ISSN: 2224-5227 (Print)
ISSN: 2518-1483 (Online)

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

№1
2026

ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

CENTRAL ASIAN ACADEMIC
RESEARCH CENTER

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE

1 (357)

JANUARY – MARCH 2026

PUBLISHED SINCE JANUARY 1963
PUBLISHED 4 TIMES A YEAR

ALMATY, NAS RK

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)
Owner: «Central Asian Academic Research Center» LLP (Almaty).
Certificate **№ KZ77VPY00121154** on the re-registration of the periodical printed and online publication of the information agency, issued on **05.06.2025** by the Republican State Institution «Information Committee» of the Ministry of Culture and Information of the Republic of Kazakhstan
Subject area*: information and communication technologies.*
Currently: *included in the list of journals recommended by the CCSES MSHE RK in the direction of «Information and communication technologies».*
Periodicity: *4 times a year.*
http://www.physico-mathematical.kz/index.php/en/

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Меншіктеуші: «Орталық Азия академиялық ғылыми орталығы» ЖШС (Алматы).

Ақпарат агенттігінің мерзімді баспасөз басылымын, ақпарат агенттігін және желілік басылымды қайта есепке қою туралы ҚР Мәдениет және Ақпарат министрлігі «Ақпарат комитеті» Республикалық мемлекеттік мекемесі **05.06.2025** ж. берген **№ KZ77VPY00121154** Куәлік.

Тақырыптық бағыты: *ақпараттық-коммуникациялық технологиялар*

Қазіргі уақытта: *«ақпараттық-коммуникациялық технологиялар» бағыты бойынша ҚР БҒМ БҒСБК ұсынған журналдар тізіміне енді.*

Мерзімділігі: *жылына 4 рет.*

http://www.physico-mathematical.kz/index.php/en/

Academic Scientific Journal of Computer Science
ISSN 2518-1726 (Online),
ISSN 1991-346X (Print)

Собственник: *ТОО «Центрально-азиатский академический научный центр» (г. Алматы).*

Свидетельство о постановке на переучет периодического печатного издания, информационного агентства и сетевого издания **№ KZ77VPY00121154.** Дата выдачи **05.06.2025**

Тематическая направленность: *информационно-коммуникационные технологии.*

В настоящее время: *вошел в список журналов, рекомендованных КОКСНВО МНВО РК по направлению «информационно-коммуникационные технологии».*

Периодичность: *4 раза в год.*

http://www.physico-mathematical.kz/index.php/en/

CONTENTS

COMPUTER SCIENCE

МАЗМҰНЫ

ИНФОРМАТИКА

СОДЕРЖАНИЕ

ИНФОРМАТИКА

ACADEMIC SCIENTIFIC JOURNAL OF COMPUTER SCIENCE
ISSN 1991-346X
Volume 1.
Namber 357 (2026). 253–270

https://doi.org/10.32014/2026.2518-1726.412

IRSTI 77.03.09
UDC 621.7.08

© Mimenbayeva A.B.[1], Turebayeva R.D.[2], Ospanova T.T.[2]*, Aruova A.B.[1], Naizagarayeva A.A.[3], 2026.
[1]Astana IT University, Astana, Kazakhstan;
[2]L.N. Gumilyov National Eurasian University, Astana, Kazakhstan;
[3]NCJSC «S. Seifullin Kazakh AgroTechnical Research University», Astana, Kazakhstan.
E-mail: Tleu2009@mail.ru

DEVELOPMENT AND COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR URBAN TRAFFIC PREDICTION

Mimenbayeva Aigul — Senior Lecturer, Astana IT University, Astana, Kazakhstan,
E-mail: A.Mimenbayeva@astanait.edu.kz, https://orcid.org/0000-0003-4652-470X;
Turebayeva Rakhila — Senior Lecturer, L.N. Gumilyov National Eurasian University, Astana, Kazakhstan,
E-mail: turebayeva_rd@enu.kz, https://orcid.org/0009-0006-4530-1300;
Ospanova Tleugaisha — Associate Professor, L.N. Gumilyov National Eurasian University, Astana, Kazakhstan,
E-mail: Tleu2009@mail.ru, https://orcid.org/0000-0002-1729-1321;
Aruova Aliya — Associate Professor, Astana IT University, Astana, Kazakhstan,
E-mail: a.aruova@astanait.edu.kz, https://orcid.org/0000-0001-7477-3856;
Naizagarayeva Akgul — Senior Lecturer, NCJSC «S.Seifullin Kazakh AgroTechnical Research University», Astana, Kazakhstan,
E-mail: a.naizagaraeva@kazatu.edu.kz, orcid.org/0000-0002-3616-0925.

Abstract. Urban traffic congestion is a major issue in modern cities. Rapid urbanization and increasing vehicle numbers intensify this problem. Accurate traffic prediction is essential for reducing congestion. It also improves road safety and supports intelligent transportation systems. The study aims to develop and compare machine learning models for regulating urban traffic. In addition, it aims to identify the most effective approach for predicting road traffic in advance. To achieve this goal, the study involves several tasks, including data collection and preprocessing, exploratory data analysis, feature engineering, and data normalization. Additionally, three machine learning algorithms—logistic regression, support vector machine (SVM), and Random Forest—were implemented and evaluated based on performance metrics such as accuracy, precision, recall, and

F1-score. The results of the study demonstrate that all models provide satisfactory performance; however, their effectiveness varies across traffic classes. Logistic Regression achieved an accuracy of 0.87 but showed weaker performance in the "low" traffic class. The SVC model improved classification balance and reached an accuracy of 0.88. The best performance was achieved by the Random Forest model, which demonstrated an overall accuracy of 0.99 and a weighted F1-score of 0.99, indicating high reliability and strong predictive capability across all traffic categories. The practical significance of the study lies in the potential application of the proposed model in real-time traffic management systems. The integration of machine learning models, particularly ensemble methods such as Random Forest, can support decision-making processes, optimize traffic flow, reduce congestion, and improve road safety. Furthermore, the developed approach can contribute to the advancement of intelligent transportation systems and serve as a foundation for future research involving real-time data and advanced predictive models. The findings indicate that ensemble learning methods, particularly Random Forest, provide superior accuracy and reliability for urban traffic prediction tasks. The proposed model can be integrated into real-time traffic management systems to enhance decision-making and improve urban mobility.

Keywords: Machine learning, Random forest, Support Vector Machine, Logistic Regression, Traffic classification, Urban traffic management

For citations: Mimenbayeva A.B., Turebayeva R. D., Ospanova T.T., Aruova A.B., Naizagarayeva A.A. Development and comparative analysis of machine learning models for urban traffic prediction. Academic Scientific Journal of Computer Science, 2026. — No.1. – P. 253–270. DOI: https://doi.org/10.32014/2026.2518-1726.412

© Мименбаева А.Б.[1], Туребаева Р.Д.[2], Оспанова Т.Т.[2]*, Аруова А.Б.[1] , Найзағараева А.А.[3], 2026.

[1]Astana IT University, Астана, Қазақстан;
[2]Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан;
[3]«С. Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті» КеАҚ, Астана, Қазақстан.
E-mail: Tleu2009 @mail.ru

ҚАЛАЛЫҚ КӨЛІК АҒЫНЫН БОЛЖАУҒА АРНАЛҒАН МАШИНАЛЫҚ ОҚЫТУ МОДЕЛЬДЕРІН ӘЗІРЛЕУ ЖӘНЕ САЛЫСТЫРМАЛЫ ТАЛДАУ

Мименбаева Айгүл — Сеньор-лектор, Astana IT University, Астана, Қазақстан,
E-mail: A.Mimenbayeva@astanait.edu.kz, https://orcid.org/0000-0003-4652-470X;
Туребаева Рахиля — техника ғылымдарының кандидаты, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: turebayeva_rd@enu.kz, https://orcid.org/0009-0006-4530-1300;

Оспанова Тлеуғайша — қауымдастырылған профессор, Л.Н. Гумилев атындағы Еуразия ұлттық университеті, Астана, Қазақстан,
E-mail: Tleu2009@mail.ru, https://orcid.org/0000-0002-1729-1321;
Аруова Алия — қауымдастырылған профессор, Astana IT University, Астана, Қазақстан,
E-mail: a.aruova@astanait.edu.kz, https://orcid.org/0000-0001-7477-3856;
Найзағараева Акгүл — PhD, аға оқытушы, «С. Сейфуллин атындағы Қазақ агротехникалық зерттеу университеті» КеАҚ, Астана, Қазақстан,
E-mail: a.naizagaraeva@kazatu.edu.kz, orcid.org/0000-0002-3616-0925.

Аннотация. Қазіргі заманғы қалаларда қалалық көлік кептелісі – басты мәселелердің бірі болып табылады. Жедел урбанизация және көліктер санының артуы бұл мәселенің артуына әкелуде. Кептелісті азайту үшін көліктің қозғалысын дәл болжау маңызды. Сонымен қатар, бұл жол қауіпсіздігін жақсартады және ақылды көлік жүйелерін қолдайды. Зерттеудің мақсаты – қалалық көлік қозғалысын реттеуге арналған машинамен оқыту модельдерін әзірлеу және салыстыру. Сонымен қатар, зерттеу жол қозғалысын алдын ала болжауда ең тиімді тәсілді анықтауды мақсат етеді. Осы мақсатқа жету үшін зерттеу деректерді жинау және алдын ала өңдеу, зерттеушілік деректер талдауы, ерекшеліктерді инженерлеу және деректерді нормалау сияқты бірнеше тапсырманы қамтиды. Сонымен қатар, логистикалық регрессия, қолдау векторлық машинасы (SVM) және Random Forest сияқты үш машинамен оқыту алгоритмі іске асырылып, дәлдік, нақтылық, қайта табу көрсеткіші және F1-шот сияқты өнімділік көрсеткіштері бойынша бағаланды. Зерттеу нәтижелері барлық модельдердің қанағаттанарлық өнімділік көрсеткенін көрсетеді; алайда олардың тиімділігі жол қозғалысы кластары бойынша әртүрлі. Логистикалық регрессия 0,87 дәлдікке жетті, бірақ «төмен» жол қозғалысы класында әлсіз өнімділік көрсетті. SVC моделі жіктеу тепе-теңдігін жақсартып, 0.88 дәлдікке жетті. Ең жақсы нәтиже Random Forest моделінде тіркелді: ол жалпы 0.99 дәлдік пен 0.99 салмақталған F1-балл көрсетті, бұл барлық жол қозғалысы категориялары бойынша жоғары сенімділік пен мықты болжау қабілетін көрсетеді. Зерттеудің практикалық маңызы ұсынылған модельді нақты уақыттағы жол қозғалысын басқару жүйелерінде қолдану мүмкіндігінде жатыр. Машинамен оқыту модельдерін, әсіресе Random Forest сияқты ансамбль әдістерін біріктіру шешім қабылдау процестерін қолдап, қозғалыс ағынын оңтайландырып, кептелісті азайтып, жол қауіпсіздігін арттыруға мүмкіндік береді. Сонымен қатар, әзірленген тәсіл интеллектуалды көлік жүйелерін дамытуға үлес қосып, нақты уақыттағы деректер мен жетілдірілген болжамдық модельдерді қамтитын болашақ зерттеулер үшін негіз бола алады.

Түйін сөздер: машиналық оқыту, кездейсоқ орман, тірек векторлар әдісі, логистикалық регрессия, трафикті жіктеу, қалалық трафикті басқару

© Мименбаева А.Б.[1], Туребаева Р.Д.[2], Оспанова Т.Т.[3]*, Аруова А.Б.[1], Найзағараева А.А.[3], 2026.

[1]Astana IT University, Астана, Казахстан;

[2] Евразийский национальный университет им. Л.Н. Гумилёва, Астана, Казахстан;

[3] НАО «Казахский агротехнический университет им. С. Сейфуллина», Астана, Казахстан.

E-mail: Tleu2009@mail.ru

РАЗРАБОТКА И СРАВНИТЕЛЬНЫЙ АНАЛИЗ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ДЛЯ ПРОГНОЗИРОВАНИЯ ГОРОДСКОГО ТРАФИКА

Мименбаева Айгуль — Сеньор-лектор, Astana IT University, Астана, Казахстан,
E-mail: A.Mimenbayeva@astanait.edu.kz, https://orcid.org/0000-0003-4652-470X;
Туребаева Рахиля — кандидат технических наук, Евразийский национальный университет им. Л.Н. Гумилёва, Астана, Казахстан,
E-mail: turebayeva_rd@enu.kz, https://orcid.org/0009-0006-4530-1300;
Оспанова Тлеугайша — ассоциированный профессор, Евразийский национальный университет им. Л.Н. Гумилёва, Астана, Казахстан,
E-mail: Tleu2009@mail.ru, https://orcid.org/0000-0002-1729-1321;
Аруова Алия — ассоциированный профессор, Astana IT University, Астана, Казахстан,
E-mail: a.aruova@astanait.edu.kz, https://orcid.org/0000-0001-7477-3856;
Найзагараева Акгул — PhD, старший преподаватель, НАО «Казахский агротехнический университет им. С. Сейфуллина», Астана, Казахстан,
E-mail: a.naizagaraeva@kazatu.edu.kz, https://orcid.org/0000-0002-3616-0925.

Аннотация. Транспортная загруженность является одной из ключевых проблем современных городов. Быстрая урбанизация и рост числа транспортных средств существенно усиливают нагрузку на дорожную инфраструктуру. Точное прогнозирование транспортных потоков играет важную роль в снижении заторов, повышении безопасности дорожного движения и развитии интеллектуальных транспортных систем. Целью исследования является разработка и сравнительный анализ моделей машинного обучения для прогнозирования городского трафика и определения наиболее эффективного подхода к его раннему прогнозированию. Для достижения поставленной цели выполнены этапы сбора и предварительной обработки данных, разведочного анализа (EDA), инженерии признаков и нормализации данных. В рамках исследования реализованы и сравнены три алгоритма машинного обучения: логистическая регрессия, метод опорных векторов (SVM) и Random Forest. Оценка эффективности моделей проводилась с использованием стандартных метрик классификации: accuracy, precision, recall и F1-score. Результаты показали, что все модели демонстрируют удовлетворительную производительность, однако их эффективность различается в зависимости от уровня интенсивности трафика. Логистическая регрессия достигла точности 0,87, однако показала более слабые результаты для класса с низкой интенсивностью движения. Модель SVM улучшила баланс классификации и достигла точности 0,88. Наилучшие результаты продемонстрировала модель Random Forest с общей точностью 0,99 и взвешенным F1-score 0,99, что свидетельствует о высокой надёжности и способности модели эффективно прогнозировать трафик во всех категориях. Полученные результаты подтверждают, что методы ансамблевого обучения, в частности Random Forest, обеспечивают более высокую точность и устойчивость прогнозирования городского трафика. Предложенная модель может быть интегрирована в системы управления дорожным движением в режиме реального времени для оптимизации принятия решений и повышения мобильности в городах.

Ключевые слова: машинное обучение, случайный лес, метод опорных векторов, логистическая регрессия, классификация трафиков, управление городским движением

Introduction. Nowadays, due to the increasing number of vehicles and urbanization, traffic congestion has become widespread in cities around the world. Accurate and efficient prediction of traffic flow on city roads plays a crucial role in reducing the negative consequences of traffic congestion. In other words, timely and accurate forecasting of road traffic is extremely important. This helps optimize traffic flows and reduce traffic jams and delays on the roads. Traffic forecasting systems are crucial for predicting such traffic jams and hazardous situations. They help to warn drivers and regulate traffic lights. Thus, such a system reduces the likelihood of accidents, which are common on city roads. Traffic jams caused by vehicles contribute to air pollution due to the increased emissions of harmful substances from stationary vehicles. Accurate and precise traffic forecasting enables the reduction of costs resulting from traffic congestion. (Lv et al., 2015).

Machine learning has become an essential tool for analyzing and predicting traffic patterns due to its ability to process large datasets and identify complex patterns. Machine learning algorithms can analyze historical traffic data, current updates, weather conditions, and other relevant factors to generate accurate predictions (Xie et al., 2020).

The adaptability and learning abilities of machine learning models allow us to constantly improve the accuracy of predictions. Consider three popular machine learning algorithms: logistic regression, random forest, and the support vector Machine method. Logistic regression is easy to implement and interpret; it works well on small datasets and in binary classification tasks. A random forest is an ensemble method that uses multiple decision trees to improve prediction accuracy and reduce the risk of overfitting. It is effective when working with large and complex data. The support vector machine (SVM) method is especially useful for classification problems with high-dimensional data and can be adapted for nonlinear decision boundaries using sound tricks. Each of these algorithms has unique advantages and features that make them suitable for different types of data and tasks (Gao et al., 2024).

The object of the study is traffic flows and their characteristics, such as the number of cars that affect the overall transport situation. The number of cars is measured by the number of vehicles located on a certain length of road. This parameter determines the current state and dynamics of traffic, and also helps to analyze and predict the traffic situation.

The purpose of this study is to develop, compare, and evaluate the effectiveness of various machine learning algorithms for predicting traffic based on data obtained from a specially collected dataset, as well as integrating the best model into a web application for real-time use.

Objectives of the Research Work:

- Collection and preprocessing of data, including traffic parameters.
- Development of machine learning models using logistic regression, random forest and support vector machine.
- Evaluating the effectiveness of models using accuracy, completeness, and F1-measure metrics.
- Compare model performance and identify the best model for predicting traffic.
- Integration of the selected model into a web application for real-time traffic prediction.

The results of the study can be used to further develop intelligent transport systems and improve the management of urban traffic flows. Scientifically based conclusions about the most effective methods of predicting traffic can be in demand when creating and improving traffic management technologies, which in turn helps to reduce traffic jams, improve road safety, and improve the environmental situation in cities. Thus, this research represents an important contribution to the development of the field of intelligent transport systems and can be used as a basis for further research in this area.

Literature review

Traffic flow management in the city is an important part of overseeing the city's overall infrastructure. This factor affects every aspect, from the efficiency of residents' daily city travel to the speed of emergency response. The year-on-year increase in population and the rapid growth in the number of vehicles on city roads have become a traffic congestion problem.

To address this, researchers in the field use machine learning and big data analysis methods to offer more efficient traffic prediction and management. The proposed literature review examines several recent studies that have investigated the technologies used in intelligent transportation systems in this area.

Bu (2019) explores the application of deep learning, a sophisticated machine learning approach, to predict urban traffic. The study highlights the use of big data analytics to process and analyze the vast amount of data generated by urban traffic sensors, cameras, and GPS devices. Special attention is paid to the effectiveness of convolutional neural networks (CNN) and recurrent neural networks (RNN) in modeling and forecasting traffic flows.

CNNs are excellent at processing spatial data, which makes them suitable for analyzing images from street traffic cameras, while RNNs excel at analyzing temporal data, ideal for time series such as speed and volume of traffic. The main findings indicate that deep learning models, when trained on large and diverse datasets, can predict traffic conditions with high accuracy. The study also discusses the computing requirements and infrastructure needed to implement these models in real-time applications for intelligent transportation systems. The integration of deep learning into traffic management systems requires high computational resources, since models must be trained on large amounts of data and executed in real time to process incoming information from sensors and cameras. However, with proper configuration and optimization, deep learning can be effectively applied in real time to improve urban traffic management and ensure traffic safety and efficiency.

Convolutional neural network (CNN) models can process spatially oriented data very well. Therefore, such models are well-suited for analyzing images obtained from street traffic cameras. RNN networks, on the other hand, are well-suited for analyzing data that changes over time. These networks are often used in such tasks to analyze time series of data, such as traffic speed and flow.

Furthermore, in the proposed review, the authors discussed the computational requirements necessary to implement the aforementioned models in real-time applications of intelligent transportation systems. Primarily, there are some challenges in implementing neural network models in urban traffic management systems. Specifically, training models with large volumes of data requires extensive computational resources. Information about vehicle traffic continuously received from sensors and cameras requires constant processing. However, by properly configuring and optimizing the model's architecture, deep learning can be applied efficiently in real time. This reduces the challenges in urban traffic management and enhances safety and efficiency.

Hanan Almuhalfi (2024) has conducted an in-depth study of the machine learning models used in such tasks. In the study, the author identifies three main types of these models and highlights the advantages of each. The author notes that supervised learning models are well-suited for short-term traffic forecasting. Meanwhile, unsupervised learning helps detect anomalies in irregular traffic patterns. This article notes that the effectiveness of machine learning models often depends on the crucial role of effective feature selection.

Shahzad Ahmad's (2025) research focuses on a type of deep learning model, the Stacked Autoencoder (SAE), for traffic flow prediction. This model effectively encodes the most important features of the traffic, and further notes that traffic flow models can be used for prediction by reconstructing their samples from scratch. The author used data collected under various traffic conditions during the study. Based on this data, the SAE model, Methods such as Support Vector Machines (SVM) and decision trees were considered. The experimental results proved that the SAE model demonstrated higher test accuracy than the other two models. Despite the computational complexity, this model can be scaled well. In other words, this model was particularly suitable for collecting and processing the large amounts of data encountered in city traffic.

In the next study, Xie et al. (2020) investigated urban traffic prediction using machine learning methods based on spatio-temporal data. The study employed models such as graph convolutional networks, Bayesian networks, and ensemble methods. The study employed models such as Bayesian networks and ensemble methods. The obtained results demonstrated that integrating spatial and temporal data significantly improves the accuracy of urban traffic flow prediction. The authors propose implementing deep learning models to achieve the best results in Intelligent Transportation Systems.

In conclusion, although recent years of research in this area have demonstrated the effective use of machine learning algorithms, real-time applications demand improved rapid decision-making. In particular, when forecasting traffic flows in large city centers, the scaling issues that arise in large-scale traffic flows must be addressed. Such calculations should necessarily include weather, special city events, and infrastructural changes (Datia et al., 2024). This issue leads to the use of distributed computing methods and cloud solutions for managing large-scale traffic (Teng et al., 2024).

Furthermore, the integration of effective machine learning models into real-time road traffic management systems is essential. Moreover, it is desirable for such systems to be accessible to the general public. It would also be beneficial for them to be designed with user-friendly interfaces and equipped with visualization tools.

Therefore, to predict and monitor city traffic, machine learning algorithms require improvement, and more research is needed in this field.

Materials and methods

To address the challenges encountered in urban traffic congestion management, this study conducted an extensive analysis of the dataset and employed three machine learning methods.The dataset used in the study was obtained from official traffic flow data from the Humanitarian Data Exchange (HDX) platform, and employed for research purposes.

This study aims to leverage advanced data analytics and machine learning techniques to predict traffic patterns and optimize traffic flow in urban settings.

Exploratory Data Analysis.

EDA is the initial stage of data analysis, the purpose of which is to gain an understanding of the data structure and identify the main characteristics. In this section, we apply various visualization and statistical analysis methods.

Data visualization

- Histograms: used to display the distribution of data for each variable.
- Scatter plots: used to show the relationship between two variables in the data.
- Heat maps: allow for the visualization of the correlation between given variables.

Statistical analysis

- Mean (μ): is calculated by dividing the sum of the data in the dataset by the total number of values:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (1)$$

- Standard deviation: measures the distribution of data in the dataset around their mean (Moumen et al., 2023):

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2} \qquad (2)$$

Pearson's correlation

Pearson Correlation

Pearson's r correlation is a statistical measure that indicates the linear relationship and its direction between input variables. This measure can take values from -1 to 1. Here, a value of 1 indicates the strongest positive correlation, -1 indicates the strongest negative correlation, and 0 indicates the complete absence of a linear relationship between the variables. The covariance is normalized by the product of the standard deviations of the variables to calculate this coefficient. A high positive coefficient indicates that as one variable increases, the other increases, while a high negative coefficient indicates that as one variable increases, the other decreases.

$$r = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{\sqrt{\sum_{i=1}^{N}(x_i - \mu_x)^2 (y_i - \mu_y)^2}} \qquad (3)$$

Feature engineering.

After completing the EDA, we move on to feature engineering, which aims to create new features that will improve the quality of our models.

1. Creation of Temporal Features. To create time features, we extracted the hour from the timestamp to reflect the daily traffic patterns. We also extracted the day of the week to account for weekly traffic variations. In addition, the month was extracted to reflect the monthly trends.

2. Data Normalization. Bringing data to a common scale to enhance model performance.

Normalization is a technique used to scale the values of a feature to a range between 0 and 1. This is done by subtracting the minimum value of the feature and then dividing by the range (difference between the maximum and minimum values).

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad (4)$$

Model Training.

After preparing the data, we proceed to train machine learning models. We consider three main algorithms: Logistic Regression, Random Forest, and Support Vector Machine (SVM).

Logistic Regression

Logistic regression is a statistical method used to model the probability of an observation belonging to a certain category or class. It is widely used in classification problems where it is necessary to predict the probability of an observation belonging to two or more classes. Instead of a direct relationship between features and the target variable, as in conventional linear regression, logistic regression uses a logistic function to convert the sum of weighted features into probability (Tripathi et al., 2023).

$$\sigma(z) = \frac{1}{1 + e^{-z}} \quad (5)$$

Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees during training and outputs the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Each tree in this model (Figure 1) is trained on a random partition of the training dataset (Jedhe et al., 2022). This method prevents overfitting and enhances the model's robustness. When predicting future data, each tree in the model independently forecasts the target variable. The final prediction aggregates the predictions obtained from all trees. Then, it calculates their mode or mean value to determine the final prediction. This ensemble Random Forest method has become widely popular due to its high accuracy, versatility, and ability to work efficiently with large datasets (Razali et al., 2021).

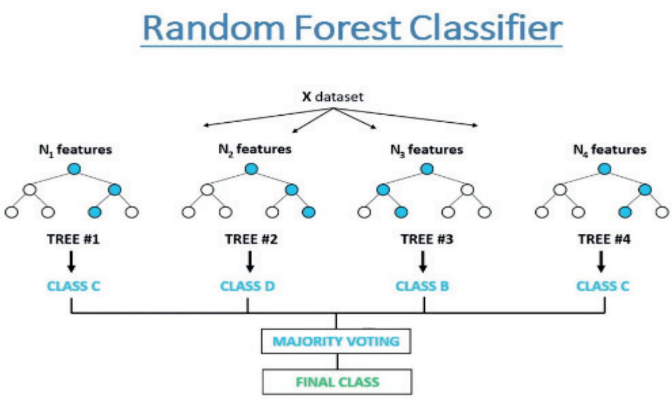

Figure 1 — Illustration of the Random Forest classification process.

Support Vector Machine (SVM)

The Support Vector Machine (SVM) method is a machine learning algorithm that builds a hyperplane in the feature space to maximize the separation of two classes of data. It uses reference vectors - the data points closest to the separating hyperplane (Oyewola et al., 2022). SVM is effective even in the case of non-linearly separable data, due to the use of nuclear functions to transition to a higher-dimensional space (Balachander et al., 2022). The C regularization parameter in SVM controls the trade-off between maximizing gap and minimizing classification errors, making it a flexible tool for a variety of classification tasks (6).

$$f(x) = w^T x + b \quad (6)$$

Results

The results of the study present the performance of the proposed machine learning models and the analysis of traffic distribution patterns. The findings are illustrated using statistical metrics and visualizations to demonstrate differences between traffic categories.

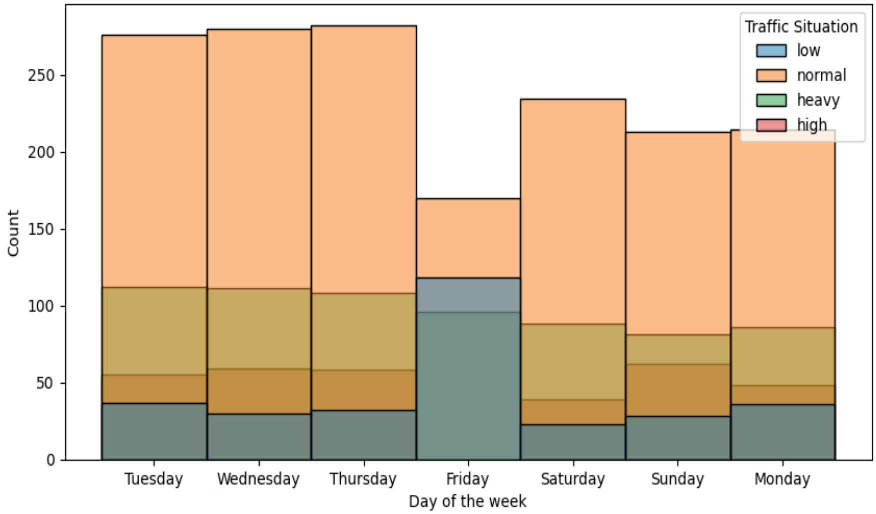

Figure 2 — Distribution of traffic situations.

The bar chart in Figure 2 displays the distribution of traffic situations across different days of the week. Each bar represents a day and is divided into segments corresponding to different traffic situations: low, normal, heavy, and high. The chart shows that traffic is predominantly normal across most days, with a notable decrease in traffic on Fridays. The data indicates variability in traffic patterns, which can be crucial for traffic prediction models. This visualization helps in understanding how traffic conditions fluctuate throughout the week, providing insights into peak and off-peak traffic days.

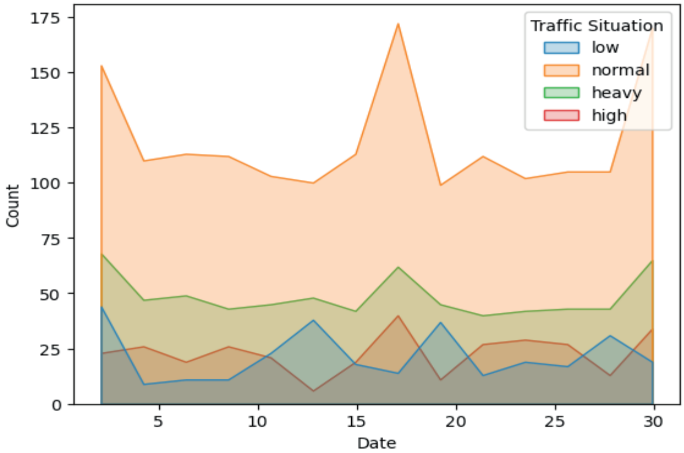

Figure 3 — Daily traffic situation over a month.

The graph (Figure 3) presents the daily traffic situation over a month, categorized into four types: low, normal, heavy, and high. The y-axis represents the count of traffic instances, while the x-axis indicates days of the month. The plot shows significant variability, particularly noticeable with 'high' traffic, which spikes dramatically towards the middle of the graph. This visualization helps in understanding the fluctuations in traffic conditions over time and can be used to predict traffic patterns or plan traffic management strategies.

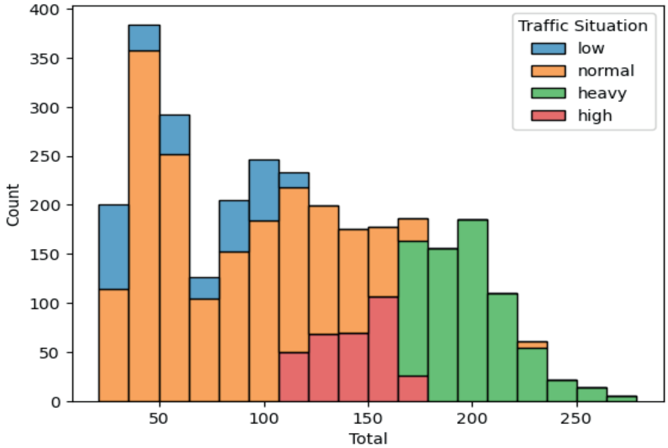

Figure 4- Distribution of traffic situations over a range of total counts.

The stacked bar (Figure 4) chart shows the distribution of traffic situations over a range of total counts. Each bar represents a total traffic count, segmented into four categories: low, normal, heavy, and high. The chart provides a clear view of how traffic conditions stack up against each other within various total count ranges. For instance, lower total counts tend to have a higher proportion of 'low' and 'normal' traffic conditions, while higher total counts include more instances of ‘heavy’ and ‘high’ conditions. This graph is useful for analyzing the composition of traffic conditions relative to the overall traffic volume.

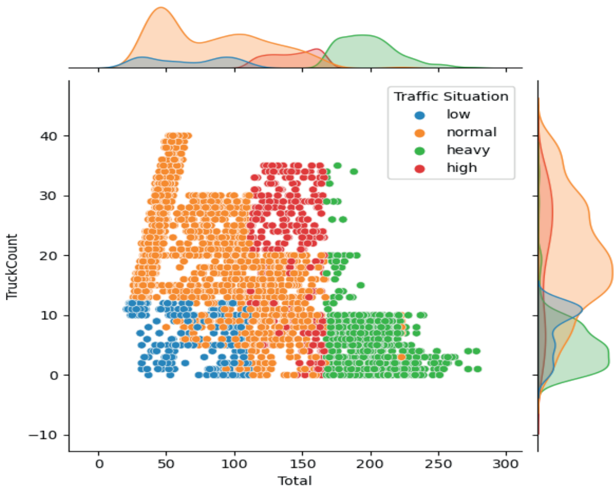

Figure 5 — Traffic situation in respective total counts and truck count.

The scatter plot in Figure 5 displays individual data points categorized by traffic situation against their respective total counts and truck counts. Each category (low, normal, heavy, high) is marked by a distinct color. The plot includes marginal histograms on the top and right, illustrating the distribution of total counts and truck counts, respectively. The scatter plot reveals a positive correlation between the total traffic counts and the severity of traffic conditions, suggesting that higher traffic volumes are often associated with more severe traffic situations.

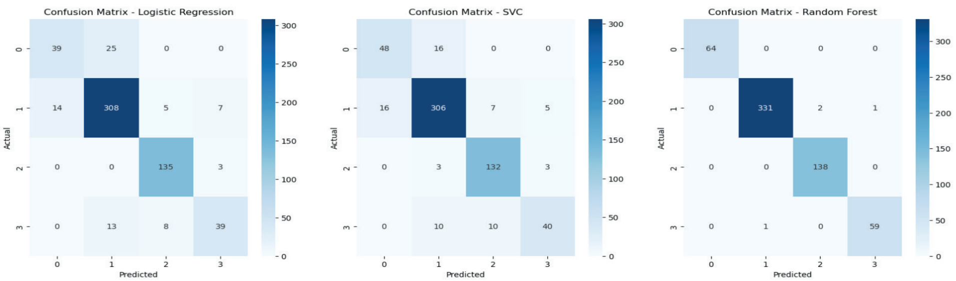

Figure 6 — Confusion matrices of models.

The image in Figure 6 shows three error matrices (confusion matrices) for three different machine learning models: logistic regression, SVC, and random forest. Each matrix shows the number of correct and incorrect predictions for four classes (0, 1, 2, and 3). For logistic regression, it can be noted that the model has difficulties classifying classes 0 and 3, often confusing them with other classes. The second matrix for the SVC model shows that it does a better job with class 0 compared to logistic regression, but also has some errors in the classification of classes 2 and 3. The third matrix for a random forest demonstrates that this model does the best job with classification, making almost no errors for classes 0 and 2, and has the least number of errors for other classes.

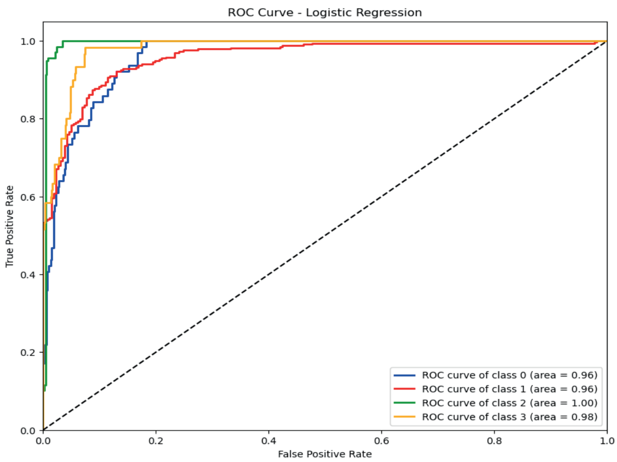

Figure 7 — ROC curves of the logistic regression model.

The graph of ROC curves in Figure 7 for logistic regression shows curves for four classes. The model shows high classification quality for all classes, with an area under the curve (AUC) from 0.96 to 1.00. Class 2 has an ideal classification with AUC=1.00, and classes 0 and 1 have AUC=0.96. This indicates that logistic regression copes well with the classification task, although for classes 0 and 1, improvements are possible.

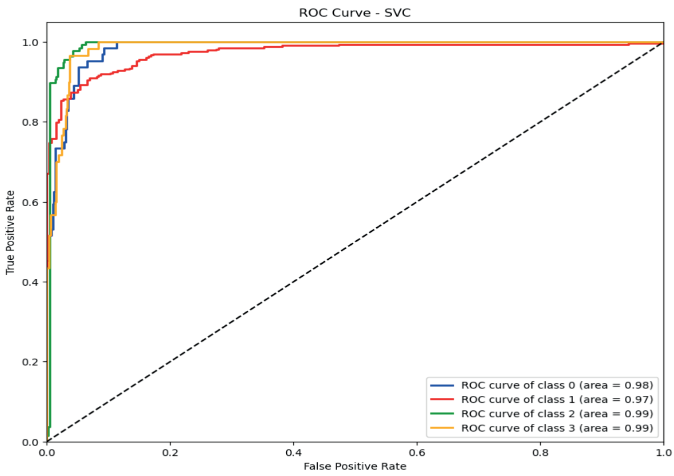

Figure 8 — ROC curves of the SVC model.

The ROC curve of the SVC model shown in Figure 8 demonstrates that it is very well classified across all given vehicle data classes. Here, the AUC values range from 0.97 to 0.99. The AUC values for class 0 were 0.98, and for class 2 they were 0.99. That is, this SVC model showed slightly better results compared to the previous logistic regression model. This SVC model is very reliable for such classification tasks.

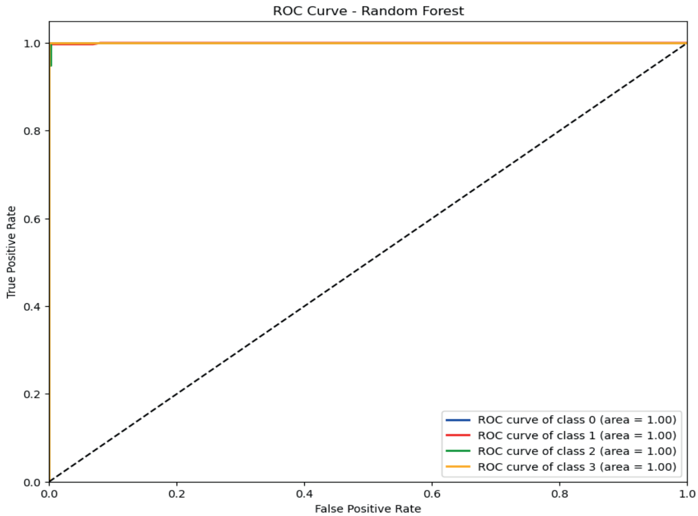

Figure 9 — ROC curves of the Random Forest model.

In the following graph in Figure 9, the ROC curves based on the results of the Random Forest model are shown. The AUC values for all classes are close to 0.99. That is, the trained model's classification quality is very high. As a result, the Random Forest model demonstrates much better discriminative ability for all classes compared to the previous two models (Figure 10).

Classification Report - Logistic Regression

	precision	recall	f1-score	support
low	0.74	0.61	0.67	64
normal	0.89	0.92	0.91	334
heavy	0.91	0.98	0.94	138
high	0.80	0.65	0.72	60
accuracy			0.87	596
macro avg	0.83	0.79	0.81	596
weighted avg	0.87	0.87	0.87	596

Classification Report - SVC

	precision	recall	f1-score	support
low	0.75	0.75	0.75	64
normal	0.91	0.92	0.91	334
heavy	0.89	0.96	0.92	138
high	0.83	0.67	0.74	60
accuracy			0.88	596
macro avg	0.85	0.82	0.83	596
weighted avg	0.88	0.88	0.88	596

Classification Report - Random Forest

	precision	recall	f1-score	support
low	1.00	1.00	1.00	64
normal	1.00	0.99	0.99	334
heavy	0.99	1.00	0.99	138
high	0.98	0.98	0.98	60
accuracy			0.99	596
macro avg	0.99	0.99	0.99	596
weighted avg	0.99	0.99	0.99	596

Figure 10 — Classification report of proposed models.

Classification Report - Logistic Regression:

Logistic regression shows accuracy (precision) and completeness (recall) for classes low (0.74 and 0.61), normal (0.89 and 0.92), heavy (0.91 and 0.98), high (0.80 and 0.65). Weighted averages (weighted avg) for accuracy, completeness and F1 measures are 0.87. The overall accuracy of the model is 0.87. The model copes best with the heavy and normal classes, but has difficulties with the low class.

Classification Report - SVC:

The SVC model shows better results compared to logistic regression for classes low (0.75 and 0.75), normal (0.91 and 0.92), heavy (0.89 and 0.96), high (0.83 and 0.67). The weighted averages for accuracy, completeness and F1 measures are 0.88. The overall accuracy of the model is 0.88. The model shows the best results for all metrics, especially for the low and high classes.

Classification Report - Random Forest:

The random forest shows the best results with accuracy and completeness for all classes equal to or almost equal to 1.00. The weighted averages for accuracy, completeness and F1 measures are 0.99. The overall accuracy of the model reaches 0.99. The model classifies all classes almost perfectly, which makes it the most effective among the presented models.

Disscussion

The results of the proposed study show that machine learning models demonstrate strong performance in predicting traffic flows. This study examined logistic regression, support vector classification (SVC), and Random Forest algorithms. Among the models considered, the Random Forest model, with an overall accuracy of 99.4%, was identified as the most effective. This result is regarded as significantly good among models used in traffic flow prediction. By accurately predicting traffic flow, it is possible to make traffic light schedules more efficient and provide drivers with real-time traffic information. This improves overall traffic flow and reduces congestion. As a result, it can lead to lower emissions and provide residents with convenient access.

The comparative analysis also highlights the importance of incorporating both spatial and temporal features in traffic modeling. The inclusion of time-based variables, such as hour and day of the week, contributed to improved classification performance, confirming findings from previous studies on spatio-temporal traffic prediction.

Despite these promising results, several limitations should be noted. The dataset used in the study may not fully represent all possible traffic scenarios, particularly rare or extreme events such as accidents, road closures, or sudden weather changes. In addition, the models were evaluated in a controlled experimental setting, which may differ from real-world deployment conditions.

Furthermore, while Random Forest demonstrated near-perfect performance, such high accuracy may indicate potential overfitting or dataset bias, and therefore requires further validation on larger and more diverse datasets. Future research should focus on incorporating real-time data streams, external influencing factors (e.g., weather conditions, special events), and exploring advanced deep learning approaches to enhance model generalization and scalability.

Conclusion

In this study, machine learning models were used to predict urban traffic flow, and their performance was comparatively analyzed. The provided data were trained, tested, and evaluated using logistic regression, support vector machine (SVM), and Random Forest algorithms. Overall, the results of all three models were good. The logistic regression model used first yielded satisfactory results. However, this model is limited in handling more complex traffic patterns.

The SVC model's results were somewhat better than those of the first model, providing a more balanced classification. Among the models used, the Random Forest model proved the most reliable and highest in accuracy.

The obtained results determine the effectiveness of ensemble learning methods for traffic forecasting tasks in large cities. In addition, it was determined that the combined use of spatial and temporal data increases the model's performance. The proposed approach can be effectively applied in real-time traffic management systems to support decision-making, optimize traffic flow, and reduce congestion in urban environments.

However, in this regard, it is also necessary to note several limitations. The dataset used in this study may not cover all possible traffic scenarios on the road. In other words, situations such as sudden road closures caused by major accidents may not be accounted for. Additionally, the study did not account for external factors such as sudden weather changes, special city events, or road repairs. Incorporating these factors into the models could enhance their predictive accuracy and reliability.

Future research will focus on incorporating additional factors, such as weather and real-time data, to improve prediction accuracy. It is also important to explore scaling and advanced deep learning models for processing big city data.

References

Almukhalfi H., Noor A., & Noor T.H. (2024) Traffic management approaches using machine learning and deep learning techniques: A survey. Engineering Applications of Artificial Intelligence, 133(Part B), Article 108147. https://doi.org/10.1016/j.engappai.2024.108147 (in Eng.).

Balachander B., Kandasamy M., Dornadula V.H.R., Nirmal M., & Alanya-Beltran J. (2022) Integrating machine learning approaches in SDN for effective traffic prediction using correlation analysis. Communications in Computer and Information Science. — P. 611–622. https://doi.org/10.1007/978-3-031-07012-9_51 (in Eng.).

Bui K.-H. N., Yi H., Jung H., & Seo, J. (2019) Big data analytics-based urban traffic prediction using deep learning in ITS. In Proceedings of the International Conference on Artificial Intelligence (ICAI) (in Eng.).

Datia N., Pato M. P., Vaz, J., & Pires, J. M. (2024) Road traffic flow prediction with visual analytics. Studies in Computational Intelligence. — P. 293–318. https://doi.org/10.1007/978-3-031-46549-9_10 (in Eng.).

Deekshetha H. R., Shreyas Madhav A. V., & Tyagi A. K. (2022) Traffic prediction using machine learning. Lecture Notes on Data Engineering and Communications Technologies. — P. 969–983. https://doi.org/10.1007/978-981-16-9605-3_68 (in Eng.).

Gao M., Wei Y., Xie Y., & Zhang Y. (2024). Traffic prediction with self-supervised learning: A heterogeneity-aware model for urban traffic flow prediction. Mathematics, 12(9). — 1290 p. https://doi.org/10.3390/math12091290 (in Eng.).

Jedhe A., & Chopade S. (2022) Traffic prediction for intelligent transportation system using machine learning. International Journal for Research in Applied Science and Engineering Technology, 10(9). — P. 244–257. *https://doi.org/10.22214/ijraset.2022. 46604* (in Eng.).

Lv Y., Duan Y., Kang W., Li Z., & Wang F.-Y. (2015) Traffic flow prediction with big data: A deep learning approach. IEEE Transactions on Intelligent Transportation Systems, 16(2). — P. 865–873. *https://doi.org/10.1109/TITS.2014.2345663* (in Eng.).

Moumen I., Abouchabaka J., & Rafalia N. (2023). Adaptive traffic lights based on traffic flow prediction using machine learning models. International Journal of Electrical and Computer Engineering, 13(5). — 5813 p. https://doi.org/10.11591/ijece.v13i5.pp5813-5823 (in Eng.).

Nashaat H., Mohammed N. H., Abdel-Mageid S. M., & Rizk R. Y. (2024) Machine learning-based cellular traffic prediction using data reduction techniques. IEEE Access. https://doi.org/10.1109/access.2024.3392624 (in Eng.).

Razali N.A.M., Shamsaimon N., Ishak K.K., Ramli S., Amran M.F.M., & Sukardi S. (2021) Gap, techniques and evaluation: Traffic flow prediction using machine learning and deep learning. Journal of Big Data, 8(1). https://doi.org/10.1186/s40537-021-00542-7 (in Eng.).

Srivastava A., Aryan N., & Khurana M. (2023) A survey paper on traffic prediction using machine learning. International Journal for Research in Applied Science and Engineering Technology, 11(5). — P. 44–48. https://doi.org/10.22214/ijraset.2023.51390 (in Eng.).

Teng T., Jagadeesh G. R., Kunal T., & Chung C. C. (2024) Traffic state prediction of perturbed and non-perturbed traffic scenarios. Lecture Notes in Networks and Systems. — P. 129–144. https://doi.org/10.1007/978-3-031-47721-8_8 (in Eng.).

Tripathi A. N., & Sharma B. (2023). A deep review: Techniques, findings and limitations of traffic flow prediction using machine learning. Lecture Notes in Mechanical Engineering. — P. 21–33. https://doi.org/10.1007/978-981-99-2921-4_3 (in Eng.).

Wang Q., Chen C., Zhang L., Song X., Li H., Zhao Q., Niu B., & Gu J. (2024) Period extraction for traffic flow prediction. Lecture Notes in Computer Science. — P. 490–501. https://doi.org/10.1007/978-981-97-0798-0_29 (in Eng.).

Xie P., Li T., Liu J., Du S., & Yang X. (2020). Urban flow prediction from spatiotemporal data using machine learning: A survey. Information Fusion, 59. — P. 1–12. https://doi.org/10.1016/j.inffus.2020.01.002 (in Eng.).

Publication Ethics and Publication Malpractice
in the journals of the Central Asian Academic Research Center LLP

For information on Ethics in publishing and Ethical guidelines for journal publication see http://www.elsevier.com/publishingethics and http://www.elsevier.com/journal-authors/ethics.

Submission of an article to the journals of the Central Asian Academic Research Center LLP implies that the described work has not been published previously (except in the form of an abstract or as part of a published lecture or academic thesis or as an electronic preprint, see http://www.elsevier.com/postingpolicy), that it is not under consideration for publication elsewhere, that its publication is approved by all authors and tacitly or explicitly by the responsible authorities where the work was carried out, and that, if accepted, it will not be published elsewhere in the same form, in English or in any other language, including electronically without the written consent of the copyright-holder. In particular, translations into English of papers already published in another language are not accepted.

No other forms of scientific misconduct are allowed, such as plagiarism, falsification, fraudulent data, incorrect interpretation of other works, incorrect citations, etc. The Central Asian Academic Research Center LLP follows the Code of Conduct of the Committee on Publication Ethics (COPE), and follows the COPE Flowcharts for Resolving Cases of Suspected Misconduct (http://publicationethics.org/files/u2/New_Code.pdf). To verify originality, your article may be checked by the Cross Check originality detection service http://www.elsevier.com/editors/plagdetect.

The authors are obliged to participate in peer review process and be ready to provide corrections, clarifications, retractions and apologies when needed. All authors of a paper should have significantly contributed to the research.

The reviewers should provide objective judgments and should point out relevant published works which are not yet cited. Reviewed articles should be treated confidentially. The reviewers will be chosen in such a way that there is no conflict of interests with respect to the research, the authors and/or the research funders.

The editors have complete responsibility and authority to reject or accept a paper, and they will only accept a paper when reasonably certain. They will preserve anonymity of reviewers and promote publication of corrections, clarifications, retractions and apologies when needed. The acceptance of a paper automatically implies the copyright transfer to the Central Asian Academic Research Center LLP.

The Editorial Board of the Central Asian Academic Research Center LLP will monitor and safeguard publishing ethics.

Правила оформления статьи для публикации в журнале смотреть на сайтах:
www:nauka-nanrk.kz
http://physics-mathematics.kz/index.php/en/archive

ISSN2518-1726 (Online),
ISSN 1991-346X (Print)

Ответственный редактор *А. Ботанқызы*
Редакторы: *Д.С. Аленов, Т. Апендиев*
Верстка на компьютере: *Г.Д. Жадырановой*

Подписано в печать 31.03.2026.
Формат 60x881/8.
20,0 п.л. Заказ 1.

«Central Asian Academic Research Center» LLP
Алматы, Шевченко к-сі, 28