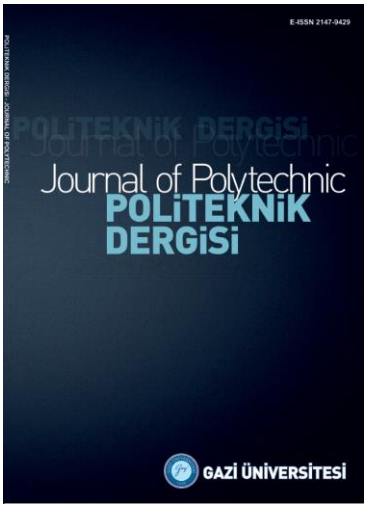

POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

Harnessing machine learning to enhance global road safety: A comprehensive review

Küresel yol güvenliğini geliştirmek için makine öğreniminden yararlanma: Kapsamlı bir inceleme

***Author**: Selma BULUT[1]*

ORCID[1]: 0000-0002-6559-7704

To cite to this article: Bulut S., "Harnessing Machine Learning to Enhance Global Road Safety: A Comprehensive Review", ***Journal of Polytechnic***, 27(6): 2127-2137, (2024).

Bu makaleye şu şekilde atıfta bulunabilirsiniz: Bulut S., "Harnessing Machine Learning to Enhance Global Road Safety: A Comprehensive Review", ***Politeknik Dergisi***, 27(6): 2127-2137, (2024).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: 10.2339/politeknik.1348075

Harnessing Machine Learning to Enhance Global Road Safety: A Comprehensive Review

Highlights

- *Road traffic accidents are the eighth leading cause of death worldwide.*
- *12,316 traffic accident records from Addis Ababa City Police Department.*
- *Random Forest (RF) outperformed other algorithms with an accuracy rate of 92.2%.*
- *The critical role of data preprocessing and the potential of machine learning in shaping effective road safety strategies.*
- *Significantly stabilized and improved dataset quality by leveraging SMOTE and Min-Max scaling for data preprocessing.*

Graphical Abstract

This study presents an analysis utilizing various machine learning algorithms, particularly the Random Forest algorithm achieving an accuracy of 92.2%, to predict traffic accidents in Addis Ababa, highlighting the significance of data preprocessing and mo del selection in achieving optimal results.

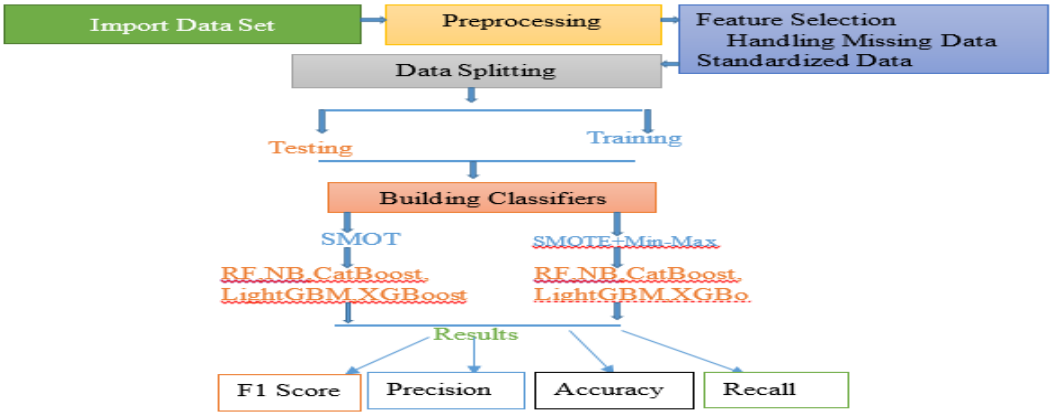

Figure. Research methodology steps

Aim

This study investigated machine learning's ability to predict traffic accidents in Addis Ababa, considering severity and causes.

Design & Methodology

In this research, 12,316 records obtained from the Addis Ababa City Police were examined using advanced pre-processing techniques like SMOTE and Min-Max scaling, and machine learning models such as Random Forest, Gaussian Naive Bayes, CatBoostClassifier, LightGBM, and XGBoost were meticulously evaluated for their effectiveness in deriving reliable insights from the data.

Originality

The research distinctively assesses a variety of machine learning models, namely Random Forest, Gaussian Naive Bayes, CatBoostClassifier, LightGBM, and XGBoost, applying them in an unprecedented manner to Addis Ababa's traffic data, with a particular focus on the significance of preprocessing.

Findings

Random Forest outperformed othe r models with a 92.2% accuracy rate, underscoring the importance of preprocessing and model-dataset compatibility.

Conclusion

Machine learning is promising for traffic analysis, but success hinges on precise preprocessing and model selection, especially in urban areas like Addis Ababa.)

Declaration of Ethical Standards

The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Harnessing Machine Learning to Enhance Global Road Safety: A Comprehensive Review

Araştırma Makalesi / Research Article

Selma BULUT[1*]

[1]Kırklareli Vocational High School, Deepartment of Computer Programming, Kırklareli University, Türkiye

(Geliş/Received : 22.08.2023 ; Kabul/Accepted : 14.10.2023 ; Early View : 07.03.2024)

ABSTRACT

As global urbanization accelerates, road safety remains a pressing concern, underscored by escalating traffic accidents and fatalities. Road Traffic Injuries (RTI) have become the eighth leading cause of death worldwide. The article delves deep into the potential of machine learning in predicting traffic accidents, their severity, and causal factors. This study comprehensively evaluates machine learning models on traffic accident records sourced from the Addis Ababa City Police Department. Comprising 12,316 records with 15 features, the dataset underwent preprocessing techniques, specifically Synthetic Minority Over-sampling Technique (SMOTE) and Min-Max scaling. Five algorithms – Random Forest (RF), Gaussian Naive Bayes, CatBoostClassifier, LightGBM, and XGBoost – were tested for their prediction accuracy. The findings spotlight the dominance of the RF model, achieving a peak accuracy of 92.2% post-SMOTE and Min-Max application. A comparative analysis with existing literature showed that while RF is a recurrently effective model across various datasets, data preprocessing and model suitability to specific datasets is paramount. This study underscores the potential of machine learning in traffic accident analysis and the nuanced choices researchers must make for optimal outcomes.

Keywords: Traffic accident analysis, machine learning, random forest, min-max scaling, comparative study.

Küresel Yol Güvenliğini Geliştirmek İçin Makine Öğreniminden Yararlanma: Kapsamlı Bir İnceleme

ÖZ

Küresel kentleşme hızlanırken, yol güvenliği, artan trafik kazaları ve ölümlerin altını çizdiği acil bir endişe olmaya devam etmektedir. Karayolu Trafik Yaralanmaları (RTI), dünya çapında sekizinci önde gelen ölüm nedeni haline geldi. Makale, trafik kazalarını, bunların ciddiyetini ve nedensel faktörleri tahmin etmede makine öğreniminin potansiyelini derinlemesine araştırmaktadır. Bu çalışma, Addis Ababa Şehri Polis Departmanından alınan trafik kazası kayıtları üzerindeki makine öğrenimi modellerini kapsamlı bir şekilde değerlendirmektedir. 15 özelliğe sahip 12.316 kayıttan oluşan veri setinde, Sentetik Azınlık Aşırı Örnekleme Tekniği (SMOTE) ve Min-Max ölçekleme başta olmak üzere ön işleme teknikleri uygulanmıştır. Beş algoritma – Random Forest (RF), Gaussian Naive Bayes, CatBoostClassifier, LightGBM ve XGBoost – tahmin doğruluğu açısından test edilmiştir. Bulgular, SMOTE ve Min-Max uygulamasından sonra %92,2'lik bir tepe doğruluğu elde eden RF modelinin hakimiyetine ışık tutmaktadır. Mevcut literatürle karşılaştırmalı bir analiz, RF'nin çeşitli veri kümelerinde yinelenen etkili bir model olmasına rağmen, veri ön işlemenin ve belirli veri kümelerine model uygunluğunun öneminin çok önemli olduğunu göstermiştir. Bu çalışma, trafik kazası analizinde makine öğreniminin potansiyelinin ve araştırmacıların optimum sonuçlar için yapması gereken incelikli seçimlerin altını çizmektedir.

Anahtar Kelimeler: Trafik kazası analizi, makine öğrenmesi, rastgele orman, min-maks ölçeklendirme, karşılaştırmalı çalışma.

1. INTRODUCTION

As people migrated from rural to urban areas, transportation becam e a problem. Traffic congestion is experienced, especially during morning and evening commutes. In the early days, this congestion originated from animals, people, and bicycles. With the advancement of technology and the invention of cars, vehicles also began contributing to this congestion.

The first known accident occurred on May 30, 1896, as a bicycle crash, and the first accident resulting in injuries occurred later the same year on August 17, involving a motorcycle. The first recorded fatal accident was on August 31, 1896, when Mary Ward tragically fell from an electric locomotive [1]. The first driver known to have died from injuries sustained in an automobile accident was on Saturday, February 12, 1898, when his electric wagon overturned [2].

Traffic accidents result in both material and emotional damage. When looking at countries, it can be seen that this rate is significantly high. For example, the Turkish

* *Corresponding Author*
e-posta : selma.bulut@klu.edu.tr

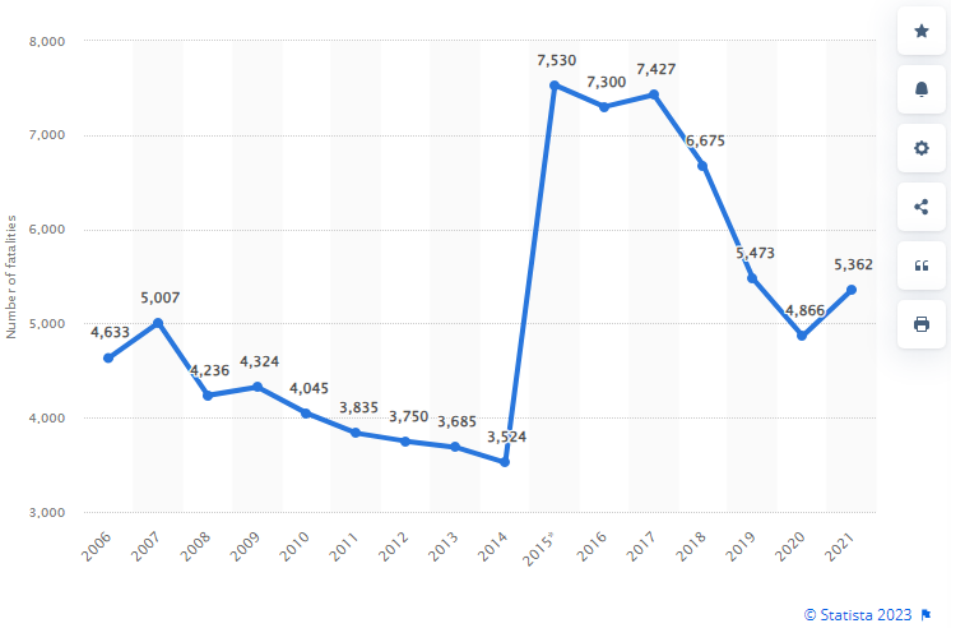

Figure 1. Traffic accident death rates in Turkey

Statistical Institute (TUIK) reports that the number of traffic accidents in Turkey reached 1.233 million in 2022. This figure is 3.9% higher than in 2021[3].As illustrated in Figure 1, there was a notable escalation in fatal road traffic incidents in Turkey during the year under review. Specifically, there was an increment of 496 such incidents, marking a 10% surge compared to the preceding year—the cumulative fatalities stemming from road traffic incidents in 2021 culminated at 5,362 deaths. "road fatalities" is defined as individuals who succumbed immediately at the accident site or within 30 days post-accident owing to the sustained injuries. It's imperative to note that these incidents invariably encompassed the involvement of at least one vehicular entity operating on a thoroughfare, be it publicly or privately accessible.

Road Traffic Injuries (RTI) are the eighth leading cause of death worldwide, resulting in 1.35 million deaths yearly. This equates to one person dying every 26 seconds on average. The Lancet's 2022 Road Safety report, addressing four main risk factors (driving under the influence of alcohol, helmet use, speed, and seatbelt use), stated that 25% to 40% of all deaths related to RTI could be prevented [4]. Today, traffic accidents are the primary cause of death for children and young adults aged 5-29 [5]. Road traffic safety analysis has been used to understand the causes of traffic accidents and to introduce safety measures, thereby saving lives [6,7].

In a call to governments by the World Health Organization, controlling speed on roads, abstaining from driving while intoxicated, using helmets on motorcycles, mandating the use of seat belts, and employing special seats for children have been pointed out as topics that directly influence the improvement of road safety. The International Road Assessment Programme (IRAP) believes that upgrading the world's roads to a 3-star standard or better would effectively contribute to achieving the United Nations Sustainable Development Goals' target of halving road deaths and injuries by 2030 [8]. This strategy will impact drivers' safety and well-being and other road users, such as pedestrians and cyclists [9].

A mere 28 nations, accounting for 449 million individuals, equivalent to seven percent of the global populace, have instituted legal frameworks encompassing the quintessential traffic safety determinants: velocity regulations, anti-intoxication driving measures, helmet mandates, seat belt enforcement, and child restraint systems. It's noteworthy that in nations with low to medium incomes, pedestrians and cyclists constitute over a third of traffic mortality victims [10]. Nevertheless, under 35% of such countries have strategized regulations to shield these susceptible road participants. According to the World Health Organization's death rate statistics in Figure 2, the global death rate from traffic accidents is 17.4 per 100,000 individuals. There is a noticeable disparity between countries based on income: while the rate stands at its highest in low-income countries at 24.1, it's at its lowest in high-income countries at 9.2 [11].

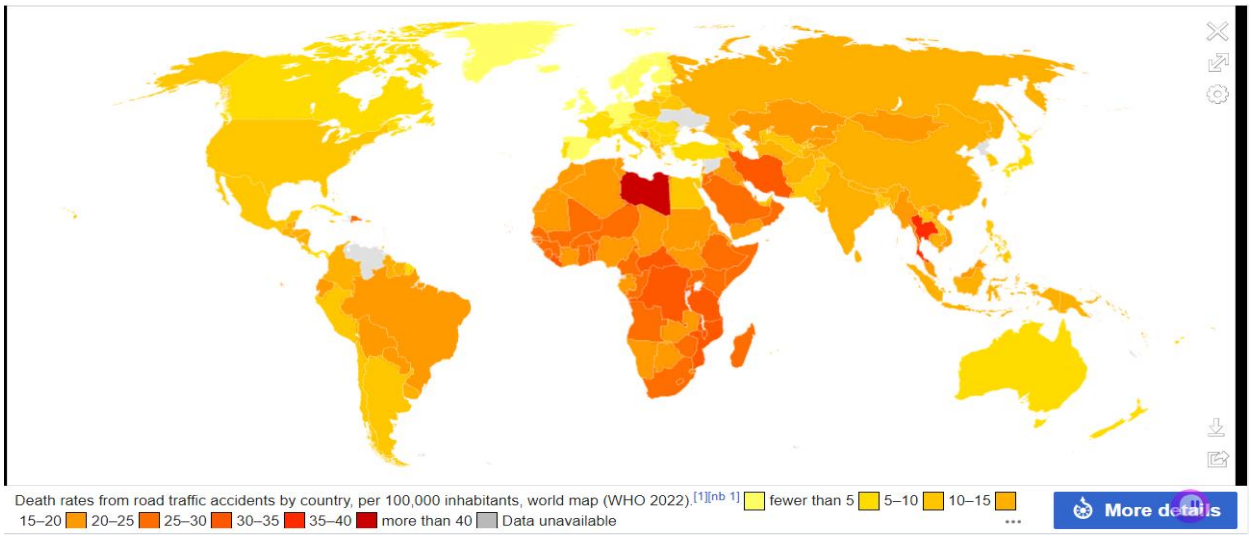

Figure 2. Traffic accident death rate per 100,000 people by country

Situated as the capital of Ethiopia, Addis Ababa continually observes an augmentation in its populace and vehicular density. Ethiopia's lamentably elevated mortality rate attributable to vehicular incidents is noteworthy, placing it among the global frontrunners in this grim statistic [12]. Data from the World Health Organization elucidates that in 2013 for every 100,000 vehicles, there were as many as 4,984 fatalities due to traffic collisions, positioning Ethiopia 24th globally. Alarmingly, the casualty rate from these accidents in Ethiopia is roughly thirty-fold compared to the incidents in the United States [13]. Further statistics from the World Health Organization in 2015 indicated a mortality rate 94 for every 100,000 inhabitants due to these accidents. This daunting figure, a staggering 79% death rate resulting from traffic incidents, undeniably categorizes Ethiopia among nations with the most perilous road conditions [11; 14].

This study conducts a critical evaluation of the application and efficacy of a range of machine learning techniques—including Random Forest, Gaussian Naive Bayes, CatBoostClassifier, LightGBM, and XGBoost—in predicting the outcomes, severities, and underlying causes of road traffic accidents. Drawing on a synthesis of research conducted between 2018 and 2023 [12,20,22,26,41,43], this investigation aims to integrate findings on the performance of machine learning models across diverse geographical regions, under varied conditions, and with different datasets, emphasizing the pivotal role of these models in enhancing the predictability and understanding of traffic incidents. By understanding the roles of factors such as driver age, vehicle type, road type, and traffic flow in predicting accidents, this research aspires to uncover patterns, strengths, and areas of improvement in using machine learning to enhance traffic safety. Additionally, this investigation aims to understand the feasibility of machine learning models in accurately identifying accident-prone hotspots, guiding road safety policies, and formulating preventative measures.

2. RELATED WORK

The global initiative to augment transportation safety has realized significant advancements with the integration of innovative technological methodologies, positioning machine learning as a cardinal instrument. Numerous investigations have entered this cross-disciplinary area, contributing vital insights.

Beshah and Hill (2010) explored the relationship between road-related factors and the severity of traffic accidents, utilizing classifiers such as Naive Bayes, Decision Tree (J48), and K-Nearest Neighbors [15]. Krishnaveni and Hemalatha (2011) used Naive Bayes, AdaBoostM1, PART, J48 Decision Tree and Random Forest Tree classifier to predict the severity of injuries occurring in traffic accidents [40]. Their pioneering work illuminated the complexities inherent in these relationships, serving as a foundational layer for subsequent nuanced investigations and shaping the methodological undertones of this research.

Chen et al. (2016) refined the focus by analyzing rollover accidents, shedding light on the dynamics and often catastrophic consequences of such incidents using SVM models [16]. Their insights are instrumental in fine-tuning the specific analytical perspectives adopted in the current study.

By acknowledging the vital interplay between traffic flow and accident prevention, Li et al. (2018) and Zeng and Huang (2020) paved the way for predictive modeling, underscoring the correlation between **Data Set**

The data for this study was sourced from the Addis Ababa City police departments. We utilized manual records of traffic accidents from 2017 to 2020. Any sensitive information in the records was meticulously excluded during the coding process to ensure confidentiality and privacy. After this cleansing process,

the dataset comprised 32 distinct features with 12,316 accident record samples [29]. A brief description of each of the fields in the dataset is given in table1.

Table 1. List of dataset estimators and ranges of values

Features	***mean***	***std***	***min***	***max***	***Definition***
Age_band_of_driver	*2,27*	*1,16*	*0*	*4*	*Indicates the age range to which the driver belongs.*
Sex_of_driver	*1,91*	*0,33*	*0*	*2*	*Represents the gender of the driver.*
Educational_level	*3,66*	*1,20*	*0*	*6*	*Represents the educational attainment level of the driver.*
Vehicle_driver_relation	*1,13*	*0,48*	*0*	*3*	*Describes the relationship between the vehicle and the driver.*
Driving_experience	*3,97*	*1,67*	*0*	*6*	*Indicates the level of driving experience.*
Lanes_or_Medians	*3,23*	*1,57*	*0*	*6*	*Represents the number or type of lanes or medians at the accident site.*
Types_of_Junction	*2 ,45*	*2,08*	*0*	*7*	*Indicates the type of junction where the accident occurred.*
Road_surface_type	*3,82*	*0,74*	*0*	*5*	*Describes the type of road surface at the accident site.*
Light_conditions	*1,32*	*0,56*	*1*	*4*	*Represents the lighting conditions at the time of the accident.*
Weather_conditions	*1,67*	*1,69*	*0*	*8*	*Describes the weather conditions at the time of the accident.*
Type_of_collision	*2,95*	*1,20*	*0*	*9*	*Indicates the type of collision that occurred.*
Vehicle_movement	*4,73*	*3,14*	*0*	*12*	*Represents the movement of the vehicle before the collision.*
Pedestrian_movement	*1,09*	*0,60*	*0*	*8*	*Describes the movement of any pedestrian involved in the accident.*
Cause_of_accident	*7,75*	*5,38*	*0*	*19*	*Indicates the main cause of the accident.*
Accident_severity	*1,83*	*0,41*	*0*	*2*	*Represents the severity of the accident.*

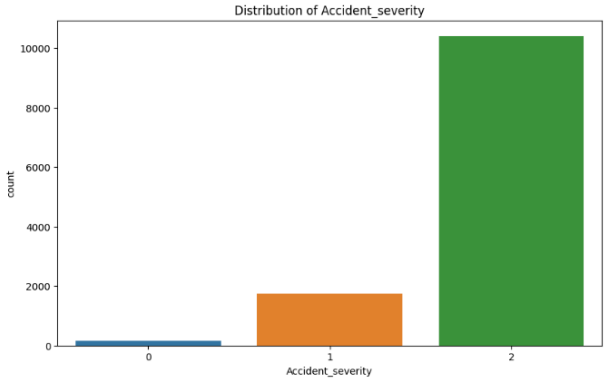

Figure 3. Traffic accident dataset classification attribute data distribution

The Traffic Accident dataset Classification attribute data distribution is shown in Figure 3.

Limitations

The dataset's imbalanced distribution of the "Accident Severity" attribute presents several challenges. Primarily, machine learning models may be predisposed to favor the majority class, which can lead to potential misclassification or overlooking of the minority classes. Relying solely on standard accuracy for model evaluation may be misleading; alternative metrics such as precision, recall, and F1-score become imperative. This imbalance also raises concerns about the model's ability to generalize to real-world scenarios, especially those that deviate from the dataset's distribution. Another potential pitfall is overfitting to the majority class, making models less robust in diverse situations. Furthermore, the dataset's representation of the broader population could be better, especially for the underrepresented severity levels. Specialized techniques might be needed to address these imbalances, like oversampling or undersampling, but they come with complexities and considerations. The sequential process followed in the research study is shown in Figure 4.

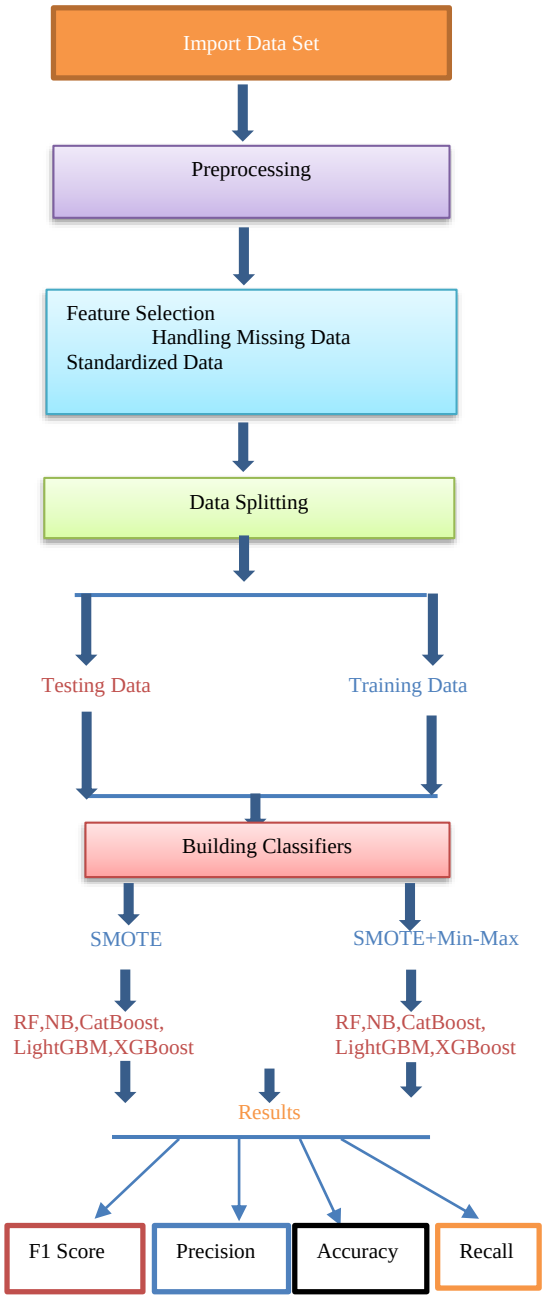

Figure 4. Research methodology steps

3. DATA ANALYSIS

The analyses were performed on Google Colab, a reputable cloud-based platform allowing for interactive Python scripting, chosen for its accessibility and robustness. In dealing with a variety of categorical attributes on differing scales, Min-Max normalization was employed to standardize the attributes to a uniform scale, promoting enhanced model convergence and improved performance [30]. This technique is essential for maintaining attribute proportionality and ensuring that no particular feature dominates the model due to its scale.

To rectify the imbalanced distribution observed in the "Accident Severity" attribute, the Synthetic Minority Over-sampling Technique (SMOTE) was utilized. SMOTE is instrumental in synthesizing new samples for the minority class, hence balancing the class distribution and mitigating model bias towards the majority class [31]. This is crucial for improving the reliability and generalizability of the model, ensuring equitable representation of all classes in the model training process.

Several machine learning models, namely, Random Forest, Gaussian Naive Bayes, CatBoostClassifier, LightGBM, and XGBoost, were employed on the adjusted dataset [34-39]. These models were selected due to their proven efficacy in handling categorical data and their adaptability to varied dataset characteristics, as evidenced in prior research. Each model's performance was rigorously evaluated based on pertinent metrics including accuracy, precision, recall, and F1-score [42], ensuring a comprehensive assessment of the model's predictive capabilities.

3.3.1 Min-max

Known as feature scaling, min-max normalization is a data preprocessing technique used to convert numerical data into a standard scale. It involves scaling the values of a variable between a specific minimum and maximum range, typically between 0 and 1.

Assuming a variable x has minimum and maximum values of min_x and max_x, respectively, the normalization formula to scale a value (y) between 0 and 1 would be as follows:

$$y' = \frac{y - min_x}{max_x - min_x}(new\_max_x - new\_max_x) + new\_min_x \quad (1)$$

In this context, y' signifies the standardized magnitude, where min_x and max_x correspond to the lowermost and uppermost values for the variable x, respectively. Conversely, new_min_x and new_max_x delineate the lower and upper boundaries of the desired normalization range [30].

3.3.2 Synthetic minority over-sampling technique (SMOTE)

In machine learning classification, imbalanced data refers to datasets where the number of instances belonging to different classes is uneven, leading to a potential bias in the classifier's performance. SMOTE, introduced by Chawla et al. in 2002, is an oversampling technique that aims to overcome the imbalance problem by generating synthetic examples for the minority class [31]. The method operates through the stochastic selection of an instance from the underrepresented class, subsequently discerning its k-adjacent entities using the k-NN algorithm. Leveraging these proximate entities, SMOTE fabricates novel synthetic instances along the vectorial pathway interlinking the minority class instance and its neighboring counterparts.

Generating synthetic examples involves selecting a minority class instance and calculating the feature-wise differences between it and its neighbors. SMOTE then

randomly chooses a number between 0 and 1, multiplying the feature-wise differences by this value. The resulting values are added to the selected minority instance, producing new synthetic examples that represent the minority class but differ slightly in their feature values. Applying SMOTE makes class distribution more balanced, as artificial models are introduced to augment the minority class. This helps to alleviate the bias caused by imbalanced data and allows the classifier to learn from a more representative dataset [32].

3.3.3 Random forest

An approach that combines several randomized decision trees and averages their predictions by summing them up has shown excellent performance in environments where the number of variables is much larger than the number of observations.

RF is a method that forms a forest consisting of numerous decision trees during training time and outputs the class, the mode of the types obtained from individual trees [33]. In the randofm forest classification paradigm, individual decision trees are cultivated utilizing distinct subsets of the training dataset. Concurrently, a stochastic selection of attributes is examined for potential bifurcation at every nodal juncture within the tree. Such inherent stochasticity is a mitigator against model overfitting, amplifying its aptitude for generalization across unseen data [34].

3.3.4 Gaussian naive bayes

The Gauss Naive Bayes classifier is a variant of the Naive Bayes algorithm that assumes that the features follow a Gaussian distribution. It is commonly used for classification tasks where the parts are continuous variables. The training data is divided by class, and each class's mean and standard deviation are calculated. Therefore, the following equation can be used to estimate the probabilities of the continuous data set [35].

$$P(X = x|C = c) = \frac{1}{\sqrt{2\pi\sigma}} e^{\frac{-(x-\mu)^2}{2\sigma^2}} \quad (2)$$

where x = variable, c = class, μ = mean, σ = standard deviation.

3.3.5 CatBoostClassifier

It is a gradient-boosting algorithm specifically designed to work effectively with categorical features. Yandex, a Russian search engine company, developed it. CatBoostClassifier uses a collection of decision trees to make predictions. It can process flat parts directly without needing one-hot or label encoding. It utilizes an ordered boosting technique, which increases the algorithm's performance by considering the order of categories [36].

3.3.6 LightGBM

LightGBM operates as a framework within the domain of gradient boosting, emphasizing tree-structured learning methodologies. Its primary objective is to promote scalability and computational efficiency. To achieve this, LightGBM incorporates the Gradient-based One-Side Sampling (GOSS) approach, selectively focusing on the most salient instances for gradient calculations, resulting in a notable reduction in memory consumption and training duration. Moreover, it endorses parallel processing and GPU-accelerated learning, facilitating accelerated training on multi-threaded CPUs and GPUs. Due to its swift computational pace, exemplary accuracy, and proficiency in managing voluminous datasets, LightGBM has garnered significant traction in machine learning [37].

3.3.7 XGBoost

XGBoost, for eXtreme Gradient Boosting, represents a widely-utilized algorithm within machine learning tailored for regression and classification challenges. This algorithm employs the gradient boosting mechanism, which amalgamates several weak predictive models, predominantly decision trees, to formulate a potent composite model. Distinctively in XGBoost, tree construction occurs concurrently across multiple processing cores, and data structuring is optimized for swift retrieval, thereby streamlining the model training process and bolstering its efficiency [38, 39].

4. MODEL PERFORMANCE AND EVALUATION

The TP, TN, FP, and FN Confusion matrix metrics provide values for correct or incorrect classification of packets in the firewall. These values were used to calculate precision, recall, f-measure, and accuracy metrics as follows:

$$\text{Precision} = \frac{TP}{(TP+FP)} \quad (3)$$

$$\text{Recall} = \frac{TP}{(TP+FN)} \quad (4)$$

$$\text{F-measure} = \frac{(2*precision*recall)}{(precision+recall)} \quad (5)$$

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (6)$$

Table 2. Confusion matrix

	Predict Class	
Actual Class	***Yes***	***No***
Yes	*TP*	*FN*
No	*FP*	*TN*

5. RESULTS

Our investigation into the diverse machine learning models on the dataset, which underwent SMOTE preprocessing, yielded the following key observations. The key findings in Table 4 emerged when we examined various machine learning models in a preprocessed SMOTE dataset.

Table 3. Analysis results of confusion matrix applied to Smote

		RF			**Gaussian NB**			**CatBoostClassifier**			**LightGBM**			**XGBoost**		
	Class	**Predicted**														
Actual	0	2065	13	7	1548	246	291	2043	27	15	1642	256	187	2035	26	24
	1	92	1731	277	1090	591	419	107	1615	378	1043	552	505	160	1550	390
	2	201	295	1568	851	501	712	116	478	1470	788	402	874	136	483	1445

Table 4. Analysis results of the dataset applied to Smote

	RF	Gaussian NB	CatBoostClassifier	LightGBM	XGBoost
Accuracy	**0,856**	0,456	0,821	0,770	0,805
F1 Score	**0,876**	0,477	0,823	0,775	0,809

The Random Forest (RF) model distinctly stood out, delivering an accuracy of 85.6% and an F1 Score of 87.6%, showcasing its ability to navigate the dataset's intricacies after SMOTE was applied adeptly. Conversely, the Gaussian Naive Bayes model faced significant challenges, with its suboptimal accuracy of 45.6% and an F1 Score of 47.7%; this might hint at the model's inherent assumptions being at odds with the dataset's nature. The gradient boosting models, namely CatBoostClassifier and XGBoost, maintained commendable consistency, each achieving accuracies around the 80% mark, emphasizing their capability to handle complex datasets. LightGBM, though proficient with a 77% accuracy, found itself slightly eclipsed by the other mentioned models. These variances in model outcomes underscore the importance of discerning model selection based on dataset nuances. Moreover, our study accentuates the value of the F1 Score as a holistic evaluation metric, especially when confronted with datasets with potential class imbalances.

Table 5. Analysis results of confusion matrix applied to Smote+Min-Max

		RF			**Gaussian NB**			**CatBoostClassifier**			**LightGBM**			**XGBoost**		
	Class	**Predicted**														
Actual	0	2068	3	14	1639	151	295	2046	12	27	2033	24	28	2048	11	26
	1	50	1735	315	1263	406	431	60	1651	389	95	1611	394	52	1668	380
	2	31	51	1982	1069	374	621	9	32	2023	5	7	2052	5	27	2032

Table 6. Analysis results of the dataset applied to Smote+Min-Max

	RF	**Gaussian NB**	**CatBoostClassifier**	**LightGBM**	**XGBoost**
Accuracy	**0,922**	0,427	0,915	0,912	0,920
F1 Score	**0,929**	0,462	0,916	0,913	0,921

Table 6 illustrates the performance of various machine learning models on a dataset treated with two preprocessing techniques: SMOTE, which addresses class imbalance, and Min-Max scaling, which standardizes feature values. The preprocessing bolstered many model performances. The Random Forest (RF) model exhibited a pronounced improvement, with an accuracy of 92.2% and an F1 Score of 92.9%, showcasing its adeptness at managing the dataset's intricacies post-processing. In contrast, Gaussian Naive Bayes lagged considerably, achieving a mere 42.7% accuracy, indicating that the model's foundational assumptions may need to be more consistent with the dataset despite the dual preprocessing. Close on RF's heels, both CatBoost and XGBoost delivered sterling performances, registering accuracies of 91.5% and 92%, respectively, underlining their robustness. LightGBM, too, asserted itself as a strong contender with an accuracy of 91.2%. The data suggests that while SMOTE and Min-Max scaling can significantly augment performance, model selection remains paramount. The dichotomy between Gaussian Naive Bayes and the ensemble methods reaffirms this, emphasizing that the right preprocessing and model synergy are crucial for optimal outcomes.

Table 7. Previous similar studies and their results

Authors	Dataset	Number of Features	Applied Models	Results
Raja et all [14]	Oromia Police Commission data (6170 records)	15 accidents attributes	BPNN, MLPNN, FFNN, RNN, RBFNN, LSTM	RNN accuracy of 97.18%
Bedane et all [12]	Addis Ababa city police departments (12316 records)	32 features	LR, NB, Decision Tree, SVM, k-NN, RF, and AdaBoost	RF achieved a 93.76% F1 score with SMOTE + PCA
Kumeda et all [20]	UK data.gov.uk	12 features	Fuzzy-FARCHD, RF, Hierarchal LVQ, RBF Network Multilayer Perceptron and NB	Fuzzy-FARCHD accuracy of 85.94%.
Çelik and Sevli [22]	Austin, Dallas, and San Antonio city of Texas (1.1 million records)	Six features	LR, XGBoost, RF, KNN, and SVM	LR accuracy of 88.1%
Krishnaveni and Hemalatha [40]	Transport Department of Government of Hong Kong (34,575 records)	Nine features	NB, AdaBoostM1 Meta classifier, PART Rule classifier, J48 Decision Tree classifier, RF	RF accuracy of 89.81%
AlMamlook et all [41]	Western Michigan University (WMU), Transportation Research Center for Livable Communities (TRCLC)		AdaBoost, LR, NB, and RF with SMOTE	RF algorithm accuracy of 75.5%
Beshas & Hill [15]		Ten features	Decision Tree(J48), NB, KNN	K-NN accuracy of 80.8281%
Ahmed et all [26]	New Zealand dataset (184314 records)	16 features	RF, DJ, Adaboost, XGBoost, L-GBM, CatBoost	RF accuracy of 81.45%
Rezashoar et all [43]	252 thousand records	32 features	NB, SVM, NN	NB accuracy of 75.10%
Our Study	Addis Ababa city police departments (12316 records	15 features	RF, Gaussian NB, CatBoostClassifier, LightGBM, XGBoost	RF accuracy of 92.2% Smote+Min-Max

6. DISCUSSION AND CONCLUSION

The application of machine learning in analyzing traffic accident data is a growing field of study, garnering diverse applications across varied geographical locales and datasets, as evident from the summarized works in Table 7. Through an in-depth comparative examination of prior works and our study, several pivotal insights and contributions to the existing literature have been deduced.

This study, juxtaposed with preceding works, accentuates the paramountcy of Random Forest (RF) in analyzing traffic accidents, contributing empirical evidence to its recurrent efficacy across different contexts and datasets. Our findings augment the understanding of RF's adaptability and robustness, offering nuanced insights on its optimal utilization in diverse settings and substantiating its prevalence in contemporary research.

Our research underscores the significance of meticulous data preprocessing and model selection in enhancing the predictive accuracy of traffic accidents, which is imperative for the development of proactive, data-driven interventions and policies aimed at mitigating traffic-related fatalities and injuries. The correlations found between different features and accident severity in our study provide a framework for targeted traffic safety measures, potentially aiding in the reduction of accidents in regions with similar traffic and road conditions.

The consistent performance of RF across different studies, including ours, signals its potential as a foundational tool for future research endeavors in traffic accident prediction. However, the variability in optimal model choices across different datasets, as exemplified by the success of RNN in Raja et al.'s study [14] and Logistic Regression in Çelik & Selvi's research [22], reinforces the necessity for context-specific model selection and customization.

Furthermore, our study highlights the critical role of balancing techniques like SMOTE in mitigating model biases and enhancing performance, emphasizing the need for balanced and representative datasets in traffic accident studies. The observed discrepancies in model performances across different studies underscore the

intricate nature of machine learning applications and the necessity for nuanced, dataset-specific approaches, negating a one-size-fits-all solution.

While this research consolidates the reliability and versatility of the RF algorithm in traffic accident studies, it also illustrates the critical interplay between data preprocessing, model selection, and contextual nuances in obtaining optimal results. The insights garnered from our study provide a stepping stone for future research, suggesting exploration into hybrid or ensemble models that amalgamate the strengths of multiple algorithms to refine predictive accuracy further. By doing so, subsequent research can contribute to the formulation of more effective, data-informed strategies for enhancing traffic safety and reducing accident-related adversities.

proficient traffic management and a reduction in congestion-related incidents [17, 18]. Their work fortifies the importance of addressing traffic flow within the present research framework.

Dong et al. (2018) and Kumeda et al. (2019) significantly contributed to the realm of predictive modeling by incorporating intricate methodologies and high-performance algorithms, setting benchmarks for methodological rigor and innovative approaches in this study [19, 20].

Al Mamlook et al. (2019) based on their 271,563 traffic accident data; AdaBoost has implemented supervised machine learning algorithms such as LR, NB, and RF. SMOTE was used to eliminate the imbalance in the data. The findings of this study showed that the RF model could be a promising tool for predicting injury severity in traffic accidents[41].

Root cause analysis and prediction took a forefront in the studies by Gan et al. (2020) and Bedane et al. (2021), addressing essential aspects such as class imbalances and refining the methodological approaches to be adopted in this research [21, 12].

Çelik and Selvi (2022) and Raja et al. (2023) broadened the analytical horizon by undertaking exhaustive comparisons of diverse machine learning techniques and innovating with tailored neural network architectures [22, 14]. Their comprehensive insights are crucial in determining the analytical breadth and depth of approaches employed in this study.

Ahmed et al. (2023) initiated a significant discourse by introducing the concept of "explainable" machine learning models, with a primary focus on enhancing the interpretability of prediction outcomes [26]. This trajectory of identifying accident-prone areas and uncovering root causes received further support from Santos et al. (2021) and Yassin and Pooja (2020), both of whom underscored the pivotal role of data-driven strategies in bolstering public safety initiatives [27, 28]. In a related context, Rezashoar et al. (2023) highlighted that the machine learning algorithms they proposed can serve as practical decision-making tools for a wide range of government departments and traffic and transportation organizations, particularly in the context of road safety measures [43]. This focus on actionable insights from complex modeling is mirrored in the goals of the current study.

Finally, each referenced study serves as a stepping stone, providing indispensable frameworks and analytical paradigms, emphasizing the importance of a diverse and nuanced understanding of traffic accidents. Together, they reinforce the methodological and contextual fabric of this research, aimed at synthesizing and advancing these multifaceted insights to explore pivotal questions surrounding road traffic safety.

DECLARATION OF ETHICAL STANDARDS

The author of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

AUTHORS' CONTRIBUTIONS

Selma BULUT: Produced specimens, conducted experiments, analysed results, wrote the manuscript.

CONFLICT OF INTEREST

There is no conflict of interest in this study.

REFERENCES

[1] Fallon I. and O'Neill D., "The world's first automobile fatality," ***Accid. Anal. Prev***., vol. 37, no. 4, pp. 601–603, (2005).

[2] "When did the first motoring fatality occur?," National Motor Museum, 11-Jan-2018. [Online]. Available: https://nationalmotormuseum.org.uk/ufaqs/when-did-the-first-motoring-fatality-occur/. [Accessed: 21-Aug-2023].

[3] "2022'de Türkiye'de artan trafik kazası sayısı.", Atlas Magazine, 31-April-2023. [Online]. Available: https://www.atlas-mag.net/en/category/pays/turquie/rising-number-of-road-accidents-in-turkey-in-2022#:~:text=The%20Turkish%20Statistical%20Institute%20(TurkStat,the%20remainder%20in%20material%20damage. [Accessed: 21-Aug-2023].

[4] Thelancet.com. [Online]. Available: https://www.thelancet.com/infographics-do/road-safety-2022. [Accessed: 21-Aug-2023].

[5] "Global status report on road safety 2018," Whoint, 17-Jun-2018. [Online]. Available: https://www.who.int/publications/i/item/9789241565684. [Accessed: 21-Aug-2023].

[6] Li L., Zhu L., and Sui D. Z., "A GIS-based Bayesian approach for analyzing spatial–temporal patterns of intra-city motor vehicle crashes," ***J. Transp. Geogr***., vol. 15, no. 4, pp. 274–285, (2007).

[7] Tola A. M., Demissie T. A., Saathoff F., and Gebissa A., "Severity, spatial pattern and statistical analysis of road traffic crash hot spots in Ethiopia," ***Appl. Sci***. (Basel), vol. 11, no. 19, p. 8828, (2021).

[8] "3 star or better," iRAP, 02-Aug-2017. [Online]. Available: https://irap.org/3-star-or-better/. [Accessed: 21-Aug-2023].

[9] Gutierrez-Osorio C., González F. A., and Pedraza C. A., "Deep Learning ensemble model for the prediction of traffic accidents using social media data," ***Computers,*** 11(9): 126, (2022).

[10] "List of countries by traffic-related death rate" Wikipedia. 12-Aug-2023. [Online]. Available: https://en.wikipedia.org/wiki/List_of_countries_by_traffic-related_death_rate . [Accessed: 21-Aug-2023].

[11] Archive.org. [Online]. Available: https://web.archive.org/web/20151020144338/http://www.who.int/violence_injury_prevention/road_safety_status/2015/en/. [Accessed: 21-Aug-2023].

[12] Bedane T. T., Assefa B. G., and Mohapatra S. K., "Preventing traffic accidents through machine learning predictive models," ***in 2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA***), (2021).

[13] Mackay M., "Global priorities for vehicle safety," ***Traffic Inj. Prev***., 4(1): 1–4, (2003).

[14] Raja K., Kaliyaperumal K., Velmurugan L., and Thanappan S., "Forecasting road traffic accident using deep artificial neural network approach in case of Oromia Special Zone," ***Soft Comput***., (2023).

[15] Beshah T.,, and Hill S., "Mining road traffic accident data to improve safety: role of road-related factors on accident severity in Ethiopia". ***In the 2010 AAAI Spring Symposium series***, (2010).

[16] Chen C., Zhang G., Qian Z., Tarefder R. A., and Tian Z., "Investigating driver injury severity patterns in rollover crashes using support vector machine models," ***Accid. Anal. Prev***., 90: 128–139, (2016).

[17] Liu M., Wu J., Wang Y., and He L., "Traffic flow prediction based on deep learning." ***Journal of System Simulation***, 30(11): 4100, (2018).

[18] Zheng J. and Huang M., "Traffic flow forecast through time series analysis based on deep learning," ***IEEE Access***, 8: 82562–82570, (2020).

[19] Dong C., Shao C., Li J., and Xiong Z., "An improved deep learning model for traffic crash prediction," ***J. Adv. Transp.***, 2018: 1–13, (2018).

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

[21] Gan J., Li L., Zhang D., Yi Z., and Xiang Q., "An alternative method for traffic accident severity prediction: Using Deep Forests algorithm," ***J. Adv. Transp.,*** 2020: 1–13, (2020).

[22] Çelik A. and Sevli O., "Predicting traffic accident severity using machine learning techniques," ***Türk Doğa ve Fen Dergisi,*** 11(3): 79–83, (2022).

[23] Ghandour A. J., Hammoud H., and Al-Hajj S., "Analyzing factors associated with fatal road crashes: A machine learning approach," ***Int. J. Environ. Res. Public Health,*** 17(11): 4111, (2020).

[24] Bhuiyan, H., Ara, J., Hasib, K. M., Sourav, M. I. H., Karim, F. B., Sik-Lanyi, C., ... and Yasmin, S., "Crash severity analysis and risk factors identification based on an alternate data source: a case study of developing country," ***Sci. Rep.,*** 12(1): 21243, (2022).

[25] Al-Mistarehi B. W., Alomari A. H., Imam R., and. Mashaqba M, "Using machine learning models to forecast the severity level of traffic crashes by R Studio and ArcGIS". ***Frontiers in the built environment,*** 8, 860805, (2022).

[26] Ahmed S., Hossain M. A., Ray S. K., Bhuiyan M. M. I., and Sabuj S. R., "A study on road accident prediction and contributing factors using explainable machine learning models: analysis and performance," ***Transp. Res. Interdiscip. Perspect***., 19(100814): 100814, (2023).

[27] Santos D., Saias J., Quaresma P., and Nogueira V. B., "Machine learning approaches to traffic accident analysis and hotspot prediction," Computers, 10(12): 157, (2021).

[28] Yassin S. S. and Pooja, "Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach," ***SN Appl. Sci***., 2(9): (2020).

[29] Bedane T. T., "Road Traffic Accident Dataset of Addis Ababa City." ***Mendeley***, (2020).

[30] Özsürünç R., "The role of data mining in digital transformation," ***in Contributions to Management Science***, Cham: Springer International Publishing, 177–190, (2023).

[31] Chawla N. V., Bowyer K. W., Hall L. O., and Kegelmeyer W. P., "SMOTE: Synthetic minority over-sampling technique," ***J. Artif. Intell. Res.***, 16: 321–357, (2002).

[32] Amirruddin A. D., Muharam F. M., Ismail M. H., Tan N. P., and Ismail M. F., "Synthetic Minority Over-sampling TEchnique (SMOTE) and Logistic Model Tree (LMT)-Adaptive Boosting algorithms for classifying imbalanced datasets of nutrient and chlorophyll sufficiency levels of oil palm (Elaeis guineensis) using spectroradiometers and unmanned aerial vehicles," ***Comput. Electron. Agric.***, 193(106646): 106646, (2022).

[33] Fai N. J., Wey W. K., Qi K. Y., Xian G. J., Chun R. J. M., and bin Abdul Salam Z. A., "Digits Classification Using Random Forest Classifier". ***Journal of Applied Technology and Innovation*** (e-ISSN: 2600-7304), 7(3): 63, (2023).

[34] Breiman L., "Random forests". ***Machine learning***, 45(1): 5–32, (2001).

[35] Gayathri, B. M., & Sumathi, C. P. ,"An automated technique using Gaussian naïve Bayes classifier to classify breast cancer," ***Int. J. Comput. Appl.***, 148(6): 16–21, (2016).

[36] Deekshitha B., Aswitha C., Sundar C. S., and. Deepthi A. K, "URL-Based Phishing Website Detection by Using Gradient and Catboost Algorithms." ***Int. J. Res. Appl. Sci. Eng. Technol***., 10(6): 3717–3722, (2022).

[37] "Welcome to LightGBM's documentation! — LightGBM 4.0.0 documentation," Readthedocs.io. [Online]. Available: https://lightgbm.readthedocs.io/en/stable/. [Accessed: 21-Aug-2023].

[38] Ramraj S.,, , UzirSunil N., R., and Banerjee S., "Experimenting XGBoost algorithm for prediction and classification of different datasets". ***International***

Journal of Control Theory and Applications, 9(40): 651-662, (2016).

[39] Memon N., Patel S. B., and Patel D. P., "Comparative analysis of artificial neural network and XGBoost algorithm for PolSAR image classification," in Lecture Notes in Computer Science, Cham: ***Springer International Publishing,*** 452–460, (2019).

[40] Krishnaveni S. and Hemalatha M., "A perspective analysis of traffic accident using data mining techniques," ***Int. J. Comput. Appl.***, 23(7): 40–48, (2011).

[41] AlMamlook R. E., Kwayu K. M., Alkasisbeh M. R., and Frefer A. A., "Comparison of machine learning algorithms for predicting traffic accident severity," in ***2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT)***, (2019).

[42] Korkmaz, A. and Buyukgoze, S. "Detection of Fake Websites by Classification Algorithms." ***Eur J. Sci. Technol***., 16: 826–833, (2019).

[43] Rezashoar, S., Kashi, E., and Saeidi, S. "Comparison of Machine Learning Algorithms for Predicting Traffic Accident Severity (Case Study: United Kingdom from 2010 to 2014)". doi.org/10.21203/rs.3.rs-3101818/v1. https://www.researchsquare.com/article/rs-3101818/v1. (2023).