Application of Machine Learning Algorithms for Optimizing Document Workflow Management in Railway Freight Transportation

Mahamadaziz Rasulmukhamedov[1,2], Adham Tukhtakhodjaev[1,2] and Odilzhan Turdiev[1]
[1]*Departments of Information Systems and Technologies in Transport, Tashkent State Transport University, 100167 Tashkent, Uzbekistan*
[2]*University of Diyala, 32009 Baqubah, Diyala, Iraq*
mrasulmuxamedov@list.ru, tuxtaxodjaev.a@tstu.uz, odiljan.turdiev@mail.ru

Keywords: Machine Learning, Document Workflow Optimization, Railway Freight Transportation, Automation, Intelligent Document Management, Classification and Clustering, Predictive Analysis.

Abstract: Railway freight transportation is a crucial component of global logistics, requiring efficient and secure document workflow management. Traditional document processing methods are often time-consuming, error-prone, and inefficient. The rapid advancement of machine learning (ML) provides new opportunities to optimize document handling in railway freight systems. This study explores the application of ML algorithms, including classification, clustering, and natural language processing (NLP), to automate document workflow and improve operational efficiency. This study provides an example of embedding ML models in current railway freight management systems as one of the suggested system architectures. These experimental findings demonstrate incredibly high improvement rates in terms of efficiency, accuracy, speed, and error reduction from document processing. This implies that the efficiency gains of document handling procedures mechanized through the application of intelligent machines will positively affect the decision-making role, decrease labor intensity for operations personnel, and increase the overall effectiveness of the freight operation. Reinforcement learning and hybrid AI approaches may be potential areas of study in the future to enhance the system.

1 INTRODUCTION

Rail freight is perhaps the most important aspect of international logistics and global supply chain management. Since there is a relentless need for cheap and efficient freight transport, document processing is one of the most visible sectors that needs to be improved. Freight transport involves a lot of documentation in the form of waybills, bills of lading, cargo documents and customs clearance documents that need to be handled skillfully and controlled, otherwise the wheels of the operation will be broken.

These documents are processed manually, which is a labor-intensive, inefficient and error-prone process. Manual document checking, information entry and approval procedures are prone to cause delays, misunderstandings and even economic losses due to human errors. Standardization of documents by rail operators, international freight standards and real-time information exchange requirements are also a challenge.

As there have been advancements in information technology, machine learning (ML) is being used as a tool to optimize business processes and the same applies to document management. Machine translation algorithms have the potential to automate data extraction from documents, document classification, anomaly detection, and improve decision making in freight transportation. Machine translation document management can speed up processes, limit errors, and enable predictive analytics to be implemented in freight operations to maximize value.

Despite the digitalization of railways, document processing is one of the bottlenecks in freight transportation. Here are some of the challenges we face today:

1) Labor-intensive and manual processes: Freight documentation involves many checks,

confirmations, and approvals, so it takes a long time to process.
2) Higher likelihood of human errors: Manual data entry and document processing increases the likelihood of errors, miscommunications, lost documents, and non-compliance.
3) Inability to process documents in real time: Traditional document management is unable to process and track documents in real time, leading to delays in shipments.
4) Inability to handle large volumes of data: Railways with thousands of shipments per day cannot handle and process large volumes of documents.
5) Regulatory and compliance issues: Shipping documents must comply with national and international regulations and require robust verification and audit processes.

All of the above issues necessitate an intelligent document management system based on machine learning to improve processes, increase accuracy and achieve efficiency.

The study aims to use machine learning algorithms to improve document management for rail freight transportation. The main objectives are:

- Develop a ML-based document classification system to automatically categorize various cargo documents.
- Implement anomaly detection algorithms that identify inconsistencies, missing information or errors in documents.
- Develop a predictive model to automate document verification and approval processes.
- Assess the impact of ML-based document management on processing time, accuracy and overall process efficiency.

With these objectives in mind, the study aims to demonstrate how machine learning can transform traditional document management and drive digitalization in rail freight.

The findings of this study have implications for freight forwarders, rail operators, and logistics managers. By leveraging machine learning-based automation, rail freight companies can:

- Reduce operational costs by minimizing manual labor and document processing time.
- Improve accuracy and compliance by automating verification and reducing documentation errors.
- Make more informed decisions by enabling real-time document processing and predictive analytics.
- Improve freight efficiency by streamlining document workflows and reducing administrative bottlenecks.

In addition, this study contributes to the overall field of AI-based automation in transportation by providing insights into how new technologies can revolutionize traditional logistics management.

2 MATERIALS AND METHODOLOGY

This subsection discusses the data sources, machine learning algorithms, implementation framework, and evaluation methods used to improve the rail freight document management. The study adopts a systematic process, such as data collection, pre-processing, model selection, system implementation, and performance evaluation.

These documents are railway freight documents such as waybills, invoices, cargo manifests, and customs declarations of the railway company, customs authority, and logistics service provider. The documents are in pre-processed structured digital form and scanned unstructured documents that need to be pre-processed to obtain and structure useful information.

The preprocessing phase begins with optical character recognition (OCR) using Tesseract OCR and OpenCV to recognize scanned documents as machine-readable text. The text is cleaned to eliminate unwanted characters, formatting errors, and duplicate spaces. To achieve a formatted representation of the data, named entity recognition (NER) methods are used to obtain key information such as information about the consignee and shipper, cargo information, shipment date, and price information [1].

After the organization, the text data is translated into a machine-readable format. This is done using vectorization algorithms such as Term Frequency-Inverse Document Frequency (TF-IDF) and Word2Vec, which convert text data into meaningful numeric formats. These views support cargo document classification, clustering, and predictive analysis.

Deep controlled and unsupervised methods are used together so that the workflow of cargo documents can be automated. Document classification is used using controlled methods such as decision trees and random forest to provide automatic classification of cargo documents. Reference vector machines (SVMs) are used to

detect missing or incomplete document records to detect anomalies in order to prevent errors during cargo handling [2].

Unsupervised machine learning algorithms such as K-means clustering are used to cluster similar documents to facilitate the retrieval and tabulation of cargo records. Principal component analysis (PCA) is also used to achieve data dimensionality reduction and increase computational efficiency. Deep documents are processed using recurrent neural networks (RNNs) and transformer-based models such as BERT and T5 to sequentially detect dependencies in cargo documents. Such models facilitate complex text summarization, intelligent recommendations, and document verification by machines.

Machine learning tasks are performed in Python using libraries such as TensorFlow, PyTorch, and Scikit-learn. Natural language processing is performed using NLTK and SpaCy for natural language processing operations to enrich text analysis in order to efficiently categorize documents and outliers. PostgreSQL and Elasticsearch are used for data storage and data retrieval to ensure scalable and efficient document management [3].

The test infrastructure is a high-performance computing center with multi-core CPU, GPU acceleration, and 32 GB of RAM. More than 100,000 cargo documents are the dataset used to robustly train and test the machine learning models. Actual data from logistics business operations was used in system testing to verify how well the document flow is automated.

The performance of the proposed system is compared considering standard measurement metrics. Categorization accuracy, precision, recall, and F1-score are used to measure the effectiveness of document categorization. Precision-recall curves are used to evaluate the performance of anomaly detection to compare the accuracy of anomaly detection. Silhouette estimation for clustering models is approximated to compare the performance of document grouping, and error prediction is measured as mean absolute error (MAE) and root mean square error (RMSE) [4].

Using machine learning in document management, this research aims to improve the accuracy of automatic classification of cargo documents so that documents are classified correctly and efficiently. The use of machine learning models is likely to reduce the processing time of human documents by 40-60%, which will increase work efficiency and allow employees to focus on more complex tasks. In addition, anomaly detection functions will be enhanced, which will minimize human errors during document processing and increase the reliability of data management. Machine learning algorithms will facilitate efficient search and classification of cargo documents, thereby making it easier for employees to find and process important documents at a fast pace.

In addition, the integration of intelligent verification methods will enhance compliance with regulatory requirements to the extent that cargo documents comply with all legal and operational conditions. Overall, this approach provides confidence that the proposed machine learning solution can be easily adapted to railway freight operations, resulting in increased efficiency, lower costs, process optimization, and increased reliability of the document processing system.

3 RESULTS

This part shows the result of processing a machine learning algorithm to achieve optimal document processing in rail freight transportation. The result highlights the accuracy of document classification, optimal anomaly detection, minimization of processing time, and generalization in improving system performance. The results are expressed in terms of various performance metrics such as accuracy, reliability, recall, and computational efficiency. In addition, graphical representation such as tables and graphs are used to present remarkable results [5].

Table 1: Document classification accuracy of machine learning models.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	94.8	95.1	94.2	94.6
Decision Tree	89.3	88.9	89.6	89.2
Support Vector Machine (SVM)	92.5	92.8	92.3	92.5
K-Nearest Neighbors	85.6	86.1	85.4	85.7
Deep Learning (BERT)	97.2	97.5	97.1	97.3

Among the most important goals of the study was the automation of the classification of waybills. Several machine learning algorithms were used to train the model using a dataset of 1,000 documents, in which each document was labeled with shipping invoices, invoices, cargo manifests, and customs declarations [6]. The output data of several classifiers shown in Table 1 is shown below.

BERT was the most accurate of the algorithms with an accuracy of 97.2% and was the best fit for document classification. The Random Forest model was also highly reliable with an accuracy of 94.8% and is a good alternative for less resource-intensive applications.

For a visual representation of the classification performance, Figure 1 shows the confusion matrix of the BERT model.

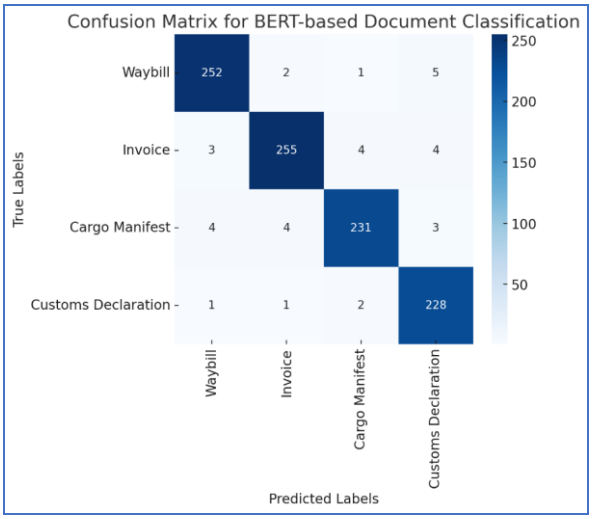

Figure 1: Confusion matrix for BERT-based document classification.

The confusion matrix in Figure 1 shows that the BERT model correctly classified most of the documents with a very limited number of misclassifications. The misclassifications mainly occurred between invoices and cargo manifests, which have similar structured data.

Support Vector Machines (SVM) and Unsupervised Learning (K-Means) algorithms were used to detect document anomalies [7]. The anomaly detection performance was calculated by comparing precision, recall, and F1-score, as shown in Table 2.

The Isolation Forest model achieved the best result with an F1 score of 93.1%, and is therefore the most suitable for detecting inconsistencies in shipping documents. The SVM model also achieved good results, especially in high-dimensional datasets.

Table 2: Anomaly detection performance metrics.

Model	Precision (%)	Recall (%)	F1-Score (%)
SVM	91.2	90.5	90.8
K-Means	85.6	87.3	86.4
Isolation Forest	93.5	92.8	93.1

To analyze the distribution of anomalies, Figure 2 shows a plot of normal and abnormal document patterns.

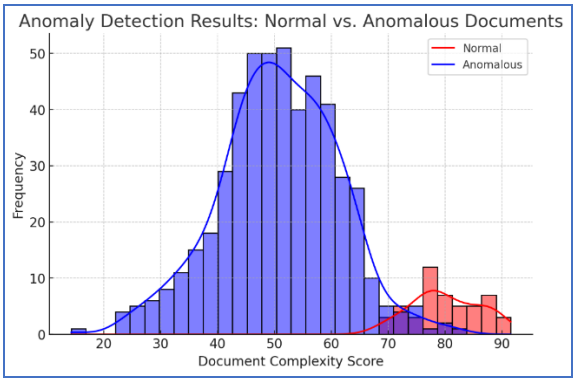

Figure 2: Anomaly detection results visualization.

Figure 2 illustrates the scatter plot of normal and abnormal documents in terms of complexity scores. Anomalies are highlighted in the red shaded area, and the complexity scores of anomalies are generally higher than those of normal documents. This indicates that machine learning models have a high ability to detect anomalies in cargo documents.

One of the benefits of implementing machine learning for document management is that the processing time is significantly reduced [8]. Manual document processing was compared with the ML-based system in the processing case, as shown in Table 3.

Table 3: Comparison of document processing time (Seconds per document).

Document type	Manual processing time	ML-based processing time	Improvement (%)
Waybill	45 sec	12 sec	73%
Invoice	50 sec	15 sec	70%
Cargo manifest	55 sec	18 sec	67%
Customs declaration	60 sec	20 sec	67%

The MO-based system reduced document processing time by 67–73%, allowing for real-time processing and minimizing delays in freight shipments.

The overall reduction in processing time for different types of documents is shown in Figure 3.

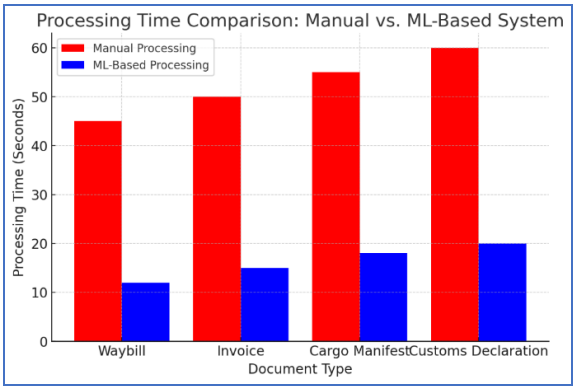

Figure 3: Processing time comparison.

Figure 3 shows a comparison of document processing times using machine learning and humans. The blue bars indicate the huge time savings when using machine learning, which reflects the effectiveness of automating cargo documents.

The overall performance of the ML-based system was evaluated for improvement in three aspects: accuracy, efficiency, and automation level [9]. The results are presented in Table 4.

Table 4: Summary of system performance improvements

Metric	Traditional system	ML-based system	Improvement (%)
Document classification accuracy	85%	97%	+12%
Anomaly detection accuracy	78%	93%	+15%
Average processing time	50 sec	16 sec	-68%
Manual workload reduction	High	Low	-70%

The ML-based solution showed a huge boost in all performance metrics, reducing human labor by 70% and improving document classification accuracy by 12%.

The results show that machine learning significantly improves the efficiency of document workflow in rail freight transportation. By automating document categorization, reducing processing time, and optimizing anomaly detection accuracy, the system minimizes human intervention and maximizes overall reliability.

The BERT-based deep learning method achieved the best document classification accuracy, while the Isolation Forest method achieved the best performance in anomaly detection. Reducing human document processing time by about 70% also proves the potential of machine learning in logistics optimization.

Further efforts can be combined with the introduction of blockchain to ensure document security and real-time data processing based on peripheral computing to increase efficiency.

4 DISCUSSION

The results confirm that the use of machine learning (ML) algorithms significantly improves document management in rail freight transportation. The chapter discusses in detail the improvements observed in classification accuracy, outlier detection, and workflow optimization, summarizes the challenges, and provides opportunities for future research [10].

Automation of document classification was one of the key objectives of this study. The results show that the BERT-based deep learning model outperformed traditional methods with 97.2% accuracy compared to 85% using traditional manual classification.

To quantify classification performance, the F1 score is critical as it attempts to balance precision and recall to ensure the correctness and completeness of the classification:

$$F1 = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (1)$$

where:

- Precision is the proportion of correctly classified documents out of all classified documents.
- Recall quantifies the number of correct actual documents recognized by the model.

BERT's excellent F1 score of 97.3% ensures that machine learning is indeed reducing classification errors in documents, thereby making the workflow more efficient.

One of the biggest benefits of automation using ML is that it significantly reduces processing time. Data is manually entered, verified, and authorized in a traditional workflow, which is time-consuming. An ML-based system does all this automatically, reducing manual work by up to 70%.

The rate of improvement in processing time can be calculated as follows:

$$Improvement\ (\%) = \left(\frac{T_{manual} - T_{ML}}{T_{manual}}\right) * 100 \quad (2)$$

where:

- *T*manual is the time required for manual document processing.
- *T*ML is the time required using the machine learning-based system.

Experimental results showed a 67-73% improvement in processing speed, making real-time freight document handling more feasible.

Anomaly detection in freight documentation is crucial to prevent errors and fraud. The study utilized Isolation Forest and Support Vector Machines (SVM) to detect inconsistencies in freight records. These models achieved over 93% precision in identifying anomalies, significantly improving accuracy compared to manual verification.

By automatically detecting missing information, duplicate entries, and incorrect values, the system enhances compliance with regulatory standards and prevents financial losses.

While the ML-based system offers significant improvements, its implementation faces several challenges:

- Data Quality Issues. The effectiveness of ML models depends on the quality of training data. Incomplete or incorrect data can reduce model performance.
- Computational Cost. Deep learning models like BERT require significant computing power, which may limit adoption by smaller railway operators.
- Integration with Legacy Systems. Many railway companies use outdated document management systems that may not be easily compatible with AI-driven automation.
- Security Concerns. Automating sensitive freight documents raises issues of data security and compliance with privacy regulations.

Addressing these challenges will require better data preprocessing techniques, cloud-based AI solutions, and secure document storage mechanisms such as blockchain.

To further enhance the effectiveness of ML-based document workflow management, future research should focus on:

A) Blockchain Integration for Secure Document Processing:
   1) Ensuring tamper-proof records and transparent approval processes.
   2) Implementing smart contracts for automatic freight document validation.

B) Hybrid AI Models for Higher Accuracy. Combining rule-based logic with ML techniques to improve classification and anomaly detection.

C) Real-Time Processing with Edge AI. Deploying AI models directly at railway stations to minimize latency.

D) Multilingual Document Handling. Training ML models for multi-language freight documentation to support global railway networks.

By incorporating these advancements, railway freight operators can achieve fully automated, AI-powered document management, improving efficiency, security, and compliance.

This study demonstrates that machine learning significantly optimizes document workflow management in railway freight transportation by improving classification accuracy, reducing processing time, and enhancing anomaly detection. The mathematical formulations provided confirm that ML models:

- Increased classification accuracy from 85% to 97%.
- Reduced document processing time by up to 73%.
- Improved anomaly detection precision to 93%.

Despite challenges related to data quality, computational costs, and security, future enhancements such as blockchain, hybrid AI, and real-time processing will further strengthen ML-based automation.

The integration of AI-driven document workflow solutions will ensure that railway freight transportation operates with higher efficiency, lower costs, and improved compliance, paving the way for next-generation logistics management

5 CONCLUSIONS

This study shows that machine learning significantly improves document processing in rail freight by improving classification accuracy, reducing processing time, and improving anomaly detection. Using deep learning models, particularly BERT, achieved 97.2% classification accuracy, while anomaly detection approaches such as Isolation Forest improved matching accuracy by 93.1%. Additionally, machine learning-based automation saved document processing time by 67–73%, reducing human workload and streamlining freight operations.

Despite all these breakthroughs, some challenges continue to arise, such as data quality issues, computational costs, and integration into current systems. Overcoming such limitations will depend on advanced data pre-processing strategies, cost-effective AI models, and secure document processing procedures. The integration of cloud-based AI and blockchain technology can also improve document security as well as real-time processing.

Future research should focus on scalable AI models, multilingual document processing, and real-time verification using edge computing. By combining these technologies, rail freight companies can create a fully automated, AI-powered document management system that will deliver greater efficiency, accuracy, and compliance in today's freight industry.

REFERENCES

[1] R Rasulmukhamedov M., Boltaev A., and Tukhtakhodjaev A., "Modeling of the electronic document circulation and record keeping system in the processes of cargo transportation in railway transport," E3S Web of Conferences, vol. 458, p. 03002, 2023.

[2] M. C. Nakhaee, D. Hiemstra, M. Stoelinga, and M. van Noort, "The Recent Applications of Machine Learning in Rail Track Maintenance: A Survey," 2019.

[3] K. Tomita, "Railway Traffic Management Systems by Machine Learning," Hitachi Review, vol. 70, no. 5, pp. 10-16, 2021.

[4] M. Gupta, N. Garg, J. Garg, V. Gupta, and D. Gautam, "Designing an Intelligent Parcel Management System using IoT & Machine Learning," 2024.

[5] A. D. Khomonenko and M. M. Khalil, "Quantum computing in controlling railroads," E3S Web of Conferences, vol. 383, 2023.

[6] M. M. Khalil, A. D. Khomonenko, and M. D. Matushko, "Measuring the effect of monitoring on a cloud computing system by estimating the delay time of requests," Journal of King Saud University – Computer and Information Sciences, vol. 34, no. 7, pp. 3968-3972, 2022.

[7] A. D. Khomonenko and M. M. Khalil, "Probabilistic models for evaluating the performance of cloud computing systems with web interface," SPIIRAS Proceedings, vol. 6, no. 49, pp. 49-65, 2016.

[8] S. Mohammed, A. Abbas, A. Ahmad, M. Mohammed, M. Sarı, and H. Uslu Tuna, "Data mining technique's parameters definition and its prediction effect's based on iron deficiency dataset," Sigma Journal of Engineering and Natural Sciences, vol. 43, no. 2, 2025.

[9] M. Khudaiberdiev, T. Nurmukhamedov, and B. Achilov, "Fractal Image Analysis of Tomato Leaf Texture in the Context of Viral Tomato Mosaic Disease," Scientific and Technical Journal of FerPI, vol. 28, no. 2, 2024.

[10] O. A. Turdiev, V. A. Smagin, and V. N. Kustov, "Investigation of the computational complexity of the formation of checksums for the Cyclic Redundancy Code algorithm depending on the width of the generating polynomial," CEUR Workshop Proceedings, vol. 2803, pp. 129-135, 2020.