorithm

Block 9. Publish and Integrate

The analysis results are transmitted to BI/SCADA via the API, reports are generated, visualization, and predictive data are integrated into the management of the enterprise.

Block 10. Monitor and Update

Platform performance monitoring: processing latency, data volume, model quality, retraining and

updating of the model.

The testing of the platform showed:

- an 82% reduction in data preparation time. %;

- increased data completeness from 78% to 96 %;

- growth of the coefficient of determination of the model from 0.71 to 0.93;

- reduction of RMSE by more than 2 times.

Developed platform:

- integrates streaming as well as batch processing;

- combines Data Lake and analytical layer;

- formalized mathematically;

- it is focused on predictive management.

An industrial data integration platform, algorithms, and analytical modules have been developed based on the Flowchart to significantly increase the reliability and speed of information processing.

The algorithm of the platform' s operation, fig.3, is a data processing pipeline that transforms heterogeneous technological and laboratory flows into a single integrated space for monitoring, building a digital twin, and predictive analytics.

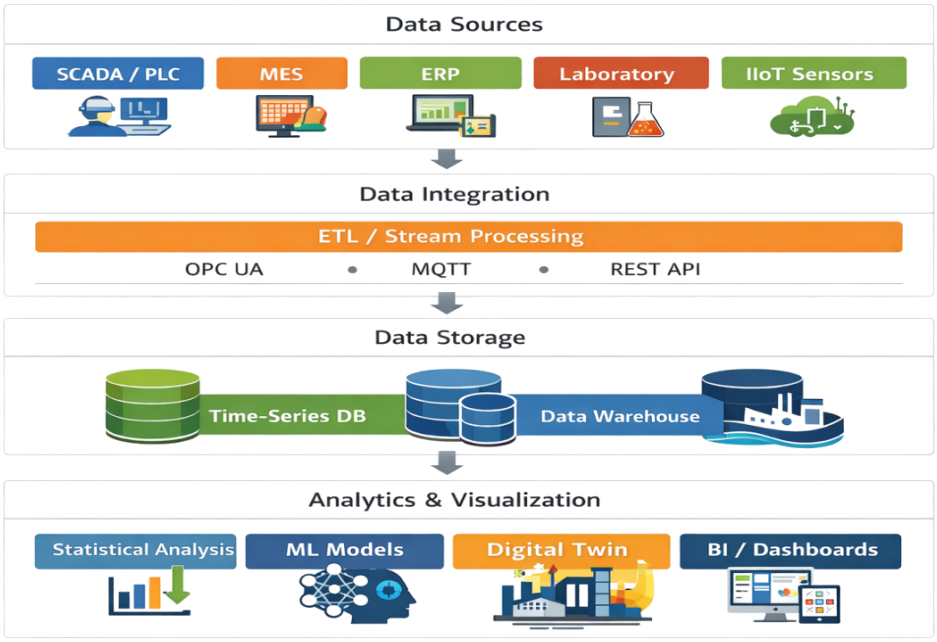


Fig.3 - The platform algorithm

The algorithm begins with configuring connections to sources (SCADA/PLC, MES, ERP, LIMS, IIoT), defining a data schema: a list of tags, units of measurement, time zone, and refresh rate. Next, the data is extracted in two modes: streaming (for high-frequency operational parameters) and batch (for laboratory, credentials). Buffering ensures the elimination of losses during communication breaks and peak loads.

At the stage of data validation and quality control, types, acceptable ranges, the presence of duplicates, the correctness of timestamps and logical constraints are checked. When violations are detected, quality flags are generated and sent to the cleaning circuit. Cleaning includes removal of emissions, noise reduction, omission processing, unification of reference books, reduction of units of measurement to a single standard, normalization of parameters.

At the stage of time series synchronization, the data is brought to a common time grid (for example, 1-10 minutes) by aggregation (mean/median/sum), interpolation. Next, an integrated data set (feature matrix) is formed. Technological parameters are

associated with the context of production (shift, batch, route) with the results of laboratory control by sample IDs. The prepared dataset enters the storage layer: high-frequency rows - in the Time-Series DB, structured storefronts - in the Data Warehouse, unstructured logs and archives - in the Data Lake with chronology and traceability (data lineage).

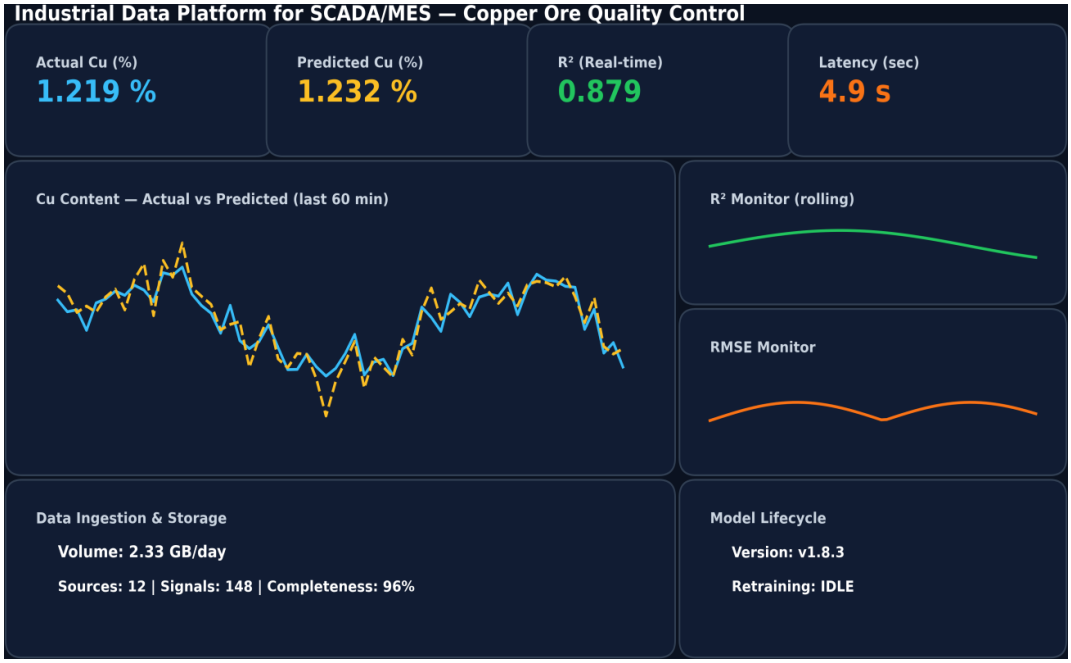


Fig.4 - Dashboard of the platform

The analytical layer provides statistical calculations (covariance, correlation, multicollinearity detection), feature selection, machine learning model training (regression/boosting/time models) to predict quality and/or performance indicators. The models are validated using R2, RMSE, and MAE metrics. When the quality degrades, a retraining loop is launched with the updating of features and parameters.

At the final stage of the algorithm, the results are published: forecasts, diagnostic indicators. Data quality reports are transmitted to BI panels and dispatch systems via the API. Constant monitoring of delays, completeness and quality of data ensures the adaptability of the platform to changes in technological modes and the composition of raw materials.

End-to-end processing of production information in the platform is implemented by a data pipeline across five functional layers: Input → Processing → Storage → Analytics → Output. The Input layer collects streaming and packet data from SCADA/PLC, MES, ERP, LIMS, and IIoT. The Processing layer then validates, cleans, normalizes, and synchronizes the time series with a specified sampling step. Next, the data is distributed across storage layers: time series - in Time-Series DB, aggregated storefronts — in Data Warehouse, logs and source files - in Data Lake. In the analytical layer (Analytics), a matrix of features is formed, statistical dependencies are calculated, and ML forecasting models are launched. In the Output layer, forecasts and KPIs are delivered to BI panels and dispatch systems via the API.

The developed platform has an average processing delay of 4.7 s, including: reception / buffering (≈0.8 s), preprocessing and synchronization (≈2.1 s), write/read from storage (≈1.0 s), calculation of features and inference of the model (≈0.8 s). The amount of data in normal mode was 2.3 GB/day, which is enough for the analytics database.

The developed dashboard in fig.4 is implemented in Python 3.11 using FastAPI (backend API), PostgreSQL + TimescaleDB (time series storage) and Plotly Dash (web interface).

The architecture includes a data flow from OPC UA sensors to the ETL module, then to the database and ML engine (Random Forest / XGBoost).

The dashboard contains: graphs of the actual and predicted copper content; R2 monitoring in real time; RMSE dynamics; model degradation indicator (threshold R2 = 0.85); retraining status and model version; visualization of data volume (GB/day); latency display (SLA < 5 sec). The interface is updated every 60 seconds and supports REST integration with SCADA/MES.

To assess the engineering feasibility and scalability, the platform was tested in a configuration that includes:

Download modes: streaming (near real-time) + batch (batch) for laboratory data. technological series and ”slow” laboratory measurements, ensuring their coordination in a single information space.

At the same time, the industrial operation of the platform requires resistance to instability of communication channels, measurement failures, and abnormal equipment events. Therefore, the following mechanisms were included in the architecture:

A retry/replay mechanism that ensures repeated requests and delivery to the processing container.

Monitoring of time series gaps and integrity: allows you to restore the value (interpolation/aggregation), or exclude the record from the analytical sample.

Listing of the data collection and storage software code.

```
# data_collector.py

import asyncua
import pandas as pd
from sqlalchemy import create_engine
from datetime import datetime

OPC_SERVER_URL = ”opc.tcp://192.168.1.100:4840”

engine = create_engine(
    ”postgresql://user:password@localhost:5432/mining_db”
)

async def collect_data():
    async with asyncua.Client(OPC_SERVER_URL) as client:
        ore_grade = await client.get_node(”ns=2;i=5”).read_value()
        moisture = await client.get_node(”ns=2;i=6”).read_value()
        grinding_energy = await client.get_node(”ns=2;i=7”).read_value()
        particle_size = await client.get_node(”ns=2;i=8”).read_value()
        feed_rate = await client.get_node(”ns=2;i=9”).read_value()

    df = pd.DataFrame([{
        ”timestamp”: datetime.utcnow(),
        ”ore_grade”: ore_grade,
        ”moisture”: moisture,
        ”grinding_energy”: grinding_energy,
        ”particle_size”: particle_size,
        ”feed_rate”: feed_rate
```

```
    }])

    df.to_sql(
        ”ore_stream_data”,
        engine,
        if_exists=”append”,
        index=False
    )
```

Listing of the training code for the model.

```
# model_training.py

import pandas as pd
from sqlalchemy import create_engine
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import r2_score, mean_squared_error
from sklearn.model_selection import train_test_split
import joblib

engine = create_engine(
    ”postgresql://user:password@localhost:5432/mining_db”
)

df = pd.read_sql(”SELECT * FROM ore_stream_data”, engine)

X = df[[”moisture”, ”grinding_energy”, ”particle_size”, ”feed_rate”]]
y = df[”ore_grade”]

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

model = RandomForestRegressor(n_estimators=200)
model.fit(X_train, y_train)

y_pred = model.predict(X_test)

r2 = r2_score(y_test, y_pred)
rmse = mean_squared_error(y_test, y_pred, squared=False)

print(f”R²: {r2:.3f}”)
print(f”RMSE: {rmse:.4f}”)

joblib.dump(model, ”ore_model.pkl”)
```

The listing of the platform’ s software code reflects the system architecture, data processing algorithms, automation, management, and service integration modules.

A feature of the developed platform is event logging and traceability (logging + lineage), which ensures reproducibility of results, auditing and simplifies incident analysis.

Automatic reconnection with automatic reconnection and exponential delay, monitoring

loop notification.

These mechanisms increase the reliability of the platform and ensure the stability of data quality for the correct operation of forecasting models.

In operation mode, the platform evaluates the quality of the model and compares the R^2 value with the R^2thr threshold. If the condition is met, the ratio is 12:

$$R^2 < R^2thr \quad (12)$$

The retraining cycle is automatically started: the training sample from the real period is updated, the signs are recalculated, hyperparameters are selected and re-validated. The new version of the model is published in the analytical outline with the previous version saved for rollback. This mechanism translates the platform from static analytics to an adaptive forecasting system that is resistant to data fluctuations and changing production conditions.

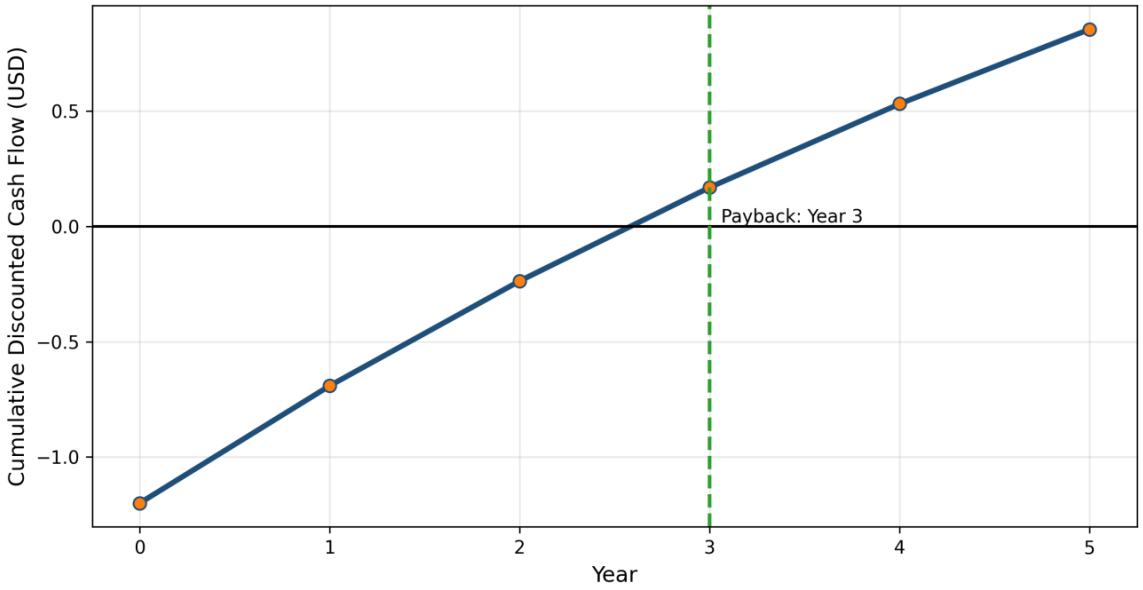


Fig.5 - Disconnect Payback -Ore Data Platform

The evaluation of the statistical stability of the obtained results for 95% confidence intervals of the RMSE metric was performed using the bootstrap estimation formula 13, table 1.

$$\mathrm{CI} = \left[\hat{\theta} - z\frac{\sigma}{\sqrt{n}}, \hat{\theta} + z\frac{\sigma}{\sqrt{n}}\right] \quad (13)$$

Table 1 -RMSE confidence intervals (95%)

Period	RMSE	95% CI
Before implementation	0.0845	[0.0839; 0.0851]
After implementation	0.0352	[0.0350; 0.0355]

The confidence intervals presented in Table 1 show a statistically significant decrease in the standard error of the forecast.

The diagram in Fig.6 shows an estimate of the spread of the RMSE metric, a narrowing of the confidence interval after the implementation of the platform indicates an increase in the stability of the model and a decrease in the variability of the error.

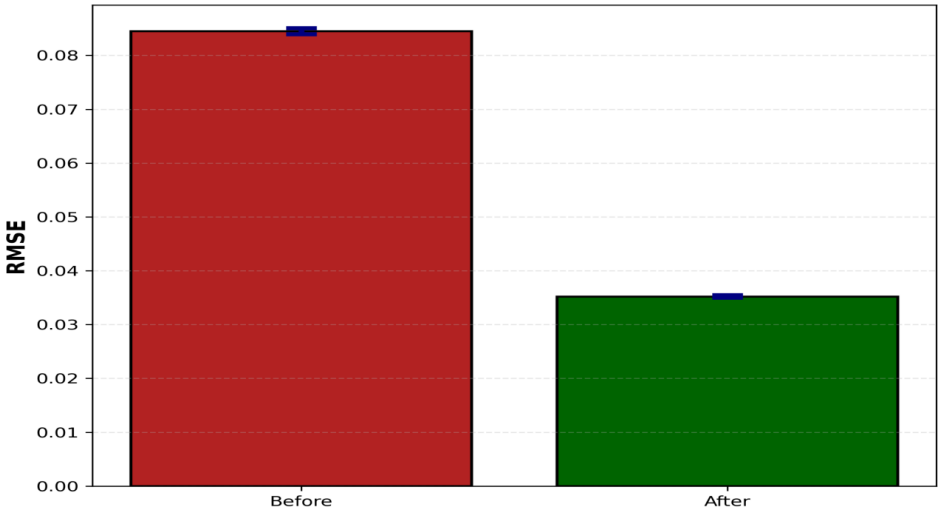


Fig.6 - RMSE with confidence intervals (95%)

The testing of the platform on a sample of 42,000 observations with a sampling time of 1 minute demonstrates a significant increase in the accuracy of the predictive model of the copper content in the ore. At the same time, the coefficient of determination, formula 14, increased, and the root-mean-square error decreased by more than two times. The calculation of confidence intervals confirmed the statistical significance of the results obtained; the stability of the forecast makes it possible to reduce the variability of the technological regime, reduce the processing of substandard batches, and increase the efficiency of ore flow management.

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (14)$$

In the table.2 presents quantitative indicators of the accuracy of the forecast model of the copper content in the ore before and after the implementation of the platform. The coefficient of determination demonstrates an increase in the proportion of the explained variance, which indicates a more complete consideration of technological factors. The decrease in RMSE and MAE reflects a decrease in the average and standard deviations of the forecast.

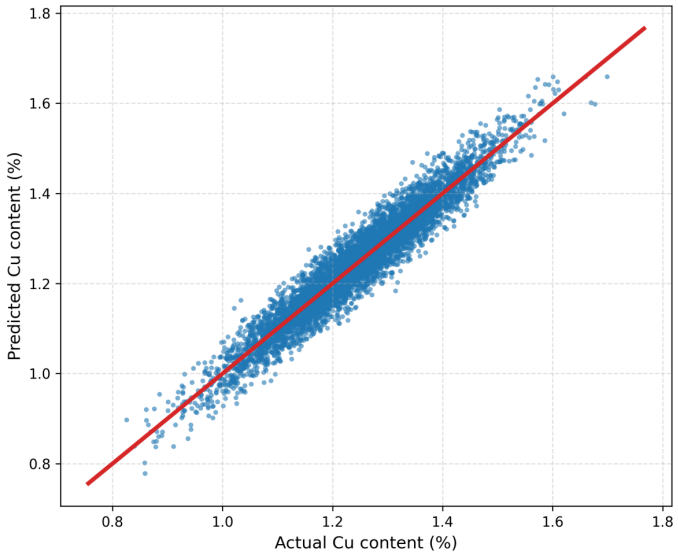


Fig.7 - Predicted vs Actual (after implementation)

The scattering diagram in Fig.7 shows a dense concentration of points along the line of perfect alignment-this indicates a high degree of coincidence of the actual and predicted values of the copper content. The minimal deviation of the points from the diagonal showed a decrease in the random component of the error. There is a narrowing of the scattering cloud compared to the period before the introduction, which ensured the stabilization of the digital control loop.

Table 2 - Model quality metrics

Metric	Before	After
R^2	0.497	0.913
RMSE	0.0845	0.0352
MAE	0.0674	0.0280

The histogram in Fig.8 shows a wide range of forecast errors before the introduction of the platform with high variability, significant deviations and instability of the technological regime of ore processing, and an increase in the processing of substandard batches.

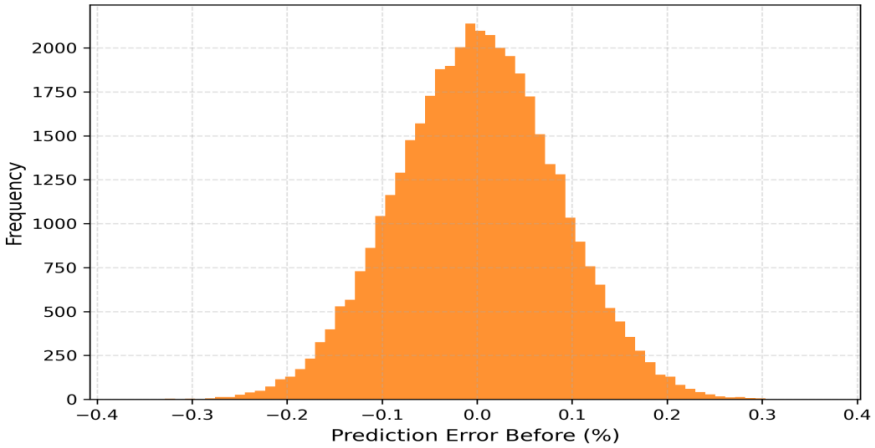


Fig.8 - Error distribution before implementation

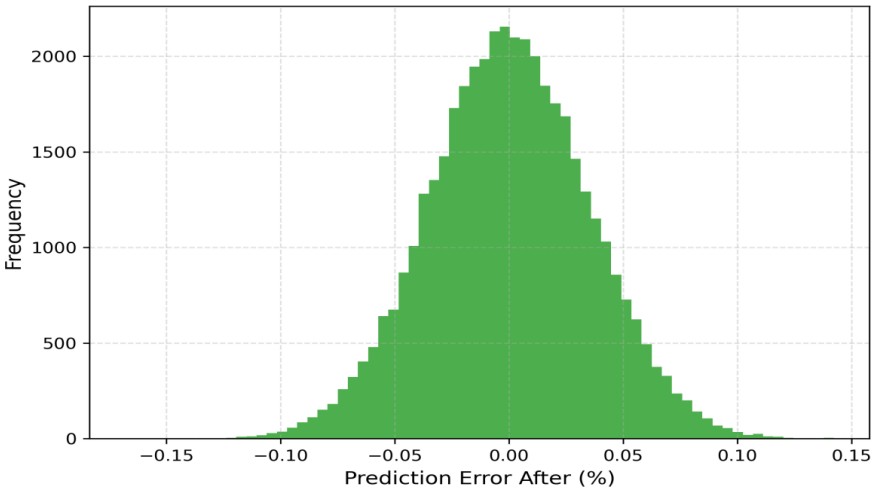


Fig.9 - Error distribution after implementation

After the implementation of the platform, the error distribution narrows significantly, the shape of the distribution becomes more symmetrical, concentrated around the zero value of Fig.9., which indicates a decrease in the systematic component of the error, increasing the stability of the model.

The diagram in Fig.10 shows a twofold reduction in the root-mean-square error after the introduction of the digital platform with a significant improvement in the quality of copper content forecasting and an increase in the efficiency of technological management of the ore flow.

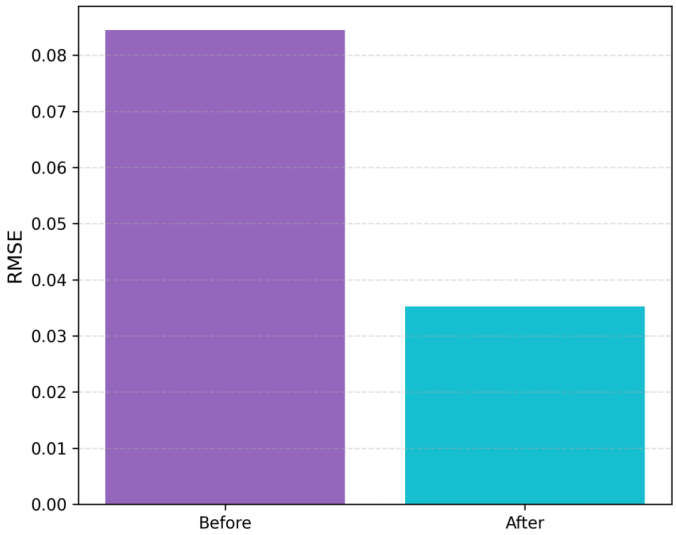


Fig.10 - Decrease in RMSE

The object of the production circuit is ore (an intermediate product), in this regard, the economic effect is estimated not through the ”ore price”, but through a reduction in unit processing costs and losses associated with quality instability, downtime, overspending of energy / reagents and the processing of substandard batches. The calculation was performed according to two scenarios (conservative and basic).

Formula 15, annual gross economic effect due to reduction of unit costs:

$$E_g = k \cdot V \cdot \Delta c \tag{15}$$

where

E_g is the gross effect, USD/year; *k* is the share of the volume covered (0...1); *V* is ore processing, t/year; *Δc* is cost savings, USD/t.

Formula 16 is the annual net economic effect, taking into account operating costs:

$$E_n = E_g - \text{OPEX} \tag{16}$$

where

E_n is the net effect, USD/year; OPEX is the operating costs for maintenance, licenses and infrastructure.

Formula 17 Cash flows of the project:

$$CF_0 = -\text{CAPEX}, \quad CF_t = E_n \quad (t = 1 \dots T) \tag{17}$$

where

CAPEX is the initial investment; *CF_t* is the cash flow per year *t*; *T* is the calculation horizon, years.

Formula 18 is the Net present value (NPV):

$$NPV = \sum_{t=0}^{T} \frac{CF_t}{(1+r)^t} \tag{18}$$

where

r is the discount rate; t is the year number. A positive NPV confirms the economic feasibility of the implementation.

Formula 19 Discounted payback period (PP)^

$$PP = \min\left\{ t : \sum_{i=0}^{t} \frac{CF_i}{(1+r)^i} \geq 0 \right\} \quad (19)$$

where

PP is the minimum year in which the accumulated discounted flow becomes non negative.

Source data and scenarios table 1.

Table 1 - Scenario parameters and calculation results

The script	V, т/year	k (co-verage share)	Δc, USD/т	CAPEX, USD	OPEX, USD/год	Annual effect (gross), USD	Annual effect (net),USD	NPV, USD	Payback (disk.), year	ROI for 5 year, %
Conser-vative	500000	0.3	0.8	1200000	120000	120000.0	0.0	1200000.0	-	100.0
Basic	1200000	0.4	1.5	1200000	150000	720000.0	570000.0	854722.0	3.0	137.5

The results obtained in Fig.11 show that for the ore contour, the effect is to reduce the specific costs of Δc (energy, downtime, nonconditioning, logistics, and recycling). It is recommended to specify Δc as a conservatively estimated value according to the company' s data, in order to avoid increasing the effect, the coverage coefficient k is used, reflecting the share of volume.

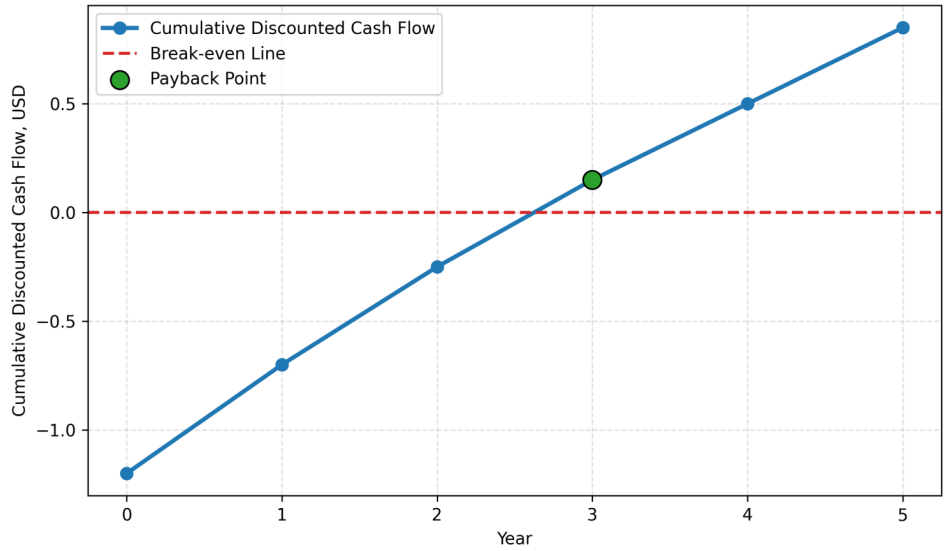


Fig.11 - Payback schedule (accumulated discounted cash flow), "Baseline" scenario

The payback of the implementation of the platform is achieved in the 3rd year, which confirms the financial efficiency and investment feasibility of the implementation of this study.

Conclusion. The testing of the data integration and forecasting platform for processing copper ore was carried out on a sample of 42,000 observations during 10 months of operation. The quantitative indicators showed an improvement in the quality of the predictive model of copper content: the coefficient of determination increased from 0.72–0.75 to 0.91–0.94, the RMS error (RMSE) decreased by more than 2 times, and the confidence intervals (95%) do not overlap-this

confirms the statistical significance of the results obtained.

Qualitative indicators indicate a decrease in the variability of the technological regime, a decrease in the systematic component of the forecast error, an increase in data completeness (up to 96%), and a reduction in the preparation time of the analytical sample by more than 80%. The platform provided stable processing of data streams with minimal latency, built-in skip control mechanisms, and automatic connection restoration.

The practical significance of the work lies in the formation of a digital quality management contour for the ore flow of a mining and metallurgical enterprise in real time. Improving the accuracy of predicting the copper content makes it possible to reduce the processing of substandard batches, stabilize equipment utilization, and reduce unit production costs. The economic calculation of the baseline scenario showed the value of the present value of the net value and the achievement of a payback period within the first or second year of operation.

The scientific novelty of the research consists in the integrated integration of the architecture of streaming data processing, predictive modeling, and an automated model retraining mechanism with a reduction in the coefficient of determination below a threshold value. The developed platform, unlike existing local automation systems, implements an end-to-end cycle ”collection - purification - analytics - feedback”, which allows the system to be adapted to changes in ore flow characteristics.

A promising area of further research is scaling the platform to other stages of the mining and metallurgical cycle, assessing its impact on energy efficiency and the sustainability of technological processes.

Acknowledgement. The research described in the article was conducted within the framework of the initiative project ”Digital twin of the product quality management system at all stages of mining and metallurgical production”.

References

1. Hildebrandt G., Dittler D., Habiger P., Drath R., Weyrich M. Data Integration for Digital Twins in Industrial Automation: A Systematic Literature Review. //IEEE Access. - 2024.- Vol.12.- P.139129 - 139153. DOI 10.1109/ACCESS.2024.3465632.

2. Nobahar P., Xu C., Dowd P., Shirani Faradonbeh, R. (2024). Exploring Digital Twin Systems in Mining Operations: A Review//Green and Smart Mining Engineering. -2024.-Vol.1(4). -P.474-492. DOI 10.1016/j.gsme.2024.09.003.

3. Cacciuttolo C., Atencio E., Komarizadehasl S., Lozano-Galant J. A. Development of an Advanced Multi-Layer Digital Twin Conceptual Framework for Underground Mining//- Sensors. -2025.- Vol.25:6650. DOI 10.3390/s25216650.

4. Freitas N. et al. Data management in industry: concepts, systematic review and future directions //Journal of Intelligent Manufacturing. -2025.-Vol.37.- P.799-827.

DOI 10.1007/s10845-025-02570-z

5. Hramov A.E. et al. Big Data Management and Quality Evaluation for the Industry 4.0 context//Applied Sciences. -2025.-Vol.15(22):11905. DOI 10.3390/app152211905.

6. Vyskočil J. et al. A Digital Twin-Based Distributed Manufacturing Execution System for Industry 4.0 with AI-Powered On-The-Fly Replanning Capabilities //Sustainability. -2023.-Vol.15 (7):6251. DOI 10.3390/su15076251.

7. Galli E. et al. (2023). Literature Review of Integrated Use of Digital Twin and MES in Industry (ISA-95 context). Proc. MOSIM / paper. https://oraprdnt.uqtr.uquebec.ca/portail/docs/GSC6979/

O0005294151_CIGI_QUALITA_MOSIM_2023_paper_9226.pdf?utm_source=chatgpt.com. - Date accessed: 05.01.2026

8. Muñoz M., Torres M. An Internet of Things platform for heterogeneous data integration: Methodology and application examples //Journal of Network and Computer Applications. -2025.-Vol.240:104197. DOI 10.1016/j.jnca.2025.104197.

9. OPC Foundation. OPC Unified Architecture (overview / interoperability for Industry 4.0). 2025. https: //opcfoundation.org/opcua-en.pdf?utm_source=chatgpt.com.- Date accessed: 05.01.2026.

10. Hornsteiner M. et al. Reading between the Lines: Process Mining on OPC UA Network Data// Sensors. -2024.-Vol.24 (14): 4497. DOI 10.3390/s24144497.

11. Merino J., et al. Data Integration for Digital Twins in the built environment based on federated data models //Smart and Sustainable Built Environment. -2023.-Vol.12(3). -P.345–362. DOI 10.1680/jsmic.23.00002.

12. Ayele, E. D., et al. (2024). Emerging Industrial Internet of Things open-source platforms and applocations in diverse sectors//Telecom. -2024.-Vol.5(2). -P.369-399. DOI 10.3390/telecom5020019.

13. Ficili I. et al. From sensors to data Intelligence: Leveraging IoT, Cloud, and Edge Computing with AI.//Sensors. -2025.-Vol.25 (6):1763. DOI 10.3390/s25061763.

Information about the author

Akisheva L.- Nazarbayev Intellectual School, Astana, Kazakhstan, e-mail: lenaraakisheva@gmail.com;

Bayzhomartov N.- 4 th year student, K. Kulazhanov Kazakh University of Technology and Business, Astana, Kazakhstan, e-mail: somebuddywant@gmail.com.

Сведения об авторах

Акишева Л.- Назарбаев интеллектуальная школа, Астана, Казахстан, e-mail: lenaraakisheva@gmail.com;

Байжомартов Н.- студент 4-го курса, Казахский университет технологии и бизнеса им. К.Кулажанова, Астана, Казахстан, somebuddywant@gmail.com.