Sinopfbd, 2025; 10(2): 354-379 https://doi.org/10.33484/sinopfbd.1659592

Sinop Üniversitesi Fen Bilimleri Dergisi
Sinop Uni J Nat Sci
E-ISSN: 2564-7873

https://dergipark.org.tr/tr/pub/sinopfbd

Traffic Accident Analysis and Prediction Using Machine Learning Models in Türkiye from 2021 to 2024

Md Al Amin HOSSAIN[1], Humar KAHRAMANLI ÖRNEK[2] and Tahir SAĞ[2]

How to cite: Hossain, A. A., & Kahramanlı Örnek, H., Sağ, T. (2025). Traffic accident analysis and prediction using machine learning models in Türkiye from 2021 to 2024. *Sinop Üniversitesi Fen Bilimleri Dergisi, 10*(2), 354-379. https://doi.org/10.33484/sinopfbd.165592

Research Article

Corresponding Author
Tahir SAĞ
tahirsag@selcuk.edu.tr

ORCID of the Authors
A. H: 0000-0003-3382-5300
H.K.Ö: 0000-0003-2336-7924
T.S: 0000-0001-8266-7148

Received: 17.03.2025
Accepted: 06.08.2025

Abstract
Traffic accidents represent a major challenge to public safety and urban development. In recent years, the number of road traffic accidents has been increasing due to the rising global population and the growing number of vehicles, leading to numerous fatalities and injuries. This study examines traffic accidents in five major cities of Türkiye from 2021 to 2024, aiming to identify trends and predict future accidents using linear regression and random forest regressor models. Data for this analysis were obtained from the Gendarmerie General Command of the Ministry of Internal Affairs, Republic of Türkiye. To evaluate model performance, key metrics such as Mean Absolute Error, Mean Squared Error, and R-squared were utilized. The results indicate significant variations in accident patterns across cities, months, and years. Furthermore, findings highlight the effectiveness of machine learning models in predicting traffic incidents with high accuracy. Among the two models, the random forest regressor outperforms linear regression in terms of evaluation metrics. Moreover, the analytical results indicate an upward trend in accidents, fatalities, and injuries across the five cities, particularly in Ankara and İzmir. These predictive and analytical insights can provide valuable guidance for policymakers and researchers in formulating effective strategies to mitigate traffic accidents and enhance road safety.

Keywords: Linear regression, Machine learning, random forest, traffic accident

Türkiye'de 2021-2024 Yılları Arasında Makine Öğrenmesi Modelleri Kullanılarak Trafik Kazası Analizi ve Tahmini

[1]Selcuk University, Graduate School of Natural and Applied Sciences, Konya, Türkiye

[2]Selcuk University, Department of Computer Engineering, Konya, Türkiye

Öz
Trafik kazaları, kamu güvenliği ve kentsel gelişim açısından önemli bir sorun teşkil etmektedir. Son yıllarda, dünya nüfusunun artması ve araç sayısının çoğalması nedeniyle trafik kazalarının sayısı yükselmiş, bu da çok sayıda ölüm ve yaralanmaya yol açmıştır. Bu çalışma, 2021-2024 yılları arasında Türkiye'nin beş büyük şehrindeki trafik kazalarını inceleyerek eğilimleri belirlemeyi ve doğrusal regresyon ile rastgele orman regresyon modelleri kullanarak gelecekteki kazaları tahmin etmeyi amaçlamaktadır. Bu analiz için veriler, Türkiye Cumhuriyeti İçişleri Bakanlığı Jandarma Genel Komutanlığı'ndan temin edilmiştir. Model performansını değerlendirmek için Ortalama Mutlak Hata, Ortalama Kare Hatası ve R-kare gibi temel değerlendirme ölçütleri kullanılmıştır. Sonuçlar, kazaların şehirler, aylar ve yıllar arasında önemli

farklılıklar gösterdiğini ortaya koymaktadır. Ayrıca bulgular, makine öğrenimi modellerinin trafik kazalarını yüksek doğrulukla tahmin etme konusundaki etkinliğini vurgulamaktadır. İki model arasında, rastgele orman regresörünün, değerlendirme ölçütleri açısından doğrusal regresyondan daha üstün performans gösterdiği belirlenmiştir. Bunun yanı sıra, analiz sonuçları beş şehirde, özellikle Ankara ve İzmir'de, kaza, ölüm ve yaralanma sayılarında artış eğilimi olduğunu göstermektedir. Bu öngörüsel ve analitik bulgular, trafik kazalarını azaltmak ve yol güvenliğini artırmak için politika yapıcılar ve araştırmacılar için yol gösterici olabilir.

Anahtar Kelimeler: Doğrusal regresyon, makine öğrenimi, rastgele orman, trafik kazası

Introduction

Traffic accidents remain a critical public health and safety concern worldwide, with significant social and economic implications. According to the Highway Traffic Law, a traffic accident is defined as an event that occurs on roadways involving one or more moving vehicles, resulting in death, injury, or property damage. According to the World Health Organization (WHO), over 1.19 million people lose their lives and nearly 50 million people were injured in 2023 [2]. Moreover 1.35 million people die in traffic accidents each year [2]. On the other hand, approximately 80% of roads do not meet a minimum safety rating for pedestrians [2]. In Turkey, where urbanization and population density are increasing rapidly, traffic-related incidents have emerged as one of the leading causes of mortality and morbidity. In 2023, Türkiye witnessed 1 million 314 thousand 136 traffic accidents among its 1 million 79 thousand 65 accidents were with material loss and 235 thousand 71 were with death or injury [3]. On average, 644 road accidents associated with death or injury occur daily, along with 18 deaths and 961 injuries, which are rapidly growing and motorizing [3]. Addressing this issue requires a comprehensive understanding of historical trends and patterns in traffic accidents, as well as effective predictive tools to anticipate future occurrences and develop proactive measures [4]. The primary problem addressed in this research is the lack of localized and detailed predictive analysis for traffic accidents in Türkiye, specifically in major urban centers such as Istanbul, Ankara, Izmir, Bursa, and Konya. This study distinguishes itself from existing literature by utilizing a city-specific, government-verified dataset that spans multiple years (2021–2024) and provides monthly accident records. Most previous studies have focused on national-level data, limiting their relevance for urban safety planning. Moreover, there is a limited application of machine learning (ML) models for predicting traffic accidents based on historical data in the Turkish context. For this purpose, the following research questions are proposed:

- RQ1: Can ML models effectively predict traffic accidents in major Turkish cities using historical and incident-based attributes?
- RQ2: How do the performances of Random Forest Regressor (RFR) and Linear Regression (LR) models compare in the context of localized traffic accident prediction?

The need for this analysis and prediction stems from the alarming increase in traffic accidents in urban areas, which poses risks to public safety and strains healthcare and emergency response systems. Accurate prediction models can assist policymakers, urban planners, and law enforcement agencies in identifying high-risk periods and areas, thereby enabling targeted interventions to mitigate accidents and improve road safety. To emphasize the study's originality, we collected traffic accident data directly from the Gendarmerie General Command (GGC) of the Ministry of Internal Affairs of the Republic of Türkiye [5], ensuring high reliability and detail. The dataset includes monthly records of accidents, deaths, and injuries reported between 2021 and 2024 across the five cities. Using this data, the study evaluates statistical patterns and then implements and compares two ML models: LR and RFR. These models are assessed using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and R-squared (R^2), providing insights into their predictive accuracy and applicability. The benefits of this research are multifaceted. First, it provides a localized analysis of traffic accident trends in five major Turkish cities, facilitating the development of tailored solutions for each urban area. Second, it underscores the efficacy of ML models in predicting traffic accidents, thereby contributing to the expanding body of literature on artificial intelligence applications in public safety. Finally, the study presents a replicable methodology that can be adapted to other cities or regions, both within Türkiye and globally, to enhance traffic management strategies. The key contributions of this paper are as follows:

- Creation of a monthly, city-level traffic accident dataset from 2021 to 2024 using official Turkish sources.
- A comprehensive analysis of traffic accident trends in Istanbul, Ankara, Izmir, Bursa, and Konya during this period.
- Evaluation and comparison of LR and RFR models for predicting accidents, deaths, and injuries.

The rest of this paper is organized as follows: Section 2 reviews related works. Section 3 describes the methods including data collection, dataset introduction, and prediction models. Section 4 presents results and discussion, and Section 5 concludes with future work.

Related works

Numerous studies have analyzed traffic accidents using ML models. Previous research highlights the importance of data-driven insights to enhance road safety. While studies focus on specific cities or regions, a comprehensive analysis of multiple urban areas in Türkiye remains underexplored. Saffet Erdogan's [6] study analyzed traffic accident and mortality rates across Turkey's provinces using spatial analysis and geographically weighted regression (GWR). The results revealed significant clustering of high accident and mortality rates in regions connecting major provinces like Istanbul, Ankara, and Antalya. GWR provided better predictions compared to ordinary least squares, highlighting its effectiveness in uncovering local patterns for targeted road safety interventions. The findings suggest the need for focused safety measures in high-risk provinces to improve overall road safety management

in Türkiye. Ali Kemal Çelik and Erkan Oktay [7] conducted a retrospective analysis of 11,771 traffic accidents in Erzurum and Kars Provinces, Türkiye, categorizing them into fatal, injury, and no-injury cases. The study used multinomial logic analysis to identify factors such as driver-age, education, and road type that significantly increase the likelihood of fatal injuries. This research is notable for its comprehensive dataset and pioneering application of an unordered response model in Türkiye to analyze traffic injury severity. In their study Sungur et al. [8] highlights Türkiye's significant share of global traffic accident fatalities, stressing the need to prioritize traffic accidents on the public health agenda. They identify the human factor, particularly drivers, as the leading cause of accidents, with issues like speeding, substance use, and fatigue being critical contributors. Despite legislative improvements, the enforcement of these laws remains insufficient, underscoring the need for stronger implementation strategies. Kaygisiz et al. [9] analyze the influence of urban built environments on traffic accidents in Eskişehir, Türkiye, using binary logit models and count data regression. Their findings highlight the significance of road segment properties, traffic flow characteristics, and land use on accident occurrence. The study proposes proactive urban planning measures to enhance traffic safety and emphasizes the need for coordinated urban design and transportation planning. Chukwutoo C. Ihueze and Uchendu O. Onwurah [10] analyzed road traffic accidents in Anambra State, Nigeria, using ARIMA and ARIMAX modeling techniques to predict crash frequency. The ARIMAX model, which incorporated human, vehicle, and environmental factors, demonstrated superior accuracy compared to ARIMA based on multiple performance metrics. Their findings highlight the importance of considering diverse factors in predictive modeling to improve road safety strategies. Ozen's [11] study analyzes traffic safety across Türkiye's seven regions over an 11-year period, utilizing both relative and absolute crash rates. The findings reveal regional stability in rankings despite differences in safety measures and highlight a significant increase in fatal and injury crash rates over time. Kumeda et al. [12] in their study using the UK's 2016 traffic accident dataset, this research explored several ML algorithms, with Fuzzy-FARCHD achieving an 85.94% accuracy. The work demonstrates the role of advanced ML techniques in deciphering complex traffic data for better prediction accuracy. Ali Kemal Erenler and Burak Gümüs [13] explored and analyzed traffic accidents in Türkiye between 2013 and 2017, revealing that male individuals and motorcyclists face higher risks of fatal and non-fatal accidents. The study recommends stricter laws and targeted education for vulnerable groups like two-wheeler and tractor drivers to reduce accidents. Using a simulated traffic accident dataset, Al Mamlook et al. [14] compared six techniques, with Random Forest (RF) emerging as the best-performing model at 82.6% accuracy. The findings underscore the strength of ensemble methods like RF in predictive tasks. This research also illustrates the value of simulated datasets in testing and validating ML models. In another study where Labib et al. [15] analysis of 43,089 traffic accidents from Bangladesh (2001–2015) revealed AdaBoost as the best-performing algorithm with 80% accuracy. This study highlights the importance of region-specific factors in determining the effectiveness of predictive models. Qu et al. [16] done an experiment on a Chinese

city's traffic accidents data from 2006 to 2016 using the integration approach of Genetic algorithm and XGBoost model. The prediction models gain 94% accuracy compared to other models. Alkheder et al. [17] investigated 5,740 traffic accidents in Abu Dhabi using Decision Tree, Bayesian Network, and linear Support Vector Machine models. Among these, the Bayesian Network achieved a modest accuracy of 66.18%. The research highlights the potential and limitations of probabilistic models in accident severity prediction. Yassin and Pooja [18] implemented a hybrid approach combining K-means clustering and RF was used to predict road accident severity, achieving a high accuracy of 99.86%. This study highlights the effectiveness of ensemble methods in handling accident-related datasets and provides a benchmark for ML-based traffic analysis. Chen and Chen [19] examining Taiwan's road traffic data (2015–2019), RF was shown to outperform logistic regression and classification trees. The research affirms the utility of RF in accident severity prediction across different geographies. Sangare et al. [20] employed the integrating version of Gassian mixture modeling and SVM on different size road accidents data points and finally the model acquires 85.53% accuracy. Utilizing RF outpacing classifiers Bokaba et al. [21] investigated 46692 South African road traffic accidents data, and the experiment showed superior accuracy, precision and recall value. Adem Korkmaz [22] proposed leverages ML techniques to predict the severity of traffic accidents in Türkiye, using models like RF, CatBoostClassifier, and LightGBM. RF emerged as the most effective model, emphasizing factors such as engine capacity, driver age, and vehicle type. The findings underscore ML's potential for targeted interventions and sustainable urban planning. Kuyumcu et al. [4] investigated traffic accidents in Türkiye between 2015 and 2021 using CHAID, RF, and Naïve Bayes models to predict drivers' casualty levels. They found that driver fault, gender, education, age, alcohol use, and road surface conditions significantly influenced accident severity, with the CHAID model yielding the most accurate results. Table 1 below summarizes key studies in this domain.

Table 1. *Summary of related works on traffic accident prediction and analysis*

No.	Study location	Methodologies utilized	Data focus/period	Highlights and limitations	Ref.
1	Türkiye (Provinces)	GWR, Spatial Analysis	National level	Emphasized clustering, no ML-based prediction	[6]
2	Erzurum & Kars	Multinomial Logistic Regression	11.771 cases	Factor analysis, no forecasting	[7]
3	Türkiye (General)	Descriptive Analysis	General policy level	Emphasized human factors, no predictive modeling	[8]
4	Eskişehir, Türkiye	Binary Logit, Regression Models	Urban road segments	Analyzed built environment influence, no accident prediction	[9]
5	Nigeria (Anambra State)	ARIMA, ARIMAX	Time-series	Time-based modeling, not ML-centric	[10]
6	Türkiye (7 Regions)	Crash rate comparisons	11 years	No ML models used	[11]
7	UK	Fuzzy-FARCHD, Others	2016	ML comparison, focused on accuracy only	[12]
8	Türkiye (2013–2017)	Statistical Summary	Fatal vs non-fatal analysis	Identified at-risk groups, no prediction	[13]

Table 1. ...continued

9	Simulated Dataset	RF, others	Simulated	Found RF best, but not real data	[14]
10	Bangladesh	AdaBoost, Others	2001–2015	AdaBoost was best, regional study	[15]
11	China	GA + XGBoost	2006–2016	94% accuracy, lacks interpretability	[16]
12	UAE (Abu Dhabi)	DT, BN, SVM	5.740 cases	Bayesian network limited to 66.18% accuracy	[17]
13	India	K-means + RF	NA	99.86% accuracy using hybrid ensemble	[18]
14	Taiwan	RF, Logistic Regression, Decision Tree	2015–2019	RF outperformed others	[19]
15	Côte d'Ivoire	GMM + SVM	Various	Achieved 85.53% accuracy	[20]
16	South Africa	RF	46,692 cases	High recall and accuracy, no comparative models used	[21]
17	Türkiye	RF, CatBoost, LightGBM	NA	Focus on severity; compared only classifiers	[22]
18	Türkiye (2015–2021)	CHAID, RF, Naïve Bayes	Nationwide driver data	No city-wise or temporal trend analysis	[4]
19	5 Major Cities in Türkiye	LR, RF Regression	Monthly data (2021–2024)	City-based, interpretable regression, government data, ML-based forecasting	Our Study

Material and Method

This section describes the systematic approach undertaken to achieve the research objectives and validate the findings. It outlines the data collection process and dataset preprocessing, experimental design, prediction models, and computational tools employed to ensure the reliability and reproducibility of the results.

Data Collection and Dataset Description

The data collection process involved gathering relevant and comprehensive statistical data from reliable sources to ensure accuracy of the analysis. For this study, traffic accident data spanning several years was obtained from the official database of the GGC of the Ministry of Internal Affairs of Türkiye [5] from 2021 to 2024, which provides access to well-structured and detailed records. The dataset includes yearly accident statistics (2021–2024) of each month for Istanbul, Ankara, Izmir, Bursa, and Konya. Each dataset consists of 48 samples (rows) and 6 attributes (columns) and is named by corresponding city name. Summary of these datasets presented in Table 2. To ensure consistency and comparability, each dataset underwent standard preprocessing, which included checking for missing or anomalous values, formatting date entries uniformly, and normalizing numeric ranges where applicable. The selected variables, Death and Injury Accidents (DIAc), Property Damage Accidents (PDAc), Deaths, and Injuries, were chosen due to their frequent use in prior traffic prediction research [4, 7, 22] and their direct relation to accident severity. These variables were used as predictors, while Total Accidents was used as the target variable. The data were then split into training and testing sets with an 80%-20% ratio,

which is standard practice to assess model generalizability. No oversampling or under sampling was applied as the dataset was already balanced across months and cities.

Table 2. *Summary of Datasets*

Names of Datasets	Ankara, Istanbul, Izmir, Bursa, Konya
Number of samples (rows)	48
Number of Attributes (columns)	6
Attributes name	Year, Month, DIAc, PDAc, Deaths, and Injuries
Source	GGC of the Ministry of Internal Affairs of Türkiye's official website [5].

Prediction Models

This subsection describes the implementation of two prediction models: the Random Forest Regressor (RFR) and Linear Regression (LR), to estimate traffic accident keenness. Both models were applied to the prepared dataset, and their performance was evaluated using key regression metrics, including MAE, MSE, and R^2. These metrics provided a comprehensive assessment of each model's accuracy, error rates, and explanatory power in predicting accident strength. To improve reproducibility, we used Scikit-learn's standard implementation for both algorithms. All models were trained using default hyperparameters, without tuning, to focus on baseline performance comparison.

Random forest regressor (RFR): Random Forest (RF) is an ensemble model comprising K decision trees, denoted as $h\ (x;\ \theta_k)$, where x represents the input vector of length p, and θ_k are independent and identically distributed random parameters. In regression, where the outcome Y is numerical, the prediction of the RF is the unweighted average of all trees as follows [23, 24]:

$$\bar{h}(x) = \frac{1}{K}\sum_{K=1}^{K} h\ (x;\ \theta_k) \tag{1}$$

As $K \to \infty$, the Law of Large Numbers ensures that the prediction error converges as follows:

$$E_{XY}\left[\left(Y - \bar{h}\ (X)\right)^2\right] = E_{XY}\left[Y - E_\theta h\ (X;\theta)\right]^2, \tag{2}$$

where the right-hand side represents the generalization error PE_f . The average prediction error for an individual tree $h\ (X;\theta)$ theta) is defined as:

$$PE_t\ = E_\theta[E_{XY}\left[(Y - h\ (X;\theta))^2\right]] \tag{3}$$

Assuming trees are unbiased ($E_Y\left[E_X[h\ (X;\theta)]\right] = Y$), the RF's total error can be expressed as:

$$PE_f\ = \rho \cdot PE_t\ , \tag{4}$$

where ρ is the weighted correlation between residuals of different trees. This formulation demonstrates how Random Forest achieves robust predictions while minimizing the risk of overfitting. The Random Forest Regressor operates based on Bootstrap Aggregating, commonly known as Bagging. Each decision tree is trained using a randomly selected subset of data and features, which helps prevent overfitting. The final prediction is obtained by averaging the predictions of all individual trees, following

the approach outlined in Equation (1). We selected this model because of its robustness, ability to handle multicollinearity, and superior performance in prior studies on accident severity and prediction [13, 14, 19, 21, 22].

Linear regression (LR): Linear Regression is a statistical method used for predictive modeling and understanding the relationship between dependent (target) and independent (predicator) variables. The general single-equation LR model is a mathematical representation used to understand the relationship between a dependent variable (Y) and one or more independent variables (X_i). It encompasses both simple LR (with one independent variable) and multiple LR (with multiple independent variables). The equation is expressed as follows [25, 26]:

$$Y = a + \sum_{i=1}^{k} b_i X_i + u, \quad (5)$$

where, Y is the dependent variable (in this case total accidents), whose value we aim to predict or explain,$X_1, X_2, \ldots, X_k$ (in this case deaths or injuries) is the independent variables that influence is the intercept Y, a representing the expected value of Y when all $X_i = 0$, b_i is the regression coefficients, quantifying the change in Y for a unit change in X_i, assuming other variables remain constant, u is the stochastic disturbance term, capturing unexplained variations in Y due to omitted variables, measurement errors, or random factors. Linear regression was chosen as a baseline due to its interpretability, simplicity, and wide use in early accident prediction research [6, 9, 19]. Although k-fold cross-validation is commonly used to improve generalization and reduce model variance, it was not applied in this study due to the relatively small size of each city's dataset (48 samples). Instead, a fixed 80/20 train-test split was chosen to maintain simplicity and comparability across cities. However, future work will include cross-validation to further enhance robustness and reproducibility. This methodology ensures transparency and reproducibility by clearly describing the data preparation, feature selection rationale, modeling approach, and evaluation procedures. The pipeline in Figure 1 enhances the clarity of the research process.

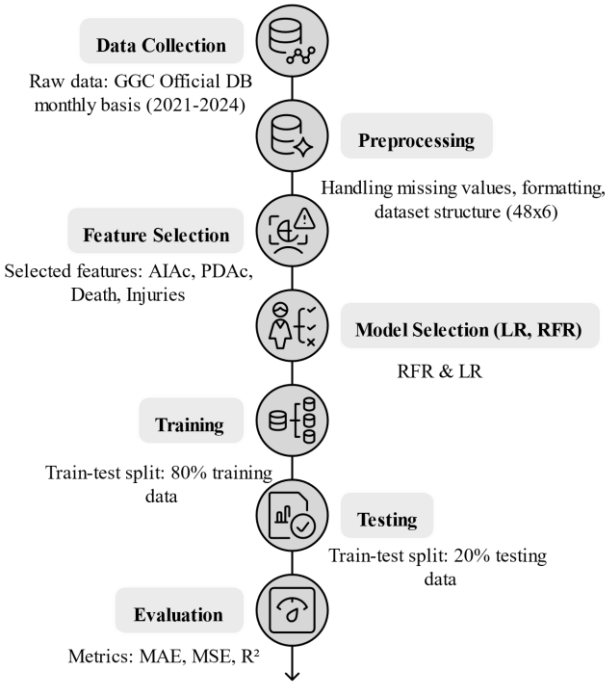

Figure 1. *Pipeline of the proposed methodology highlighting each step from data acquisition to evaluation*

Results and Discussion

The dataset was split into training (80%) and testing (20%) subsets to evaluate the performance of the predictive models. The analysis was conducted in a Google Colab pro version environment utilizing Python libraries such as Pandas for data manipulation, Matplotlib and Seaborn for visualization, and Scikit-learn for implementing ML models. The monthly and yearly accident trends were examined and then visualized to identify significant patterns and seasonal variations, which could influence the predictive accuracy of the models. Default parameters were applied during the initial stages of model training to ensure a baseline for comparison while maintaining reproducibility. The evaluation metrics, including MAE, MSE, and R^2 were calculated to assess model accuracy.

Monthly and Yearly Accident Trends

The total accidents in a month computed from the number of death-injury accidents and property-damage accidents summation. The graph of the total accident, death, and injured trends of each month from 2021-2024 for five cities is shown in Figure 2, Figure 3, and Figure 4 respectively. These graphs come from the datasets that were named by each city name and visualize the traffic accidental information of each city on a monthly basis. The results presented in Tables 3 and 4 provide valuable sagacity into the trends in accidents across five major cities in Türkiye from 2021 to 2024. Table 3 highlights the yearly statistical data for different attributes, including total DIAc, PDAc, deaths, and injured individuals. Ankara consistently records the highest number of accidents and injured individuals across the years, with a noticeable increase in total DIAc and total Injured from 2021 to 2024, indicating a worsening trend. Similarly, Izmir exhibits significant numbers for Total Injured and DIAc, with a

steady rise in fatalities, peaking in 2024 with 43 deaths. In contrast, Istanbul has relatively lower accident figures although it is the largest metropolitan in Türkiye in terms of area and population. However, it remains consistent in its fatality count, with no significant changes over the years. Bursa shows fluctuating propensity, particularly in DIAc and PDAc, while mortality and injuries remain comparatively lower than other cities. Konya displays a rigid increment in total injuries and a slight variation in DIAc and fatalities, indicating an overall moderate level of accidents compared to the larger metropolitan areas. Table 4 provides a consolidated view of total accidents in each city over the four years. Ankara and Izmir emerge as the most accident-prone cities, with consistently high totals across the years, while Bursa and Konya report comparatively fewer accidents. Interestingly, Istanbul shows a slight decrease in total accidents in 2024, which may suggest some improvement in road safety measures. These trends underline the necessity for targeted interventions and stricter enforcement of traffic regulations, particularly in Ankara and Izmir, to curb the growing number of accidents and injuries. The corresponding graph views of Table 2 and 4 depicted in Figure 4. Furthermore, each year total accidents of those five cities indicated on the map in Figure 5.

Table 3. *Yearly statistical data of five cities from 2021 to 2024 based on datasets*

City (Dataset)	Attributes	Year			
		2024	**2023**	**2022**	**2021**
Ankara	DIAc	1130	1043	874	905
	PDAc	1652	1279	1432	1384
	Death	41	27	22	22
	Injured	2092	1881	1622	1576
Istanbul	DIAc	595	633	510	487
	PDAc	1312	1581	1411	1400
	Death	15	15	22	14
	Injured	956	1029	849	821
Izmir	DIAc	1651	1377	1244	1315
	PDAc	1019	1039	1245	1327
	Death	43	37	38	44
	Injured	2568	2149	1950	1977
Bursa	DIAc	686	805	716	716
	PDAc	567	489	689	689
	Death	15	30	27	27
	Injured	1294	1488	1227	1227
Konya	DIAc	946	815	733	723
	PDAc	492	567	420	372
	Death	30	31	35	22
	Injured	1846	1587	1457	1246

Table 4. *Total Accidents in Major Cities (2021-2024)*

Year	Cities				
	Ankara	**Istanbul**	**Izmir**	**Bursa**	**Konya**
2024	2782	1907	2670	1253	1438
2023	2322	2214	2416	1294	1382
2022	2306	1921	2489	1385	1153
2021	2289	1887	2642	1348	1095

Evaluation Metrics

Model performance was assessed using following evaluation metrics:

Mean Absolute Error (MAE): MAE measures the average magnitude of errors between predicted and actual values, providing a straightforward assessment of prediction accuracy [27]. It is less sensitive to outliers than MSE and is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{6}$$

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the total number of observations.

Mean Squared Error (MSE): MSE evaluates the average squared difference between predicted and actual values. By squaring the errors, it penalizes larger deviations more heavily than MAE [27].

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{7}$$

R-squared (R^2): R^2 quantifies how well the model explains the variance in the actual data, ranging from 0 to 1, where 1 indicates a perfect fit [28], where, y_i is the mean of the actual values.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{8}$$

Figure 2. *Traffic accident trends per month (2021-2024). (a) Ankara, (b) Istanbul, (c) Izmir, (d) Bursa, (e) Konya*

Figure 3. *Accidental death trends per month (2021-2024). (a) Ankara, (b) Istanbul, (c) Izmir, (d) Bursa, (e) Konya*

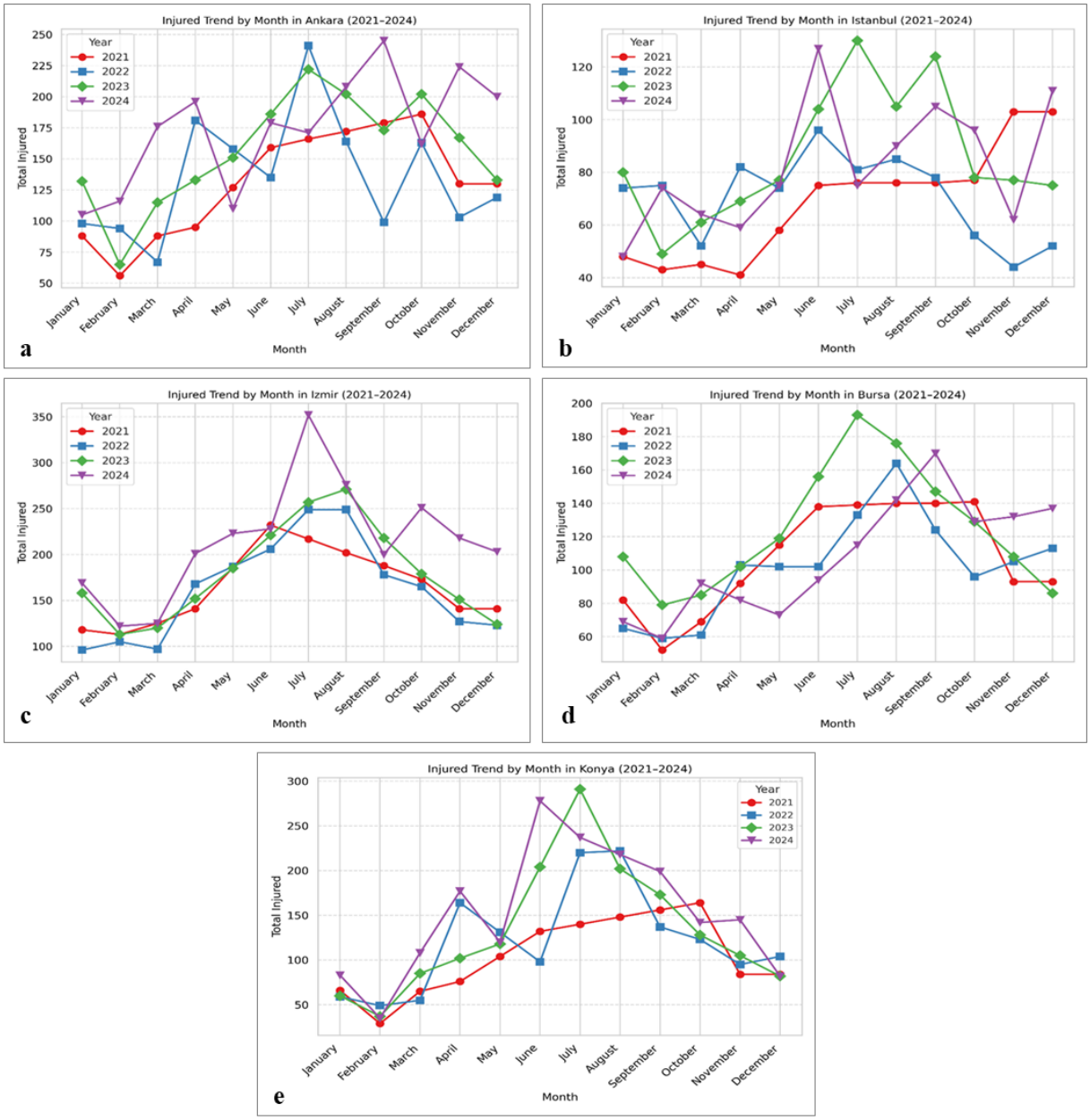

Figure 4. *Accidental injury trends per month (2021-2024). (a) Ankara, (b) Istanbul, (c) Izmir, (d) Bursa, (e) Konya*

Predictive Model Performance

The results presented in Table 5 highlight the effectiveness of the RFR and LR models in predicting accidents, deaths, and injuries across five cities. Both models demonstrate excellent predictive capabilities, as evidenced by high R^2 values and favorable error metrics. The RFR model consistently delivers high accuracy across all cities, particularly in Istanbul and Bursa, where R^2 values exceed 0.92 for all metrics. These values indicate a strong alignment between the predicted and actual data. In Ankara, RFR achieves R^2 scores of 0.84, 0.91, and 0.96 for accidents, deaths, and injuries, respectively, showcasing its reliable performance in forecasting these critical outcomes. The relatively low MAE and MSE values further confirm the model's ability to generate precise predictions. Similarly, the LR model

achieves perfect R^2 scores of 1.0 in all cities and for all metrics, indicating its strong ability to fit the given data accurately and provide reliable predictions. This consistency is particularly evident in cities like Istanbul and Bursa, where the LR model's performance mirrors the actual data trends closely.

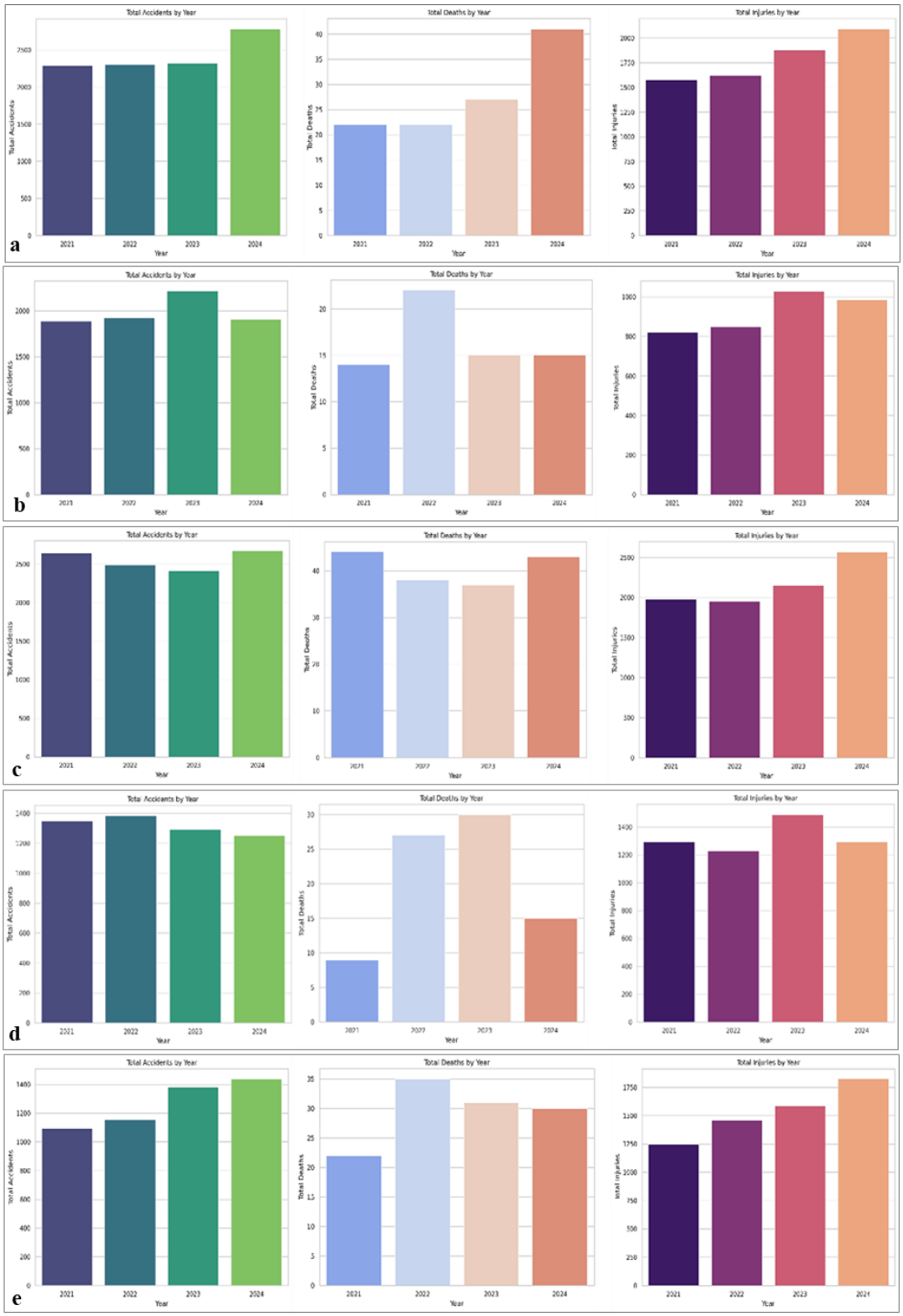

Figure 5. *Yearly accidents, deaths and injury (2021-2024). (a) Ankara, (b) Istanbul (c) Izmir (d) Bursa (e) Konya*

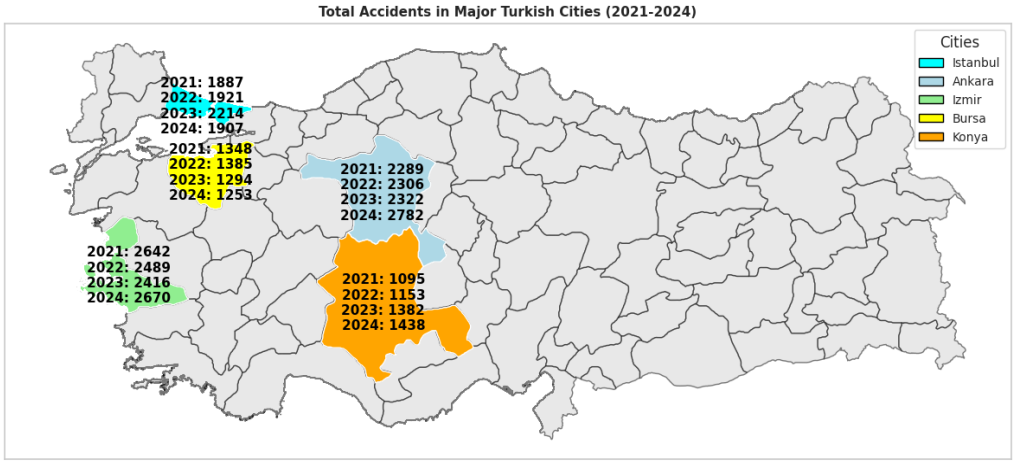

Figure 6. *Map view of total accidents of five major cities in Türkiye according to year (2021-2024)*

In Istanbul, both models excel in predicting outcomes, with RFR demonstrating remarkable precision. The predictions for deaths and injuries are particularly noteworthy, supported by near-perfect R^2 values and minimal errors. Similarly, in Izmir, the RFR model provides highly accurate results for accidents and deaths, reflecting its adaptability to varying patterns in the data. Moreover, the LR model demonstrates its adaptability by delivering accurate predictions with minimal error rates, particularly for deaths and injuries. For Bursa, both models perform exceptionally well, achieving almost perfect R^2 values. This consistency across metrics underlines the models' capability to handle diverse datasets effectively. The RFR model's ability to predict injuries and accidents with low error rates further emphasizes its reliability. In Konya, the RFR model exhibits outstanding performance, with high R^2 values and accurate predictions for all metrics. The results demonstrate the model's proficiency in capturing the unique characteristics of the city's data, ensuring dependable forecasting outcomes. Additionally, the LR model exhibits robust performance for Bursa and Konya, achieving perfect R^2 values and maintaining low error rates, which underscores its effectiveness across different datasets. Overall, the predictive models effectively highlight trends and patterns across the five cities. The RFR model, with its robust performance, stands out as a reliable tool for analyzing and forecasting critical traffic-related statistics. Tables 7 and 8 present the test set prediction values for the RFR and LR models across five cities, highlighting their capability to estimate accidents, deaths, and injuries. Both models exhibit strong alignment with the datasets, with the RFR predictions demonstrating more nuanced values across varying test cases. The RFR model provides detailed estimates for all metrics, reflecting its ability to account for complex patterns in the data. Conversely, the LR model produces consistent predictions, with some test values showing notable approximations to the observed trends, particularly in metrics like accidents and injuries. For example, in Ankara, both models predict accidents and injuries with high accuracy, while Istanbul and Izmir show robust predictions in deaths and injuries, demonstrating the adaptability of these models to different urban environments. Bursa and Konya also reflect the models' efficiency in capturing regional variations in traffic-related incidents. The table 6 and 7 corresponding

actual vs predicted accident, death and injured line and bar chart graph of Ankara, Istanbul, Izmir, Bursa and Konya presented in Figure 7, 8, 9, 10, and 11. These results underline the potential of predictive modeling to analyze urban traffic data effectively.

Table 5. *Comprehensive statistical results of the two prediction models for five cities*

City	Metric	RFR model			LR model		
		Accidents	**Deaths**	**Injuries**	**Accidents**	**Deaths**	**Injuries**
Ankara	MAE	9.16899	0.25899	5.37400	2.70006	3.94690	1.13686
	MSE	161.94572	0.38090	79.79678	1.23188	3.946900	2.42338
	R^2	0.84	0.91	0.96	1.0	1.0	1.0
Istanbul	MAE	6.797	0.01800	0.72999	1.56319	5.96613	1.42108
	MSE	77.67229	0.003240	1.47407	5.04870	5.82339	3.33214
	R^2	0.87	0.99	0.99	1.0	1.0	1..0
Izmir	MAE	10.60100	0.37999	8.097	2.84217	2.25072	2.700062
	MSE	174.40387	0.75552	111.58575	1.29246	7.19195	9.49157
	R^2	0.82	0.90072	0.96	1.0	1.0	1.0
Bursa	MAE	4.79999	0.02099	5.02699	1.84741	3.94859	2.13162
	MSE	38.28389	0.00150	41.24024	5.04870	2.15611	1.51461
	R^2	0.92	0.99	0.97	1.0	1.0	1.0
Konya	MAE	3.35300	0.07799	6.69400	5.32907	5.90167	5.82645
	MSE	36.65263	0.02241	61.53950	3.82187	4.79110	4.55393
	R^2	0.98	0.99210	0.97	1.0	1.0	1.0

Table 6. *Specific test set prediction value of RFR models for five cities*

City	Metric	Test Set									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Ankara	Accidents	211.63	157.02	166.68	215.33	223.90	144.81	241.24	202.68	189.45	207.01
	Deaths	4.55	2	1	2.02	0	1	5.11	4.82	2.01	4.04
	Injuries	176.06	129.67	86.23	172.03	95.74	75.53	132.44	200.23	109.72	96.59
Istanbul	Accidents	187.29	148.28	166.03	159.62	166.5	148.01	185.45	193.51	138.2	148.68
	Deaths	1	2	1	1	3.82	1	1	0	0	2
	Injuries	81.23	58.11	50.95	75.89	74.21	46.43	79.72	104.79	75.09	73.96
Izmir	Accidents	169.86	226.56	190.69	256.71	197.47	180.43	204.11	255.36	221.03	188.87
	Deaths	4	4	0.53	4.9	0.43	2.01	2.01	7.34	4	2.06
	Injuries	163.44	185.91	116.5	206.68	117.2	117.65	151.1	267.22	219.56	117.17
Bursa	Accidents	109.79	93.59	97.03	128.27	102.68	89.96	121.19	143.42	76.61	76.62
	Deaths	1	2.01	2.01	1	3.91	0	3.92	2	4.02	1
	Injuries	102.04	112.85	69.44	140.45	71.61	64.58	107.18	168.53	70.67	67.46
Konya	Accidents	106.67	89.41	47.46	117.58	68.6	48.95	63.81	187.07	94.43	47.77
	Deaths	4.16	2	3	2.99	1.99	0.03	4.05	4.74	2	0.03
	Injuries	155.57	105.93	52.45	142.64	67.83	44.99	65.84	212.42	124.14	45.55

Table 7. *Specific test set prediction value of LR models for five cities*

City	Metric	Test set									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Ankara	Accidents	200	160	161	216	216	115	230	204	188	188
	Deaths	5	2	1	2	0	1	7	5	2	4
	Injuries	181	127	67	172	98	56	132	202	110	94
Istanbul	Accidents	205	145	162	159	167	146	179	207	126	141
	Deaths	1	2	1	1	4	1	1	0	0	2
	Injuries	82	58	52	76	74	43	80	105	75	75
Izmir	Accidents	173	223	162	264	199	172	193	244	201	178
	Deaths	4	4	0	5	0	2	2	1	4	2
	Injuries	168	186	97	202	96	113	158	271	223	105
Bursa	Accidents	100	98	94	131	105	76	116	148	75	77
	Deaths	1	2	2	1	4	0	4	2	4	1
	Injuries	103	115	61	140	65	52	108	176	73	59
Konya	Accidents	107	88	47	117	75	43	63	204	94	48
	Deaths	5	2	3	3	2	0	4	5	2	0
	Injuries	164	104	55	148	59	29	60	202	120	49

Discussion

The findings provide a comprehensive understanding of traffic incidents and predictive modeling across five major cities from 2021 to 2024. The accident trends reveal an overall fluctuation in the number of accidents, injuries, and deaths, with variations across years and cities, as shown in Tables 3 and 4. Ankara consistently exhibited the highest number of incidents, while Bursa and Konya experienced comparatively fewer accidents. Predictive models demonstrated strong performance, with the RFR offering high precision across all metrics, while the LR model delivered consistent, albeit simpler, predictions. The test set predictions for both models aligned closely with observed data, reflecting their reliability. These findings underline the utility of data-driven approaches in analyzing traffic trends, facilitating targeted interventions to enhance road safety and urban planning. In this study, the RFR outperformed LR in predicting traffic accidents across all five cities, demonstrating higher R^2 values and lower error metrics (MAE and MSE). These results are consistent with prior studies such as [22] and [4], where RF models provided superior predictive accuracy in accident prediction tasks compared to traditional regression methods. For instance, Korkmaz [22] found that RF yielded the highest performance among several models (RF, CatBoost, LightGBM) for traffic severity prediction in Türkiye. Similarly, Kuyumcu et al. [4] showed that RF provided reliable results in predicting casualty levels in Turkish traffic accident data. Unlike some studies that focus primarily on classification (e.g., predicting severity levels), our work is based on regression analysis to predict accident numbers, which fills a different gap in the literature. Nonetheless, the consistent advantage of RF across both classification and regression contexts underlines its robustness for traffic-related ML tasks.

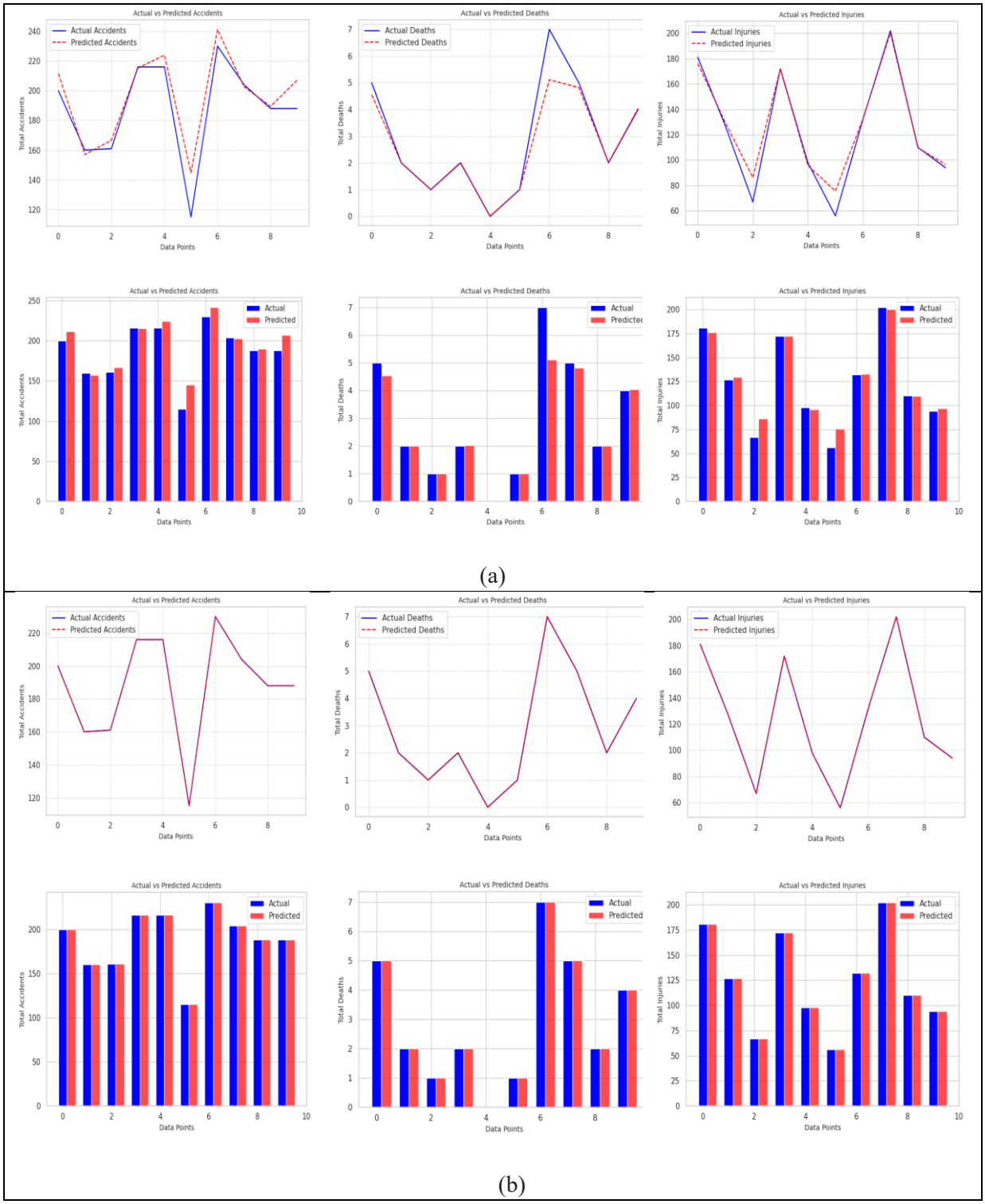

Figure 7. *Actual vs predicted accident, death and injured line and bar chart graph of Ankara (a) RFR (b)LR*

Figure 8. *Actual vs predicted accident, death and injured line and bar chart graph of Istanbul (a) RFR (b)LR*

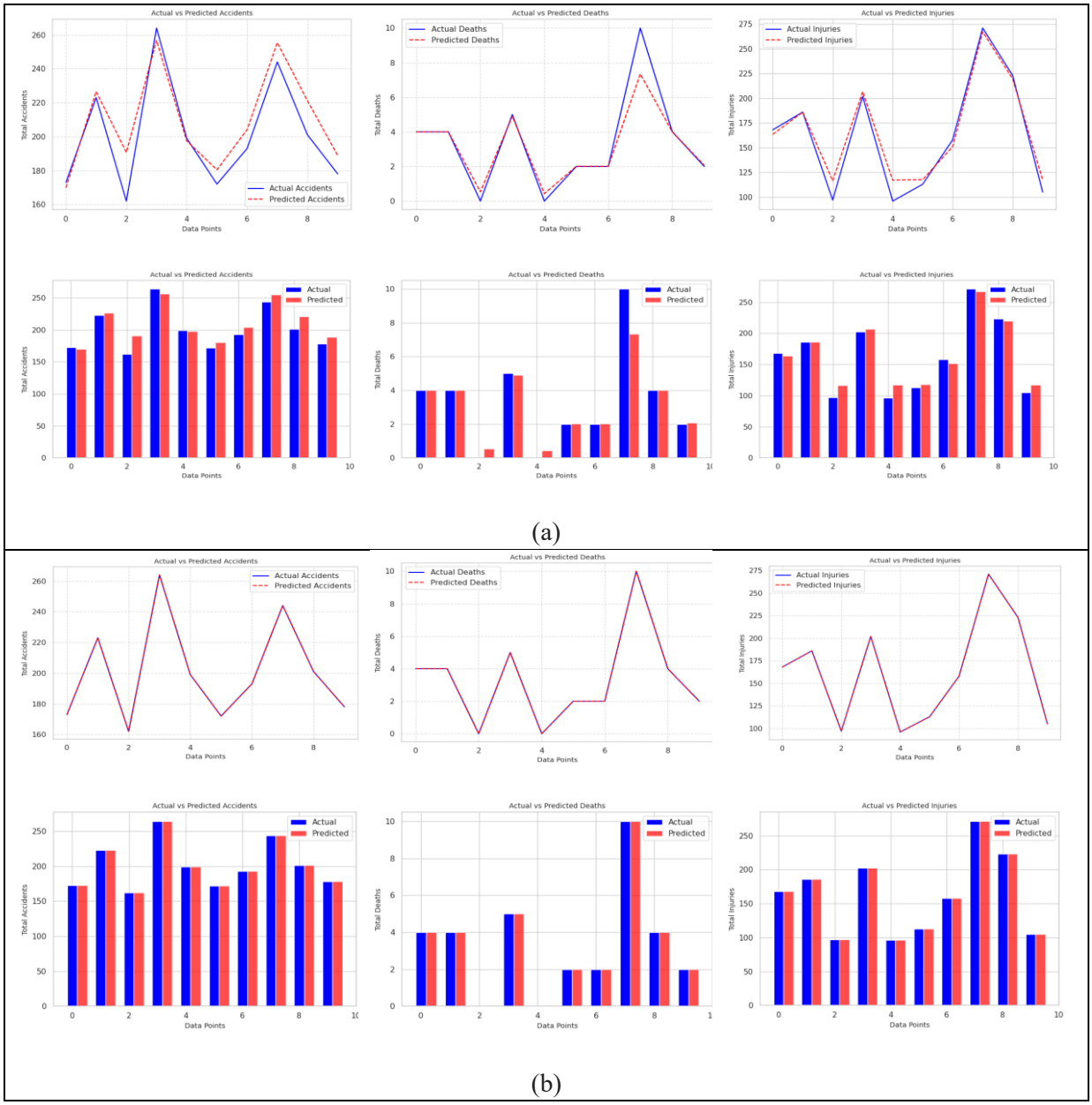

Figure 9. *Actual vs predicted accident, death and injured line and bar chart graph of Izmir (a) RFR (b)LR*

E-ISSN: 2564-7873

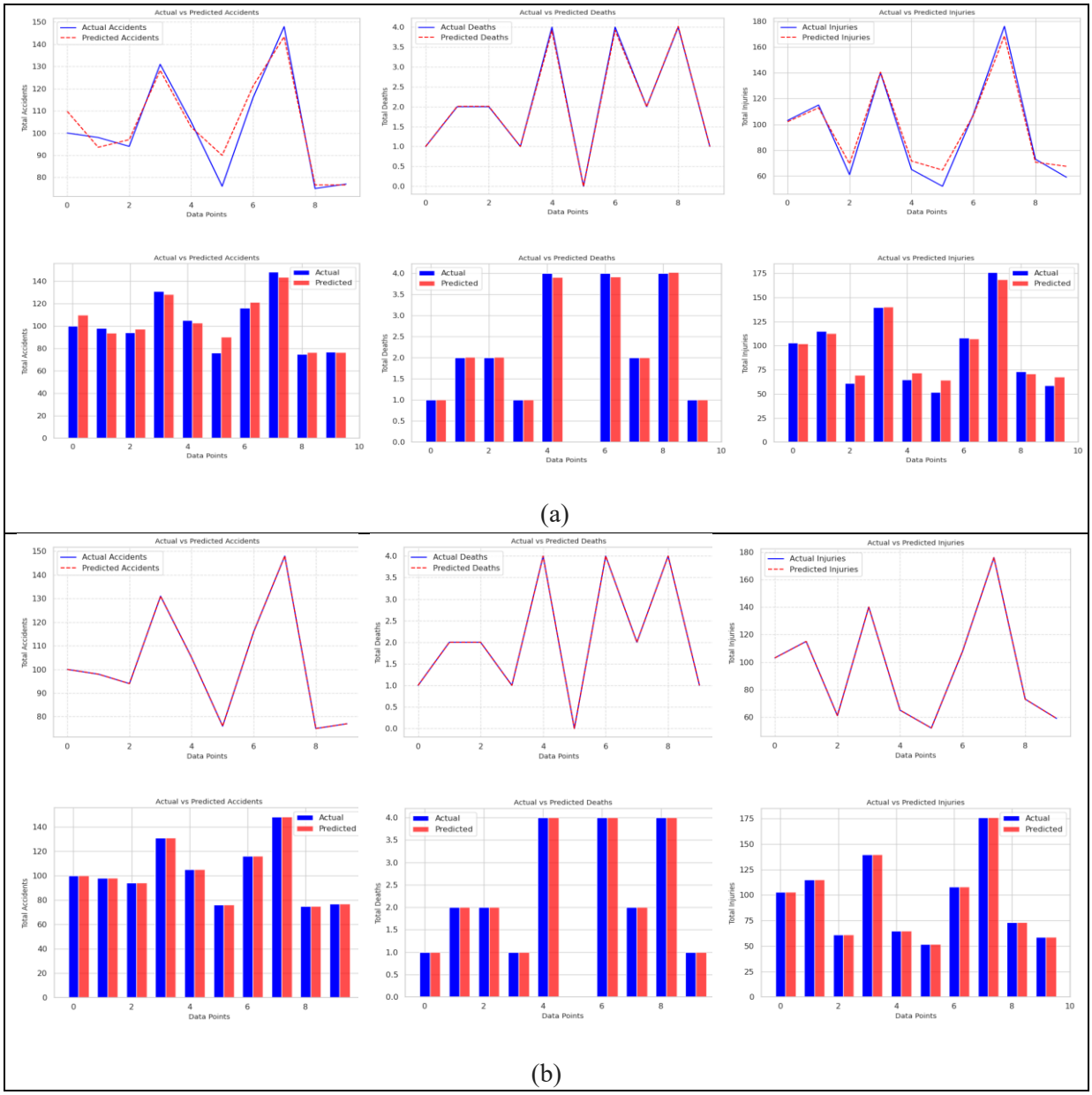

Figure 10. *Actual vs predicted accident, death and injured line and bar chart graph of Bursa (a) RFR (b)LR*

Figure 11. *Actual vs predicted accident, death and injured line and bar chart graph of Bursa (a) RFR (b)LR*

Conclusion and Future Work

The study analyzes traffic accident data from five major cities in Türkiye from 2021 to 2024, revealing significant trends in accidents, injuries, and fatalities, with Ankara consistently exhibiting the highest number of incidents, totaling 1130 accidents in 2024 alone. The linear regression and random forest regressor models are employed to evaluate accident patterns. Findings indicate significant variations in accidents across cities, months, and years, with machine learning models showing high predictive accuracy. The RFR model outperforms LR in evaluation metrics, emphasizing its effectiveness. Predictive modeling using RFR demonstrated high precision, achieving an R^2 value of 0.84 for accidents, 0.91 for deaths, and 0.96 for injuries in Ankara, indicating strong predictive capabilities.

The performance superiority of the RFR model aligns with findings in recent studies, which highlight its robustness in accident severity and frequency prediction. Unlike some studies focused primarily on spatial or environmental correlations, our model-centered approach emphasizes temporal forecasting with monthly granularity. This implies that ML, particularly ensemble-based models, offers a promising direction for real-time accident forecasting and urban safety planning. The findings can assist transportation authorities and policymakers in identifying high-risk months and allocating resources accordingly. Future work should explore the integration of additional variables, such as weather and traffic conditions, to enhance prediction accuracy. Moreover, deep learning models such as LSTM or CNN could be applied for time-series prediction, which may improve performance over traditional ML methods. Expanding the dataset to include more regions and longer time periods could provide deeper insights into accident patterns. Further comparison with other algorithms such as SVR and XGBoost is also recommended to broaden the scope of model performance analysis. The findings underscore the importance of data-driven strategies for improving road safety, suggesting that authorities can utilize these insights to develop targeted interventions to reduce traffic incidents.

Acknowledgements -

Funding/Financial Disclosure The authors declare that they have no financial interests or relationships pertaining to the publication of this article.

Ethics Committee Approval and Permissions The study does not require ethics committee approval or any special permission.

Conflicts of Interest The authors declared no conflict of interest.

Authors Contribution All authors read and approved the final manuscript. The authors contributed equally to the work.

References

[1] Aygencel, G., Karamercan, M., Ergin, M. & Telatar, G. (2008). Review of traffic accident cases presenting to an adult emergency service in Turkey. *Journal of Forensic and Legal Medicine*, *15*(1), 1-6. https://doi.org.10.1016/j.jflm.2007.05.005.

[2] World Health Organization. (2023). Global status report on road safety 2023: Summary (Licence: CC BY-NC-SA 3.0 IGO). World Health Organization. https://www.who.int/teams/social-determinants-of-health/safety-and-mobility/global-status-report-on-road-safety-2023.

[3] Turkish Statistical Institute. (2024). Road traffic accident statistics, 2023 (Issue 53479). https://data.tuik.gov.tr/Bulten/Index?p=Road-Traffic-Accident-Statistics-2023-53479&dil=2

[4] Kuyumcu, Z. C., Aslan, H., & Yurtay, N. (2024). Casualty analysis of the drivers in traffic accidents in Turkey: A CHAID decision tree model. Applied Sciences, 14(24), 11693. https://doi.org/10.3390/app142411693

[5] Gendarmerie General Command. (2025, January). *Aylık istatistik bültenleri (Monthly statistical bulletins).* https://www.jandarma.gov.tr/veriler.

[6] Erdogan, S. (2009). Explorative spatial analysis of traffic accident statistics and road mortality among the provinces of Turkey. *Journal of Safety Research, 40*(5), 341-351. https://doi.org.10.1016/j.jsr.2009.07.006.

[7] Celik, A. K., & Oktay, E. (2014). A multinomial logit analysis of risk factors influencing road traffic injury severities in the Erzurum and Kars Provinces of Turkey. *Accident Analysis & Prevention, 72*, 66–77. https://doi.org/10.1016/j.aap.2014.06.010

[8] Sungur, İ., Akdur, R., & Piyal, B. (2014). Analysis of traffic accidents in Turkey, *Ankara Medical Journal*, *14*(3), 114-124.

[9] Kaygisiz, Ö., Senbil, M., & Yildiz, A. (2017). Influence of urban built environment on traffic accidents: The case of Eskisehir (Turkey). *Case Studies on Transport Policy*, *5*(2), 306-313. https://doi.org.10.1016/j.cstp.2017.02.002.

[10] Ihueze, C. C., & Onwurah, U. O. (2018). Road traffic accidents prediction modelling: An analysis of Anambra State, Nigeria. *Accident Analysis & Prevention*, 112, 21-29. https://doi.org.10.1016/j.aap.2017.12.016.

[11] Özen, M. (2018). Comparative study of regional crash data in Turkey. *Turkish Journal of Engineering*, *2*(3), 113-118. https://doi.org.10.31127/tuje.385008.

[12] Kumeda, B., Zhang, F., Zhou, F., Hussain, S., Almasri, A., & Assefa, M. (2019). Classification of road traffic accident data using machine learning algorithms. *IEEE 11th International Conference on Communication Software and Networks,* 682-687, https://doi.org.10.1109/ICCSN.2019.8905362.

[13] Erenler, A. K., & Gumus, B. (2019). Analysis of road traffic accidents in Turkey between 2013 and 2017. *Medicina (Kaunas)*, *55*(10), 679. https://doi.org.10.3390/medicina55100679.

[14] Al Mamlook, R. E., Ali, A., Hasan, R. A., & Kazim, H. A. M. (2019). Machine learning to predict the freeway traffic accidents-based driving simulation. *IEEE National Aerospace and Electronics Conference,* 630-634. https://doi.org.10.1109/NAECON46414.2019.9058268.

[15] Labib, M. F., Rifat, A. S., Hossain, M. M., Kumar, A., & Nawrine, F. (2019). Road accident analysis and prediction of accident severity by using machine learning in Bangladesh. *7th International Conference on Smart Computing & Communications,* 1-5. https://doi.org.10.1109/ICSCC.2019.8843640.

[16] Qu, Y., Lin, Z., Li, H., & Zhang, X. (2019). Feature recognition of urban road traffic accidents based on GA-XGBoost in the context of big data. *IEEE Access, 7*, 170106-170115, https://doi.org.10.1109/access.2019.2952655.

[17] AlKheder, S., AlRukaibi, F., & Aiash, A. (2020). Risk analysis of traffic accidents' severities: An application of three data mining models. *ISA Transactions*, 106, 213-220. https://doi.org.10.1016/j.isatra.2020.06.018.

[18] Yassin, S. S., & Pooja, R. (2020). Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. *SN Applied Sciences, 2*(9), Article 1560. https://doi.org.10.1007/s42452-020-3125-1.

[19] Chen M. M., & Chen, M. C. (2020). Modeling road accident severity with comparisons of logistic regression, decision tree and random forest. *Information*, *11*(5), 270. https://doi.org.10.3390/info11050270.

[20] Sangare, M., Gupta, S., Bouzefrane, S., Banerjee, S., & Muhlethaler, P. (2021). Exploring the forecasting approach for road accidents: Analytical measures with hybrid machine learning. *Expert Systems with Applications*, *167*, 113855. https://doi.org.10.1016/j.eswa.2020.113855.

[21] Bokaba, T., Doorsamy, W., & Paul, B. S. (2022). Comparative study of machine learning classifiers for modelling road traffic accidents. *Applied Sciences*, *12*(2), 828. https://doi.org.10.3390/app12020828.

[22] Korkmaz, A. (2023). Predictive modeling of urban traffic accident severity in Türkiye's centennial: machine learning approaches for sustainable cities. *Kent Akademisi*, *16*, 395-406. https://doi.org.10.35674/kent.1353402.

[23] Segal, M. R. (2004). Machine learning benchmarks and random forest regression. *UCSF: Center for Bioinformatics and Molecular Biostatistics*, https://escholarship.org/uc/item/35x3v9t4.

[24] Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5-32. https://doi.org/10.1023/A:1010933404324.

[25] Poole M. A., & O'Farrell, P. N. (1971). The assumptions of the linear regression mode. *Transactions of the Institute of British Geographers*, *52*, 145-158. https://doi.org/10.2307/621706.

[26] Aalen, O. O. (1989). A linear regression model for the analysis of life times. Stat Med, *8*(8), 907-925. https://doi.org.10.1002/sim.4780080803.

[27] Twomey, J. M., & Smith, A. E. (1995). Performance measures, consistency, and power for artificial neural network models. *Mathematical and Computer Modelling*, *21*(1-2), 243-258.

[28] Serefoglu Cabuk, K., Cengiz, S. K., Guler, M. G., Topcu, H., Cetin Efe, A., Ulas, M. G., & Poslu Kandemir, F. (2024). Chasing the objective upper eyelid symmetry formula; R^2, RMSE, POC, MAE, and MSE. *International Ophthalmology*, *44*(1), 303, https://doi.org.10.1007/s10792-024-03157-y.