UDC 656.13 DOI: 10.30977/AT.2219-8342.2025.57.0.07

Proactive analysis of road traffic accidents in the Republic of Kazakhstan based on machine learning models and geographic information systems

Kobdikova Sh.[1], Chupekov Y.[2], Arimbekova P.[3], Nokhatov M.[4]

[1]Almaty Academy of Internal Affairs of the Republic of Kazakhstan named after M.Esbulatov, Kazakhstan
[2]Police Department of the Almaty region of the Ministry of Internal Affairs, Kazakhstan
[3]Kazakh - German University, Kazakhstan
[4]Kazakh Automobile and Highway Institute named after L.B. Goncharov, Kazakhstan

Abstract. ***Problem.*** *The article is devoted to the development of a methodology for proactive analysis of road traffic accidents (RTAs) in the Republic of Kazakhstan (RK). Traditional retrospective approaches do not provide sufficient effectiveness in preventing incidents under conditions of annually. increasing accident rates and significant socio-economic losses, which exceed USD 7 billion per year.* ***Goal.*** *The aim of this study is to provide a theoretical justification for a proactive analysis methodology based on machine learning (ML) models.* ***Methodology.*** *The proposed approach is grounded in the integration of Big Data obtained from the national digital platform TOR (Traffic Operational Response) and the application of predictive ML models such as Random Forest and XGBoost.* ***Originality.*** *The scientific novelty lies in the synthesis of ML models and GIS-based analysis to create a dynamic proactive model for RTA risk assessment, adapted for the first time to the specific data environment of Kazakhstan. The developed framework enables the prediction of both the probability and the severity of RTAs on specific road segments using dynamic influencing factors.* ***Results.*** *The results can be utilized by road infrastructure agencies and law enforcement bodies in Kazakhstan for targeted patrolling and proactive interventions.* ***Practical value.*** *It is recommended to integrate the ML module directly into the TOR platform and to establish standardized interagency data exchange procedures.*

Keywords: *proactive rta analysis; machine learning; random forest; xgboost; gis analysis; big data; TOR (Traffic Operational Response); risk prediction; Kazakhstan; road safety.*

Introduction

The problem of ensuring road traffic safety is critically important for the Republic of Kazakhstan (RK). A high level of road traffic injuries and fatalities not only leads to irreparable social losses but also significant damage to the national economy. According to the Ministry of Internal Affairs of the Republic of Kazakhstan, the year 2024 was marked by a record increase in the number of road traffic accident (RTA) [1], resulting in rising social losses and economic damage, which annually exceed 7 billion USD. A particularly alarming trend was observed in 2024, according to the statistics of the Ministry of Internal Affairs of the Republic of Kazakhstan [1].

Analysis of publications

According to the data of the Committee on legal statistic and special accounts of the General Prosecutor's office of the Republic of Kazakhstan (CLSSA GP RK), a total of 246.648 road traffic accidents were recorded in the country between 2000 and 2024. These accidents resulted in 38.648 fatalities and 312.059 injuries of varying severity [1]. Despite the measures undertaken, the overall dynamics of road traffic accidents remain of considerable concern, as evidenced by the comparison of social and transport risks (Fig. 1).

Citation: Kobdikova Sh., Chupekov Y., Arimbekova P. (2025) Proactive analysis of road traffic accidents in the Republic of Kazakhstan based on machine learning models and geographic information systems *Automobile Transport*, (57), 54–59. https://doi.org/10.30977/AT.2219-8342.2025.57.0.07

Received (надійшла) 02.10.2025
Revised (рецензована) 22.10.2025
Accepted (прийнята до друку) 15.12.2025
Published (опублікована) 30.12.2025

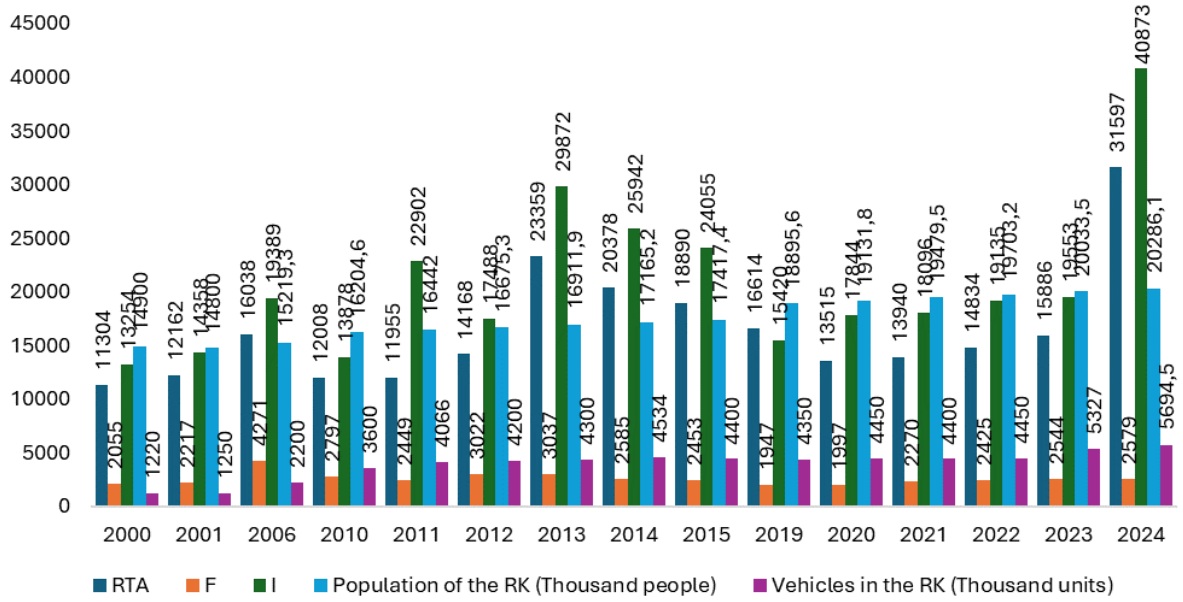

Fig. 1. Dynamics of road traffic accidents (RTA), severity of their consequences, and social transport risk expressed in coefficients

The key control able factor of accident occurrence remains the human factor (maneuvering, violations, speeding) and the condition of road infrastructure. In the context of rapidly increasing traffic volumes and urbanization, the exciting practice of responding to road traffic accidents remains largely retrospective, as analysis is conducted only after incidents have already occurred. This approach demonstrates limited effectiveness, since safety measures are introduced only after a road segment has already become a “high-risk hotspot” [2].

A shift toward proactive methods capable of predicting the occurrence of hazardous situations and enabling the implementation of preventive measures in advance is required – an approach supported by international research findings [3-5]. Proactive analysis involves forecasting the probability of road traffic accidents (RTA) on a specific road segment in the near future (for example, within the next 1-4 hours), based on a combination of static and, critically, dynamic predictors. This enables road maintenance and law enforcement agencies to respond preemptively.

The scientific novelty of the study lies in the synthesis of ML models (Random Forest, XGBoots) and GIS-based analysis to create a dynamic proactive model for assessing road traffic accident risk, adapted for the first time to the specific datasets and factors of the road network of the Republic of Kazakhstan [6], including its integration with the unified digital platform TOR (Traffic Operational Response) [6].

Traditional road safety analysis methods focus primarily on statistically identifying accident-prone locations based on historical data. Modules such as Poisson regression and the negative binormal distribution are effective for detecting static risk factors (road geometry, density of roadside infrastructure), yet they overlook the dynamic spatio-temporal dimension. Road safety is determined by the complex interaction between weather conditions and driver behavior. Retrospective approaches are unable to predict when a previously safe road segment may become critically hazardous due to, for example, sudden snowfall or peak traffic congestion. This highlights the necessity of transitioning to proactive, dynamic methods [3, 7].

The emergence of big data and the growth of computational capacity have positioned machine learning (ML) methods at the forefront of road traffic accident prediction research [2, 8, 5]. ML models, particularly ensemble approaches [9], offer a significant advantage, as they are capable of capturing nonlinear relationships, processing heterogeneous data, and effectively addressing the problem of class imbalance (road traffic accidents (RTA) are rare events compared to long periods of safe driving) [2].

Effective prediction is achieved through the use of machine learning models such as Random Forest and Gradient Boosting [9], which can process large volumes of diverse data and identify nonlinear dependencies among hundreds of predictors (weather conditions, traffic patterns, traffic violations) [2, 4].

The Random Forest model has proven to be a powerful ensemble method [9, 10]. RF, which relies on aggregating the results of numerous independent decisions trees, is highly resistant to outliers and overfitting [9, 10]. A key advantage of RF is its ability to access feature importance, allowing road safety researchers and traffic safety engineers to accurately determine which factor contributes most to crash risk [2].

Gradient Boosting Decisions Trees (GBDT), particularly their optimized implementation XGBoots (Extreme Gradient Boosting), have demonstrated superior classification and prediction accuracy in road safety applications [11], [10]. Unlike RF, where trees are generated independently, boosting algorithms iteratively correct the errors of previous trees, thereby improving overall accuracy [11]. Gradient Boosting methods have shown strong performance when integrating real-time traffic data [11].

Recent development also includes the use of deep learning approaches, such as convolutional and recurrent neural networks (CNNs and RNNs), which are particularly effective in processing high-dimensional spatio-temporal data such as video streams or highly detailed vehicle trajectory information [3, 8]. However, for an operational ML system based on structured traffic and weather datasets – such as those provided by the TOR platform – ensemble methods often provide the best balance between accuracy, interpretability, and computational efficiency [9, 10].

Special attention in the literature is devoted to models evaluating road traffic accidents severity (fatalities and injuries), which is essential for selecting appropriate preventive measures [12]. Early studies primarily relied on logistic regression. Modern approaches, including XGBoost, have also been successfully applied to severity classification tasks, enabling more detailed and informed response planning [10, 12].

Proactive analysis requires the integration of Big Data that combines both static and dynamic information. Recent trends emphasize spatio-temporal Big Data analytics [7], which fully aligns with the aims of integrating predictive models into Kazakhstan's national digital platform TOR [6].

For visualization and operational decision making, the use of Geographic information System (GIS) and the Hotspot Analysis are critically important [8]. GIS enables the transformation of numerical crash probability forecasts into intuitive Risk Maps, which can serve as a foundation for proactive interventions [8]. The synthesis of ML-based forecasting and GIS based analysis (such as Hotspot Analysis) is a key element developing effective proactive safety management system [8].

Purpose and Tasks

The purpose of this study is to develop and theoretically substantiate a methodology for proactive analysis and prediction of road traffic accident in the Republic of Kazakhstan, based on machine learning models adapted for integration on with the national digital infrastructure.

The objectives of the study include:

1. Analyzing the current state of road traffic accident in Kazakhstan and identifying key risk factors.

2. Justifying the need to transition from retrospective to proactive accident analysis.

3. Selecting and adapting machine learning algorithms for predicting the likelihood and severity of road traffic accidents.

4. Developing an algorithm for integrating the predictive model with existing data sources (TOR platform, KazHydromet).

5. Proposing practical recommendations for the use of High-Risk Maps to exchange road traffic safety in Kazakhstan.

Methodology for Proactive Analysis Based on ML and GIS

Proactive analysis requires the use of Big Data that integrates both static and dynamic information [7]. In the context of the Republic of Kazakhstan, the unified digital platform TOR (Traffic Operation Response) serves as the primary system for collecting and aggregating dynamic data [6]. TOR functions as a central hub for acquiring information from video surveillance cameras, automated traffic enforcement system, and when necessary, GPS/GLONASS trackers installed on public and commercial transport vehicles [6].

The data integration framework includes:

1. Static data:

Geocoordinates of historical road traffic accidents (Committee on Legal Statistics and Special Records of the General Prosecutor's Office of the Republic of Kazakhstan [1].

Road infrastructure characteristics (road surface type, presence of markings, lighting, geometric configuration of the segments) (JSC "NC "KazAvtoZhol" [6]).

2. Dynamic Data (Predictors):

Traffic indicators: Traffic volume, average flow speed, speed variability, recorder speeding events (TOR data [6]).

Meteorological data: Air and pavement temperature, precipitations (snow, rain), visibility, presence of ice (Kazhydromet data).

Temporal factors: Day of the week, peak hours, public holidays [2].

To enable proactive forecasting, dynamic data must be aggregated and averaged across road segments with high spatial resolution (e.g., 100-500 meters) and high temporal resolution (1-4 hours).

For road traffic accident risk prediction, ensemble models – particularly Random Forest – are proposed as a suitable choice [9]. This model is selected due to its robustness to outliers and incomplete data, as well as its ability to generate feature importance scores, which allow for identifying the key risk factors specific to Kazakhstan [2, 10].

As a primary classification algorithm for binary prediction (road traffic accident will occur/ will not occur) XGBoost may be employed [10, 11]. As demonstrated in the literature, XGBoost outperforms Random Forest in predictive accuracy in traffic accident prediction tasks, especially when large volumes of dynamic predictors are present [11].

The forecasting model is based on the eXtreme Gradient Boosted (XGBoost) algorithm which minimized the loss function through additive learning:

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \qquad (1)$$

where, $\hat{y}_i^{(t)}$ is predicted road traffic accident (RTA) risk for observation i at iteration t, and $f_t(x_i)$ is the new decision tree added at stage t to minimize the residuals.

Minimization is performed with respect to the regularized loss function $\lambda^{(t)}$:

$$\lambda^{(t)} = \sum_{i=1}^{n} l \cdot (y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \qquad (2)$$

where, l is the loss function (e.g., logistic loss binary classification), and $\Omega(f_t)$ is the regularization term controlling model complexity and preventing overfitting [11].

The predicted parameter $P_{RTA}(t,G)$ represents the probability that RTA will occur on segment G within the forecast interval t (e.g., the next 24 hours).

For decision-making, it is essential to consider not only the probability of an RTA but also its potential severity. This allows prioritizing preventive interventions on segments with both high predicted probability and high expected severity of outcomes.

An integrated risk indicator R for segment G at time t combines the predicted probability and the historical RTA severity:

$$R_{G,t} = P_{RTA}(t,G) \cdot K_{Gravity}(G) \qquad (3)$$

where, $P_{RTA}(t,G)$ – the predicted probability of an RTA, calculated by the ML model; $K_{Gravity}(G)$ - the severity coefficient, based on the average severity of RTAs (ratio of fatalities to injures) historically recorded on the given road segment [12].

The severity coefficient $K_{Gravity}(G)$ is calculated based on retrospective data on the severity of RTAs within segment G:

$$K_{Gravity}(G) = \frac{Injuries_G + Fatalities_G}{Total\_Crashes_G} \qquad (4)$$

where, a is a weighting coefficient reflecting the policy priority of preventing fatalities ($a >> 1$).

The forecasting results $R_{G,t}$ are visualized on a geospatial map of the road network. Using GIS technologies and the Hotspot Analysis method, Proactive Risks Maps are generated [8].

Hotspot Analysis makes it possible to identify statistically significant clusters of elevated risk $R_{G,t}$ on the map. Road segments are color-coded (from green to red), enabling rapid operational communication with responsible agencies.

Color interpretations:

Green: Low risk ($R < R_{low}$) no preventive measures required.

Yellow: Moderate risk ($R_{low} \le R \le R_{med}$), monitoring is necessary.

Red: High risk ($R \ge R_{high}$), immediate proactive intervention is required (targeted patrolling, roadworks) [8].

Results and Discussion

Analysis of Predictive Performance.

The theoretical validation of the model using data from Kazakhstan allows for the identification of the most significant predictors for the national road network.

Key results. The superiority of the proactive model: unlike traditional retrospective approaches, the proactive ML-based model is capable of detecting potential accident hotspots that become hazardous only under specific combinations of dynamic factors [3, 4]. This enables agencies to respond to the current risk dynamics than its historical patterns [6].

Quality metrics: based on international and domestic studies employing XGBoost [10, 11], the expected prediction accuracy is 85-90%, while the area under the ROC curve (AUC) reaches 0,88 – 0,95. These indicators confirm the model's high discriminative power [11].

Key Predictors in the Context of the Republic of Kazakhstan:

1 Speed monitoring indicator (dynamic): an increase in average traffic speed on a segment, especially when combined with adverse weather factors [4, 11].

2 Road surface condition: poor pavement quality (according to [6]), which becomes a significant risk factor when paired with weather conditions [6].

3 Weather factors: sudden temperature drops (leading to ice formation) or heavy precipitations [4].

4 Temporal factors: peak hours and nighttime periods [2].

Practical implementation of Risk Maps

The implementation of Proactive Risk Maps based on GIS ensures operational decision-making capabilities [8]. A map updated at a high frequency (ever 1 – 4 hours) allow for:

Targeted patrolling: police services receive automatically generated instructions for patrolling "red" segments during the forecasted time interval, significantly improving the effectiveness of limited resources [8]. Instead of static patrolling of historically dangerous locations, resources are redirected to areas where the risk is maximum at the present moment [7].

Preventive maintenance: road maintenance services can proactively plan de-icing operations or address surface defects in advance, relying on the forecasted peak risk level [8]. For example, a prediction indicating a high likelihood of ice formation on a specific segment, amplified by heavy traffic, triggers the proactive deployment of specialized road maintenance equipment.

Conclusions and Recommendations

Conclusions. A methodology for proactive analysis of RTAs in the Republic of Kazakhstan has been developed and sustained, based on the integration on Big Data (TOR [6], meteorological data) and predictive machine learning models [2, 5].

The scientific and practical necessity of transportation from reactive responses to forecasting RTAs has been demonstrated, which is critically important given the increasing accident rates in Kazakhstan [1, 3].

The purposed approach enables dynamic risk assessment on road segments, identification of key influencing factors, and visualization of results in the form of Proactive Risk Maps [8].

Practical Recommendations for Implementation in Kazakhstan.

Integration with TOR: it is recommended that the Ministry of Internal Affairs of the Republic of Kazakhstan initiate the development of an ML module for the TOR platform, which would utilize its aggregated data [6] as the primary input features for calculating RTA probability [3].

Interagency collaboration: ensure standardized data exchange between the Ministry of Internal Affairs (RTAs), JSC "NC "KazAvtoZhol" (road conditions [6]), and KazHydromet (weather forecasts) to improve accuracy of the predictive model.

Pilot project: conduct pilot testing methodology in Almaty or Astana – regions with the highest traffic density and RTA rates – to evaluate the effectiveness of preventive measures based on ML based forecasting [8].

Conflict of interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

1. General Prosecutor's Office of the Republic of Kazakhstan. (2025). *Statistical data on registered road traffic accidents in the Republic of Kazakhstan for 2023–2024*. Retrieved November 13, from URL https://tgstat.gkp.gov.kz.
2. Ministry of Internal Affairs of the Republic of Kazakhstan. (2025). *On the introduction of the unified digital platform Traffic Operational Response (TOR) in the cities of Kazakhstan*. Retrieved November 13, from URL https://www.gov.kz/mvd.
3. Mohammed, S., Alkhereibi, A. H., Abulibdeh, A., Jawarneh, R. N., & Balakrishnan, P. (2023). *GIS-based spatiotemporal analysis for road traffic crashes; in support of sustainable transportation planning. Transportation Research Interdisciplinary Perspectives*, **20,** 100836. https://doi.org/10.1016/j.trip.2023.100836
4. Li H, Chen L (2025) Traffic accident risk prediction based on deep learning and spatiotemporal features of vehicle trajectories. *PLoS One* **20(5):** 1-28 e0320656. https://doi.org/10.1371/journal.pone.0320656
5. Hazaymeh, K., Almagbile, A., & Alomari, A. H. (2022). Spatiotemporal analysis of traffic accidents hotspots based on geospatial

techniques. *ISPRS International Journal of Geo-Information,* **11(4),** 260. https://doi.org/10.3390/ijgi11040260.
6. Bhele, R., Dhungana, S., Chimoriya, D., Sapkota, A., & Ghorasaini, S. (2024). *Spatial and temporal analysis of road traffic accidents using GIS. International Journal on Engineering Technology,* **2(1),** 1–18. https://doi.org/10.3126/injet.v2i1.72464
7. Li, Yue & Shi, Yuanyuan & Xiong, Huiyuan & Jian, Feng & Yu, Xinxin & Sun, Shuo & Meng, Yunlong. (2024). Investigating Influence Factors on Traffic Safety Based on a Hybrid Approach: Taking Pedestrians as an Example. *Sensors.* **24(23).** 7720. https://doi.org/10.3390/s24237720
8. Almahdi, A., & Al Mamlook, R. E. (2023). *Boosting Ensemble Learning for freeway crash classification under varying traffic conditions. Sustainability,* **15(22),** 15896. https://doi.org/10.3390/su152215896
9. Li, P., & Abdel-Aty, M. (2022). *A hybrid machine learning model for predicting real-time secondary crash likelihood. Accident Analysis & Prevention, 165,* 106504. https://doi.org/10.1016/j.aap.2021.106504
10. Senasinghe, A. P., de Barros, A., Wirasinghe, S. C., & Tay, R. (2024). *Factors affecting crash severity on two major intercity roads in Western Sri Lanka: A random parameter logit approach. Journal of South Asian Logistics and Transport,* **4(2),** 41–65. https://doi.org/10.4038/jsalt.v4i2.91
11. Sun, Y., Zhu, Q., Li, S., & Qin, K. (2025). *Spatiotemporal risk mapping of statewide weather-related traffic crashes: A machine learning approach. Machine Learning with Applications,* **20,** 100642. https://doi.org/10.1016/j.mlwa.2025.100642
12. Budzyński, A., Federowicz, M., Jabłoński, A., & Hasan, W. (2023). Prediction of road accidents using machine learning algorithms. *Middle East Journal of Applied Science & Technology,* **6(2),** 64-75.

Kobdikova Sh.M.[1] Doctor of Technical Sciences, Professor of the Almaty Academy of the Ministry of Internal Affairs of the Republic of Kazakhstan named after M. Esbulatov, e-mail: shkobdikova@gmail.com
Chupekov Y.K.[2], Police Department of the Almaty region of the Ministry of Internal Affairs, e-mail: yelzhankali@gmail.com
Arimbekova P.M[3], Kazakh - German University e-mail: arimbekova@dku.kz
Nokhatov M.A.[4], doctoral student at the Kazakh Automobile and Highway Institute named after L.B. Goncharov e-mail: muzer.7aa@gmail.com

[1]Almaty Academy of Internal Affairs of the Republic of Kazakhstan named after M.Esbulatov, Kazakhstan
[2]Police Department of the Almaty region of the Ministry of Internal Affairs, Kazakhstan
[3]Kazakh - German University, Kazakhstan
[4]Kazakh Automobile and Highway Institute named after L.B. Goncharov

Проактивний аналіз дорожньо-транспортних пригод у Республіці Казахстан на основі моделей машинного навчання та геоінформаційних систем

Анотація. Проблема. *Стаття присвячена розробці методології проактивного аналізу дорожньо-транспортних пригод (ДТП) у Республіці Казахстан (РК). Традиційні ретроспективні підходи не забезпечують достатню ефективність у запобіганні інцидентам за умов щорічного зростання кількості аварій та значних соціально-економічних втрат, що перевищують 7 мільярдів доларів США на рік.* ***Мета.*** *Метою цього дослідження є надання теоретичного обґрунтування методології проактивного аналізу на основі моделей машинного навчання (ML).* ***Методологія.*** *Запропонований підхід ґрунтується на інтеграції Big Data, отриманих із національної цифрової платформи TOR (Traffic Operational Response – Оперативне реагування на дорожній рух), та застосуванні прогнозних моделей ML, таких як Random Forest та XGBoost.* ***Новизна.*** *Наукова новизна полягає в синтезі моделей ML та аналізу на основі GIS для створення динамічної проактивної моделі оцінки ризику ДТП, вперше адаптованої до специфічного інформаційного середовища Казахстану. Розроблена структура дозволяє прогнозувати як ймовірність, так і тяжкість ДТП на конкретних ділянках доріг із врахуванням динамічних факторів впливу.* ***Результати.*** *Результати можуть бути використані агентствами дорожньої інфраструктури та правоохоронними органами Казахстану для цільового патрулювання та проактивних заходів втручання.* ***Практична цінність.*** *Рекомендується інтегрувати модуль ML безпосередньо в платформу TOR та встановити стандартизовані процедури міжвідомчого обміну даними.*

Ключові слова: *проактивний аналіз RTA; машинне навчання; випадковий ліс; xgboost; гіс-аналіз; великі дані; TOR (реагування на дорожній рух); прогнозування ризиків; Казахстан; безпека дорожнього руху.*

Кобдикова Ш.М.[1], д.т.н., професор Алматинської академії МВС Республіки Казахстан імені М. Есбулатова, e-mail: shkobdikova@gmail.com
Чупеков Ю.К.[2], Департамент поліції Алматинської області Міністерства внутрішніх справ e-mail: yelzhankali@gmail.com
Аримбекова П.М.[3], Казахсько-Німецький університет e-mail: arimbekova@dku.kz
Нохатов М.А.[4], Докторант Казахського автомобільно-дорожнього інституту імені Л.Б. Гончарова, e-mail: muzer.7aa@gmail.com

[1]Алматинська академія внутрішніх справ Республіки Казахстан імені М.Есбулатова, Казахстан
[2]Департамент поліції Алматинської області Міністерства внутрішніх справ, Казахстан
[3]Казахсько-Німецький університет, Казахстан
[4]Казахський автомобільно-дорожній інститут імені Л.Б. Гончарова