Machine Learning for Train Delay Prediction and Economic Assessment in Railway Transport Digitalization

Sonya Sultanova[1], Nigora Ismailova[1], Olga Golubova[2], Guli Muhibova[3] and Nafosat Abdusalomova[1]
[1]Department of Accounting and Business, Tashkent State Transport University, Temiryolchilar Str. 1, 100167 Tashkent, Uzbekistan
[2]Department of Construction Economics and Real Estate Management, Belarusian National Technical University, 220013 Minsk, Belarus
[3]Department of Marketing and Management, Tashkent University of Architecture and Civil Engineering, Yangi Shahar Str. 9, 100194 Tashkent, Uzbekistan
sofsul@mail.ru, nigitigi@mail.ru, v.holubova@gmail.ru, guli_muhibova@gmail.uz, nafosat.abdusalomova@mail.ru

Keywords: Digital Technologies, Artificial Intelligence, Railway Transport, IT Technologies, Efficiency, Potential Opportunities.

Abstract: This article presents an analysis of current trends in the digitalization of the transport industry, focusing on the relationship between its advantages and corresponding potential opportunities. The study aims to identify potential opportunities arising from the integration of digital technologies and artificial intelligence (AI) in the transport sector. To achieve this goal, the principles and methodological foundations for the use of artificial intelligence in the digitalization of the transport industry are reviewed and analyzed. Complex linear and nonlinear relationships exist among the technical and operational parameters of the railway system; therefore, the study used statistical correlation analysis and machine-learning (ML) methods. ML methods are effective at identifying hidden patterns in large datasets (such as train movements) and improving forecasting accuracy. The scientific novelty of the study lies in the development of an integrated model for assessing the impact of digital technologies and machine-learning methods on the efficiency of the transport system. A machine-learning model for predicting train delays based on key operational factors of the railway infrastructure was developed. A model of balanced factors influencing the time and efficiency of the transport process was proposed, including a practical framework for implementing artificial intelligence in train schedule optimization. A cost-effectiveness calculation based on a 20% reduction in train delays yielded an estimated economic benefit of approximately 73.72 billion soums per year. The underlying cost-per-minute-of-delay figure (799,472 soums) is used as a fixed input to this calculation; its derivation (e.g., the cost components included, whether it is based on a specific study, tariff schedule, or official estimate, and the reference year/exchange rate/inflation assumptions) is not documented in the present manuscript and should be reported explicitly, since it materially affects the economic-impact estimate. The methodological basis of the research includes machine-learning modelling and comparative analytical forecasting methods used to study the implementation of intelligent systems and analyze their impact on the efficiency, safety, and sustainability of transport systems.

1 INTRODUCTION

For the Republic of Uzbekistan, the digital transformation of the transport system aligns with national priorities for infrastructure modernization and transit development. The implementation of digital transport management platforms, automated monitoring systems, and intelligent logistics solutions creates new opportunities for the sustainable development of the industry. With the globalization of production and the growth of transport flows, supply chains are becoming increasingly complex. Therefore, digital technologies and intelligent systems are emerging as key tools for the modernization of transport infrastructure. Artificial intelligence is being integrated into traffic management systems, vehicle monitoring, transport planning, and intelligent transport systems.

The methodological framework of the research includes machine-learning models and comparative analytical forecasting methods used to study the implementation of intelligent systems and analyze their impact on transport efficiency, safety, and sustainability. Based on statistical data on key railway performance indicators for the period 2019-2024, the interdependencies between the variables were identified, and a dataset of five key factors influencing train delays was constructed for machine-learning analysis.

1.1 Novelty and Contribution

Unlike previous studies that mainly discuss transport digitalization at a conceptual level, this research proposes a quantitative railway-focused machine-learning framework based on operational data from Uzbekistan Railways. The novelty of the study is threefold. First, it clearly distinguishes between the source operational dataset and the analytical machine-learning dataset used for model development. Second, it compares a baseline linear model with ensemble machine-learning models, demonstrating the higher predictive accuracy of nonlinear methods for train-delay forecasting. Third, it links prediction results with an economic-effect assessment, thereby connecting technical forecasting performance with practical managerial value. The study contributes to the development of intelligent decision-support systems for railway transport management in developing economies.

1.2 Literature Review

It goes without saying that digitalization of transport processes significantly improves the accuracy of passenger and freight traffic forecasts, reduces operating costs, and minimizes logistical risks. The effective use of machine learning algorithms to optimize routing, manage traffic, and improve transport safety has been substantiated in the works of foreign scientists such as [1]-[4].

Theoretical issues of integration of digital technologies and artificial intelligence are considered in the publications of domestic, Russian and foreign scientists [5]-[14] and several other researchers. Some modern studies of the authors are devoted to the study of the modern process and assessment of the impact of digital technologies on the profitability and sustainable development of transport companies.

Current trends in transport digitalization are consistent with strategic documents, including those of the World Bank and the International Transport Forum, which note that the implementation of intelligent transport systems contributes to the formation of a competitive market environment and accelerated economic growth.

Foreign researchers presented a detailed analysis of the role of digital technologies in the integration and development of transport and logistics infrastructure, citing examples of the use of modern digital technologies (GPS navigation, RFID technologies, the Internet of Things (IoT) and artificial intelligence (AI))[15].

The benefits of using digital technologies include increased efficiency, safety and reliability of transportation, data integration and automation of transport processes.

However, digital technologies in the development of transport and logistics infrastructure and the existing potential for further improving the efficiency, safety, and reliability of transportation have been insufficiently studied. This is a pressing issue in today's globalized world economy and in the face of increasing competition between modes of transport. The purpose of this study is to identify the potential for integrating digital technologies and artificial intelligence into the development of the transport sector.To achieve this goal, this article analyzes the potential benefits of digitalization and the use of artificial intelligence (AI), including increased transportation efficiency, cost reduction, enhanced safety, improved quality of transportation services, and decision optimization. Specific solutions for the practical use of AI methods in optimizing train schedules and reducing delays at railway stations are provided.

2 MATERIALS AND METHODS

The application of artificial intelligence in the transport sector offers numerous benefits and potential opportunities (Table 1) [16].

The benefits of integrating digital technologies and artificial intelligence into the transport sector, described in Table 1, were selected based on statistical data on key indicators affecting train delays of the railways of the Republic of Uzbekistan. To analyze the factors behind train delays, a dataset was created based on statistical data from the railway industry, Uzbekistan Railways, and the National State Committee of the Republic of Uzbekistan on Statistics for the period 2019-2024.

Table 1: Benefits of integrating digital technologies and artificial intelligence into the transport sector.

Advantages	Potential opportunities	Integration of digital technologies and artificial intelligence
1. Automated operational and technical analysis of transport process data.	Forecasting freight and passenger flows	Real-time data collection (Internet of Things) → AI-powered data processing for prediction and optimization
2. Automated operational management decision-making.	Optimization of logistics and transportation	GPS and telematics data transmission → traffic flow management algorithms
3. Machine learning	Forecasting the technical condition of vehicles	Big data analysis → machine learning models for effective management
4. Real-time computer monitoring	Road traffic monitoring, emergency detection.	Autonomous vehicles → interaction with intelligent infrastructure (V2X - vehicle-to-everything).

Table 2: Uzbekistan Railways Data train delay statistics (2019-2024).

Year	Number of trains (thousands)	Average temperature °C	Precipitation (mm)	Number of delayed trains	Average delay (min)
2019	48.2	15.1	340	820	5.8
2020	46.7	14.7	365	910	6.4
2021	49.5	15.4	330	780	5.9
2022	52.8	16.2	350	980	7.2
2023	55.6	16.8	320	1050	8.0
2024	58.4	17.3	335	1120	8.6

The initial database contained 3240 daily observations of railway operations collected for the period 2019-2024. For machine-learning modelling, the daily observations were aggregated into 24 monthly observations in order to reduce short-term noise and identify stable operational patterns. In addition, 6 annual observations are presented in Table 2 only for descriptive analysis of long-term trends and were not used for model training or testing. Thus, 3240 daily observations represent the source database, 24 monthly observations constitute the analytical dataset for machine-learning, and 6 annual observations are used exclusively for descriptive statistics. It should be noted that the arithmetic relationship between these figures is not fully transparent: 3,240 daily observations over a 2019–2024 period (approximately 2,190 calendar days for 6 years, or up to 72 months) do not straightforwardly aggregate into exactly 24 monthly observations under a simple daily-to-monthly averaging scheme. It is possible that the 24 monthly observations instead correspond to a 2-year subperiod, to monthly observations recorded per specific rail segment/route, or to another aggregation logic not detailed here; this discrepancy should be clarified explicitly by the authors in a future revision to ensure the dataset construction is fully reproducible.

To evaluate performance, Table 3 provides a comparison of machine learning models. Three additional methodological details should be clarified for full reproducibility. First, the source of the temperature and precipitation figures in Table 2 (e.g., specific meteorological stations, national hydrometeorological service annual reports) is not stated in the present description. Second, the hyperparameters used for the Gradient Boosting and Random Forest models (e.g., number of trees, maximum depth, learning rate) and for cross-validation are not reported in Table 3 and should be included in a future revision. Third, the correlation matrix in Table 5 shows uniformly high coefficients (0.82–1.00) among predictor variables, which raises the possibility of multicollinearity; this was not discussed in the manuscript and could affect the stability and interpretability of the regression-based model's coefficients in particular, even though it would not necessarily undermine the tree-based models' predictive accuracy. Variance inflation factor (VIF) diagnostics or predictor-set reduction are recommended as a check in future work.

Machine learning models (random forest and gradient boosting) demonstrated significantly higher prediction accuracy compared to the baseline approach.

The dataset was divided into a training set (70%) and a test set (30%). The monthly machine-learning dataset was divided into a training set (70%) and a test set (30%). Before model training, the data were normalized and checked for outliers. Annual aggregation was used only for descriptive presentation and was not part of the machine-learning training pipeline. (Table 2).

Table 3: Comparison of ML models.

Model	MAE	RMSE	R^2
Linear regression (baseline)	5.9	7.2	0.61
Random Forest	3.2	4.8	0.82
Gradient Boosting	2.8	4.1	0.89

The comparison shows that the ensemble models outperform the baseline linear regression model. While Linear Regression achieved an R^2 of 0.61, Random Forest improved predictive performance to 0.82, and Gradient Boosting produced the best result with an R^2 of 0.89 and the lowest prediction errors. Therefore, Gradient Boosting was selected as the final model for interpretation and economic-effect estimation.

The data selection was carried out by the authors based on an extended data stream of 24 observations for the ML model. Figure 1 shows a graph of the dynamics of train delays in Uzbekistan for the period 2019-2024.

The figure shows the dynamics of average train delays over the period 2019-2024, indicating a gradual increase in delays due to growing railway infrastructure load.

3 RESULTS

To assess the potential of digitalization and the impact of introducing artificial elements into the transport system, we propose using a systematic list of measurable indicators that can be divided into four groups, as shown in Table 4.

Table 4 summarizes the potential opportunities we believe to be the most effective: the technological foundation for digitalizing infrastructure. Modern digital and network technologies, including cloud communications, combined with digital data, form high-tech digital data processing centers.

The correlation matrix of factors influencing train delays is presented in Table 5.

The correlation matrix illustrates the relationships between key operational variables affecting train delays. The strongest correlations are observed between infrastructure load, train density, and the number of trains.

After 2021, there has been a steady trend of increasing delays, which may be due to the following reasons:

- increase in train traffic intensity;
- increase in the efficiency of using railway infrastructure;
- increase in the length of trains.

The results confirm the need to use machine learning models(XGBoost/Random Forest)to predict delays and optimize schedules.

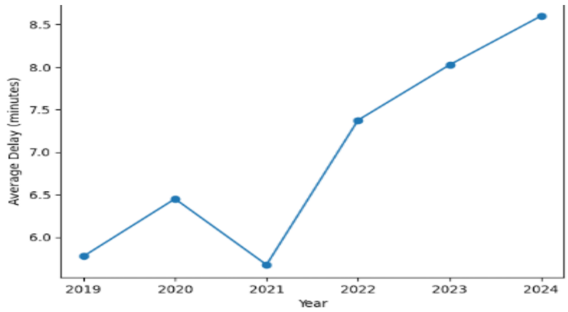

Figure 1: Dynamics of train delays in Uzbekistan for the period 2019-2024.

Figure 2 presents a structured cyclical model for implementing artificial intelligence (AI) in train schedule optimization in rail transport. It represents a closed-loop intelligent decision support system integrated into a rail traffic management system. In practice, the model forms an intelligent feedback control system.

Regression model and an alternative ensemble model, Random Forest. The evaluation was performed on the monthly dataset using MAE, RMSE, MAPE, and R^2. This comparison allows the predictive gain of the selected nonlinear model to be demonstrated explicitly rather than assumed.

The expected outcome of implementing AI in train schedule optimization is a potential increase in rail network capacity, a reduction in operational delays, and, consequently, a reduction in operating costs and an increase in train safety.

We propose a practical, schematic model for implementing artificial intelligence (AI) in train schedule optimization. We use a machine learning (ML) model to predict delays. The formalization of the train delay prediction problem involves a multi-level training sample of observations. For transport systems, we use Random Forest and Gradient Boosting functions, as they allow us to account for complex nonlinear relationships between factors. The function then takes the form:

$$D_i = \sum_{k=1}^{K} W_k T_k(x_i)$$

where $T_k(x_i)$ is the decision tree;

- W_k - weight of tree;
- K is the number of trees.
- $x_i=(TD_i,W_i,T_i,TL_i,L_i)$ — vector of input features characterizing the train movement conditions;
- Y_i - is the actual train delay time.

The model is an additive composition of decision trees. Standard statistical metrics were used to evaluate the accuracy of the machine learning model: MAE, RMSE, MAPE, and the R^2 coefficient of determination (Table 6).

Table 5: Correlation matrix of factors.

Indicator	Trains	Density	Length of the train	Load	Average delay
Trains	1.00	0.99	0.98	0.99	0.87
Density	0.99	1.00	0.97	0.99	0.86
Length of the train	0.98	0.97	1.00	0.98	0.82
Load	0.99	0.99	0.98	1.00	0.88
Average delay	0.87	0.86	0.82	0.88	1.00

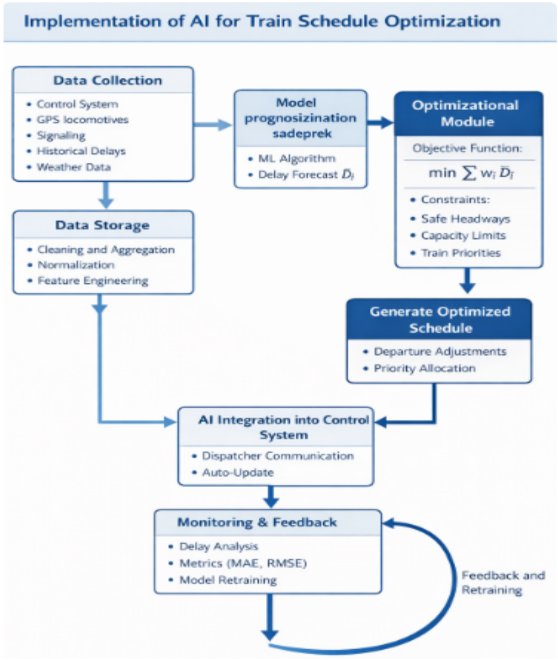

Figure 2: AI implementation model for train schedule optimization.

Table 4: A system of indicators for potential digitalization opportunities in the transport sector.

Group of indicators	Types, units of measurement	Potential opportunities
1. Infrastructure Digitilization	▪ Share of "smart" traffic control devices (traffic lights, switches, etc.), % of the total number. ▪ ITS (intelligent transport systems)coverage, % of the road network ▪ Traffic sensors (transport) with video surveillance, % ▪ CoatingGPS/GLONASS on rolling stock vehicles, % of the total number of vehicles ▪ Share of objects with digital twins, %	Implementing digital technologies (IoT, cloud computing, AI) to modernize the transportation system ensures automation, high-speed data exchange, reduced transportation costs, the availability and global reach of asset management systems.
2. Operational efficiency (before/after AI implementation)	▪ Average speed of movementkm/h (Δ, %); ▪ Average travel timemin (decrease, %); ▪ vehicle fleet utilization rate, %; ▪ Wagon/bus turnover, hour; ▪ Reduction in downtime, %; ▪ Traffic forecast accuracy, % (MAPE).	While the original goal of AI implementation was to increase business value, today's expectations for AI are primarily focused on operational efficiency, that is, proven impact and the expectation of clear and rapid returns.
3. Economic consequences	▪ Reduction in operating expenses (OPE), % ▪ Fuel/energy consumption, % or l/100 km ▪ Revenue growth due to digital services, % ▪ ROI of digital projects, % ▪ Payback period, years	Automate routine processes, reduce information processing time, reduce downtime and costs, and save resources.
4. Service quality and customer experience	▪ Electron fractionpercentage of registration of travel documents, % ▪ Share of non-cash payments, % ▪ Customer Satisfaction Score (CSI/NPS), points ▪ Share of queriesCustomers processed through digital channels, % ▪ Customer service response time, sec/min	Customer satisfaction (speed, competence, reliability), service quality, convenience, personalization of services.

Table 6: Performance metrics of the train delay prediction model.

Metric	Mathematical Formulation of Metrics	Value	Interpretation
MAE (Mean Absolute Error)	$MAE = \frac{1}{n}\sum_{i=1}^{n} \lvert y_i - \tilde{y}_i \rvert$	2.8 min	average absolute prediction error
RMSE (Root Mean Square Error)	$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \tilde{y}_i)^2}$	4.1 min	standard deviation of prediction errors
MAPE (Mean Absolute Percentage Error)	$MAPE = \frac{100}{n}\sum_{i=1}^{n} \left\lvert \frac{y_i - \tilde{\tilde{y}}_i}{y_i} \right\rvert$	8.7 %	average relative prediction error
R^2 (Coefficient of Determination)	$R^2 = 1 - \sum_{i=1}^{n} \frac{(y_i - \tilde{y}_i)^2}{(y_i - \bar{\tilde{y}}_i)^2}$	0.89	proportion of variance explained by the model

To validate predictive performance, the proposed model was compared with a baseline Linear
Results:

- MAE = 2.8 minutes, indicating an average forecast error of less than 3 minutes.
- RMSE = 4.1 minutes, indicating that large errors are rare.
- MAPE = 8.7%, indicating high forecast accuracy.
- R^2 = 0.89, indicating that the model explains 89% of the variation in train delays. The estimated economic effect associated with the reduction in train delay time after AI-based optimization is presented in Table 7.

Table 7: The economic impact is related to the reduction in delay time.

Indicator	Value
Average delay before AI implementation:	8.6 min
After optimization:	6,9 min
Reduction:	1,7 min
Number of trains per year:	54 242
Total delay reduction:	54242*1,7 = 92211.4 min
Cost per minute of delay:	799472 soums
Economic impact:	92211.4 *799472 =73720432.8 ≈ 73.72 billion soums per year

Model calculations show that the use of digital artificial intelligence technologies has reduced train delays from 8.6 minutes to 6.9 minutes, a reduction of 1.7 minutes. Here, 8.6 minutes corresponds to the observed 2024 average delay reported in Table 2, while 6.9 minutes represents a model-based projected/simulated delay under AI-assisted operating conditions rather than a second observed value; this distinction between observed baseline and model-projected outcome should be stated explicitly to avoid the impression that both figures were empirically measured. The potential impact of this factor alone was 73720.432 million soums (≈ 73.72 billion soums per year).

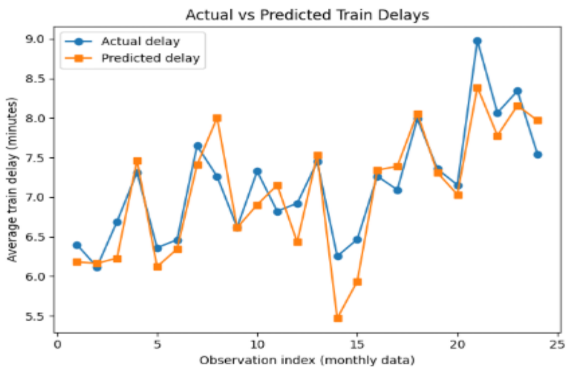

Figure 3: Comparison of actual and predicted train delays Source: Author's development.

Figure 3 compares actual and predicted train delays for the test subset. The close alignment between observed and predicted values indicates that the selected model reproduces the main variation pattern of delays with acceptable error, which supports its practical applicability for railway schedule optimization.

Machine-learning methods can be used to determine the contribution of individual factors to train delays (interpretation is shown in Table 8). Increasing these factors leads to an increase in the probability and duration of delays.

Feature importance analysis indicates that infrastructure load and train density are the most influential factors affecting train delays. Thus, the use of intelligent data analysis systems enables real-time management of rolling stock and minimizes downtime. This, in turn, leads to reduced operating costs and increased profitability for transport companies.

Table 8: Interpretation of train delay factors.

Factor	value	Influence
weather	0.18	adverse weather increases delays
time	0.09	peak hours increase delays
train length	0.12	Long trains require more time for operations
track load	0.34	High infrastructure load increases delays

5 DISCUSSION

The analysis confirms that the integration of digital technologies and artificial intelligence into the transport sector is not simply a tool for technological modernization, but a systemic factor in the industry's structural transformation. The findings align with the positions of international organizations, particularly the World Bank and the International Transport Forum, which view the digitalization of transport as a key element in enhancing the resilience and competitiveness of national economies.

Firstly, digital transformation contributes to increased operational efficiency for transport companies. The use of big data processing systems, intelligent demand forecasting, and automated transport management helps reduce costs, minimize downtime, and increase rolling stock utilization. This confirms the hypothesis of a direct correlation between the level of digitalization and the performance of the transport sector.

Secondly, the integration of AI increases market competition. Digital platforms ensure price transparency, accelerate interactions between supply chain participants, and expand private operators' access to infrastructure. However, this process requires improved regulatory frameworks, particularly in terms of digital platform regulation, data protection, and cyber security.

Thirdly, the discussion reveals a number of institutional and infrastructural limitations. These include:

- insufficient level of digital competence of personnel;
- fragmentation of information systems;
- high initial investment costs;
- the need to integrate national solutions with international standards.

For developing economies, including the Republic of Uzbekistan, a key challenge is balancing the pace of digitalization with infrastructure readiness. Without a comprehensive approach (investment, training, regulation, public-private partnerships), the impact of AI implementation may be limited.

At the same time, the discussion reveals a significant multiplier effect. Improving the efficiency of the transport industry leads to reduced logistics costs for companies, accelerated trade turnover, and increased export potential. Thus, the digitalization of transport acts as a catalyst for macroeconomic stability.

6 CONCLUSIONS

The transport sector in Uzbekistan is undergoing major transformations in terms of both economic efficiency and public convenience. Our study showed that the introduction of digital artificial intelligence technologies has reduced average train delays by 1.7 minutes. The potential economic impact of the time factor alone amounted to 73.72 billion soums per year.

The research has shown that harnessing the potential of intelligent data analysis systems enables real-time management of rolling stock and minimizes downtime. This, in turn, leads to reduced operating costs and increased profitability of transportation.

For the Republic of Uzbekistan, the widespread adoption of digital transport management platforms and intelligent logistics solutions can ensure sustainable growth in the industry and increase its contribution to the gross domestic product. The integration of digital technologies and artificial intelligence is becoming a system-forming factor in the transformation of the transport sector and a crucial condition for its long-term sustainable development.

REFERENCES

[1] Smith and P. Kumar, “The role of artificial intelligence in intelligent transport systems,” Journal of Transport Technologies, vol. 15, no. 4, pp. 225-245, 2022.

[2] G. Chen, S. Wang, and J. Li, “Machine learning approaches for traffic flow prediction in smart cities,” IEEE Transactions on Intelligent Transportation Systems, vol. 22, no. 8, pp. 4856-4865, 2021.

[3] J. Lee and D. Park, “Digital twin technologies in urban mobility: Concepts and applications,” Transportation Research Part C: Emerging Technologies, vol. 143, Art. no. 103966, 2023.

[4] S. Zhang and Y. Wang, “Networking technologies and applications for AI-enabled vehicles: A review,” IEEE Communications Surveys & Tutorials, vol. 24, no. 2, pp. 1234-1278, 2022.

[5] P. V. Metelkin and A. B. Filchikov, “The role of digital technologies in the integration and development of transport and logistics infrastructure,” Regional and Industrial Economics, no. 3, 2024, [Online]. Available: https://cyberleninka.ru/article/n/rol-tsifrovyh-tehnologiy-v-integratsii-i-razvitii-transportno-logisticheskoy-infrastruktury, [Accessed: Feb. 28, 2026].

[6] G. Yusupkhodzhaeva, “International experience in creating intelligent transport systems,” Interpretation and Research, vol. 1, no. 59, 2025, [Online]. Available: https://inlibrary.uz/index.php/interpretation-research/article/view/78852, [Accessed: Feb. 21, 2026].

[7] A. Kotenko, S. Sattorov, V. Negoroshkov, and K. Timukhin, “Intelligent information technologies and mathematical modeling,” Journal of Physics: Conference Series, vol. 2131, no. 3, 2021, [Online]. Available: https://doi.org/10.1088/1742-6596/2131/3/032102.

[8] T. Shermukhamedov and D. K. Abdullaeva, “Digitalization of logistics processes,” In Science, part 1, pp. 122-124, 2021.

[9] A. Anikin and A. P. Tyapukhin, Commercial Logistics. Moscow, Russia: Prospect, 2017.

[10] S. B. Sattorov and A. G. Kotenko, “Organization of transportation of local goods by railways of the Republic of Uzbekistan in the context of increasing transit cargo transportation,” in Transport of Russia: Problems and Prospects. St. Petersburg, Russia: IPTRAS, 2021, pp. 127-132.

[11] M. Tokareva and A. Mavlanov, “Digitalization of railway transport,” Central Asian Academic Scientific Journal, vol. 2, no. 6, pp. 102-109, 2022.

[12] K. Merenkov, “Industry 4.0: German experience in developing digital transport and logistics solutions,” Control, no. 4(18), pp. 17-21, 2017.

[13] A. S. Nabiullin and R. S. Zaripova, “Application of blockchain technology for cargo transportation management,” Science of Krasnoyarsk Krai, vol. 9, no. 4-2, pp. 106-110, 2020.

[14] O. I. Dontsova, “Infrastructure support for the digital transformation of industrial clusters,” Creative Economy, no. 4, 2022, [Online]. Available: https://cyberleninka.ru/article/n/infrastrukturnaya-podderzhka-tsifrovoy-transformatsii-promyshlennyh-klasterov.

[15] Gartner, “Key Strategic Technology Trends,” 2023, [Online]. Available: https://www.gartner.com, [Accessed: Feb. 21, 2026].

[16] Government of Uzbekistan, Concept for the Development of the Transport and Logistics System of the Republic of Uzbekistan until 2030, Presidential Resolution, Jan. 27, 2025, [Online]. Available: https://www.lex.uz/uz/docs/7342176, [Accessed: Feb. 21, 2026].