Machine Learning Models established toward the Car Smash Injury Difficulty

Lixuan Zhang, Chang Li, Lee Chen, Don Chen, Zheng Xiang, Bing Pan
Faculty of Computer Science and Information System, Universiti Teknologi MARA (UiTM), Malaysia

ABSTRACT

Car crash can cause serious and severe injuries that impact people every day. Those injuries could be especially damaging for elderly drivers of age 60 or more. The goal of this research is to investigate the risk factors that contribute to crash injury severity among elderly drivers. This is accomplished by designing accurate machine learning based predictive models. Na ̈ive Bayesian (NB), Decision Tree (DT), Logistic Regression (LR), Light-GBM, and Random Forest (RF) model are proposed. A set of influential factors are selected to build the five predictive models to classify the severity of injuries as severe injury or non- severe injury. Michigan traffic data of the elderly population is used in this paper. Data normalization and Synthetic Minority Oversampling Technique (SMOTE) as injury classes balancing technique are used in the pre-processing phase. Results show that the Light-GBM achieved the highest accuracy among the five tested models with 87%. According to the Light-GBM model, the three most important factors that impact the severity of injuries are the driver's age, traffic volume, and car's age.

KEYWORDS: machine learning, traffic accident prediction, feature selection

1.0 INTRODUCTION

The World Health Organization (WHO) road safety technical package of 2017 [1-7] indicates that traffic crash lead to a loss of over 1.2 million lives yearly and injure millions more worldwide. For this reason, road traffic crash are currently estimated to be the ninth leading cause of death among all age groups globally. The United Nations (UN) highlights of world population aging in 2017 state that the global population of age 60 years or over numbered 962 million, which is more than double the same figure in 1980 when there were 382 million. The elderly population is expected to double again by 2050 to reach nearly 2.1 billion [8-13]. This demographic shift would challenge a large number of countries to ensure that their transportation systems are able to handle the change. One of the expected consequences of the demographic shift is that the elder's usage of mobility is projected to increase. According to the National Highway Traffic Safety Administration facts of 2015, the number of elder drivers over the age of 65 who are licensed to drive vehicles has increased significantly [13-16]. Although elder drivers are less prone to vehicle crash in terms of numbers than younger drivers, injuries are more. A study of car crashes fatal statistics in European Union (EU) roads shows that the fatality rate in elderly drivers is increasing [17-21]. The car crashes caused by the speed factor cause higher risk on seniors than same speed car crashes of younger driver [1-5]. In the near future, transportation systems worldwide would increasingly face the challenge of how to reduce traffic crash risk factors caused by senior drivers. A transportation safety specialist's main concern is how to determine the factors that are most likely to cause traffic crash [22-27]. The main objective of this study is to identify the main causality factors of older drivers' crash by building a machine learning-based model to predict the traffic crash injuries sever- ity of older people over 60. Highlighting the most effective risk factors can help transportation specialists better adapt transportation systems to reduce future crash and their severity. The proposed traffic crash risk factors prediction is achieved through applying five machine learning classification approaches, Na ̈ive Bayesian(NB), Decision Tree(DT), Logistic Regression (LR), Light-GBM, and Random Forest (RF) on Michigan freeways crash data to help Michigan transportation agencies face future challenges. The performance of the five methods is presented with the importance of crash risk factors according to the model with the highest accuracy. Several studies have investigated machine learning ap- proaches in transportation-related traffic crash either through employing the various types of Neural Networks or by ap- plying other machine learning techniques such as RF, NB, etc. The outcomes of those studies have had a wide range of performance results due to the differences in the data applied under each study [28-33]. Predicting injury severity risk factors by applying machine learning was always a topic of interest since machine learning algorithms proved their performance in prediction-related problems. A two-layer Neural Network was tested to predict the injury severity level but with relatively low accuracy of 73.5% [22-27]. Recently, researchers in [31-38], utilized 500 road crash samples from the Traffic Insurance Information Center (TRAMER) dataset in Istanbul. In their study, they focused on factors such as vehicle type (i.e., car, truck, bus, etc.), time zone, and we at her conditions. Ada Boost, CART, C4.5, Na ̈iveBayes, OneR, and IBk machine learning models are tested. The authors found that CART, IBk, C4.5, and Naive Bayes algorithms performed the best with an up to 81.5% accuracy. Although thorough data preprocessing was applied, results did not achieve higher than 81.5% because of the limited number of samples. On the other hand, 146,322 samples from the crash2014 dataset and 194,477 samples from the Casualties2014 dataset that belong to the Department for Transportation in the UK were investigated in [34-40]. The authors used H2O and WEKA mining tools to investigate the performance of Na ̈ive Bayes, C4.5, Random Forest, AdaboostM1, and Bagging classifiers. Results show C4.5 and Random Forest achieved the best accuracy of 87%. The study was for any age population, but it showed that age was the second most important factor in crash prediction. An agreement of RF outperformance was proved. Researchers investigated the accuracy of C4.5, IB, and RF machine learning methods in predicting the severity of injury in crash with focus on age and gender. They experiment with the Miami metropolitan area dataset in Florida provided by the Florida Department of Transportation (FDOT), which includes crash between 2008 and 2012. Projects evaluate and compare different approaches to modeling crash severity as well as investigating the effect of risk factors on the fatality outcomes of traffics crashes using machine learning- based driving simulation. They developed prediction models to identify risk factors of traffics crashes can be targeted to reduce accident. The RF model demonstrated the best performance from among the six different techniques with an accurate 82.6% researchers studied the prediction of crash risks and drivers' performance among elderly drivers in Japan that are at least 70 years old [1-16]. They use LR analysis and Support Vector Machine (SVM). Their experiments are conducted based on the license renewal test data from the National Police Agency of Japan. This test dataset consists of a driving simulator test and an on-road test. Authors state that non-linear analysis of LR classifies most of the cognitively impaired drivers. Additionally, authors indicate that unimpaired drivers have a strong tendency of becoming impaired and can be predicted earlier compared the accuracy of J48, ID3, CART, and Na ̈ive Bayes classifiers in predicting the severity of injury in traffic crash. Authors conduct their experiments on 3050 records of the UK traffic crash dataset with 12 factors. They conclude that the J48 performed the best with up to 96% accuracy. At the data preprocessing phase, the year of the crash was eliminated since it was considered as an irrelevant factor. Also, age was considered a redundant factor; therefore, it was removed. Authors in,compared the accuracy of AdaBoost, Logistic Regression (LR), Naive Bayes (NB), and Random Forests (RF) classifiers in predicting the severity of injury in traffic crash. The experimental results show that the accuracy of Random Forests is higher than the other three methods [17-24]. The results indicate that Random Forests has the potential to provide the best classification for predicting the injury severity of traffic crash with an average accuracy 75.50%. Using Convolutional Neural Network (CNN), the authors of proposed a TAP-CNN, a traffic crash prediction model. They studied 21 factors such as weather, traffic flow, and light to test their algorithm. They use a dataset of a section from the I-15 highway in the United States. Their algorithm is trained using 300 training samples and tested using 100 testing samples. Authors claim a 78.5% prediction accuracy that is 7.7% better prediction accuracy than the traditional Backprop- agation neural network model. The research is considered as a promising work but did not detail the data preprocessing phase. In like manner, compared between Na ̈ive Bayes, Deep Learning, and Gradient Boosting Trees classifiers perfor- mance in classifying crash' severity. Big data conducted using 1,018,204 records from DGT Spanish traffic agency with 58 factors or features. The results indicate that the Deep Learning classifier provides the best accuracy of 87%. Other related studies such as projects, presented Vehicle Crash Foresight System (VAFS) to predict future crash using a random DT algorithm. The system is tested using a dataset from the UK with crash between 1979 and 2016 with 25% for testing. Authors claim that their system achieves 75% prediction accuracy [25-31]. Researchers compared the accuracy of crash prediction of machine learning models against statistical models. The study is based on the State of Florida dataset with 326 freeway segments and 5538 crashes. Researchers use Multinomial Logit and Ordered Probability statistical models versus K-Nearest Neighbor, DT, SVM, and RF machine learning models. The study concludes that ma- chine learning models are better than statistical models in terms of prediction accuracy. Also, the RF model had the best prediction accuracy among all other models [32-40].

2.0 RESEARCH METHODOLOGY

In this paper, two machine learning models are applied to predict the severity of injury in traffic crash of elderly drivers. The methodological approach of this study includes three main phases as illustrated in Fig. 1. Data pre-processing phase takes place first before the training phase to make the raw data applicable. A cross-validation technique is used to ensure good training. Those models are tested and evaluated using various performance metrics in the third phase. The following subsections briefly discuss the research methods of the paper:

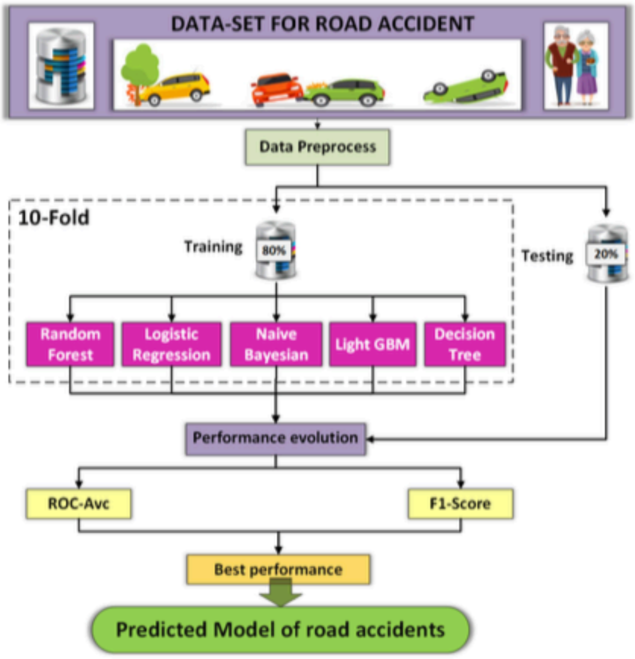


Fig. 1: Methodology Block Diagram

A. Data Source and Description

The Michigan Traffic Crash Facts (MTCF) dataset, which is provided by the Office of Highway Safety Planning in Michi- gan [19] is used. This dataset includes traffic crash records in the period from 2010 to 2017. A subset of 106,274 records that covers drivers of ages 60+ is selected. The age histogram is shown in Fig.2. This subset includes a comprehensive list of various environmental, roadway, vehicle, and human factors that were recorded by the police investigators at the crash scene. In this research, these factors are referred to as crash features. Car crash will be classified either as severe injury or as non-severe injury.

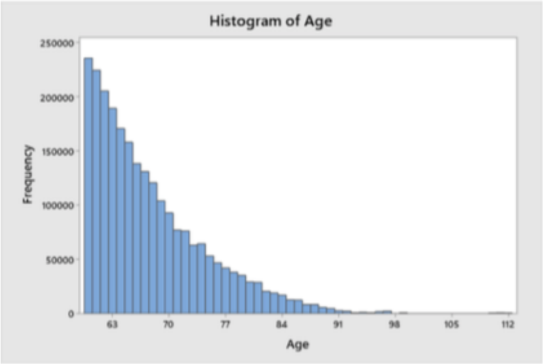


Fig. 2: Age histogram for 60+ drivers in the Michigan traffic

B. Data Pre-processing Phase

Table I demonstrates the statistical information for the selected factors of the 60+ driver's records in Michigan traffic data. In the pre-processing phase missing values were recov- ered by considering mean and mode values since the used dataset has a low missing rate. Table II lists nine crash features that are considered in this study, where N refers to continuous numerical values and C refers to non-continuous categorical values. The Car Age feature is calculated and added to the dataset by computing the difference between the crash year and the car manufacturing year.

TABLE I: Data statistical description

Variable	N	N*	Percent	Mean	StDev	Minimum	Median	Maximum	Skewness
Injuary	84914	48	99.944	0.12344	0.32895	0	-	1	-
LightCondition	106085	0	100.000	2.1335	1.6616	1.0000	1.0000	6.0000	0.92
Speed Limit	105536	549	99.482	66.902	6.741	5.000	70.000	90.000	-2.40
Vehicle year	39403	66682	37.143	2006.7	6.51	1900.0	2007.0	2018.0	-2.26
Age	36280	0	100.000	66.864	6.242	60.000	65.000	89.000	1.13
Alchol Ind	103552	2533	97.612	0.01064	0.10261	0.000000	0.000000	1.00000	9.54
Traffic volume	106085	0	100.000	50580	28680	2308	48700	111047	0.19
Crash Year	106085	0	100.000	2013.5	1.81	2010.0	2013.0	2017.0	0.27
GenderNum	106085	0	100.000	0.46967	0.49908	0.00000	0.00000	1.00000	0.12

The values of the features range differently; therefore, maximum and minimum standardization for Speed Limit, Age, Car Age, and Traffic Volume features are applied and then scaled to the range of [0, 1]. This normalization makes the weight of each feature dimension have the same influence on the objective function and improves the convergence speed of the iterative solution. SMOTE (Synthetic Minority Oversampling Technique) is applied to the dataset to solve the problem of having an unbalanced dataset that could cause the applied models to be overfitted. The SMOTE is proven to be a powerful balancing technique [20]. The original dataset contains 10,517 sever injury records and 74,585 of non-sever injury records. After applying SMOTE, the dataset became balanced by having equal number of both injury types. After dealing with outliers and missing values followed by normalizing and balancing the whole dataset, it is randomly divided into two experimental datasets of 80% training dataset and 20% testing dataset. The training dataset is applied during the second phase to train the machine learning model, while the testing dataset is used during the third phase to evaluate the performance of trained models.

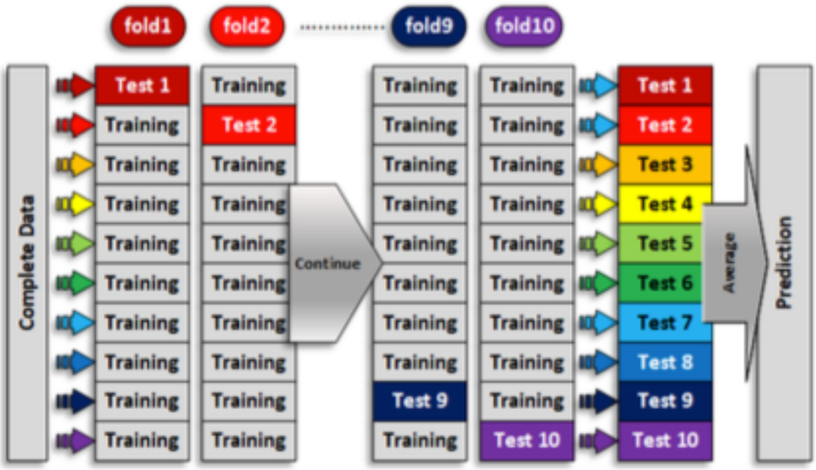
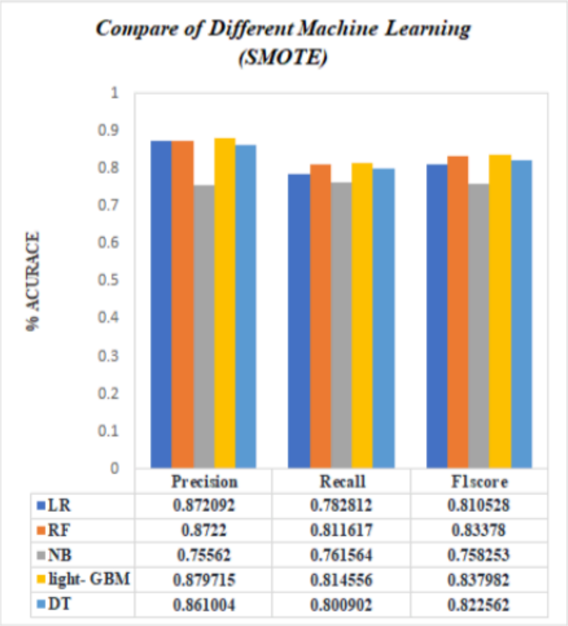


Fig. 3: Implemented Ten Cross Validation Process

3.0 RESULT

In order to evaluate the prediction performance of a model in regard to the binary data of 1 (injury) and 0 (no injury), the trained model was applied to the test data to compute its area under the curve (AUC) value. Therefore, both ROC and con- fusion matrix measures are used, and AUC was calculated. To compare the performances of the five models, several metrics are used; Precision, Recall/Sensitivity, F-Measure, ROC, and AUC. Fig. 4 shows that the highest Precision, Recall, and F- Measure are 0.879%, 0.814%, and 0.837% respectively to light GBM followed by 0.872%, 0.811%, and 0.833% respectively to RF, then by 0.0.873%, 0.782%, and 0.810% respectively to LR. All these results are achieved with SMOTE method and 10-folds cross-validation. The best result of our model can achieve the F1 score of 0.837. Therefore, the results of models accuracy show that the Light-GBM and RF classifier achieves the highest performance of 87.54% and 87.16% respectively, while the NB classifier achieves the lowest performance of 79.56%. In particular, the ROC performance has increased by 0.48% over the best ROC performance, which is achieved by the Light-GBM classifier with 87.54%.



	Precision	Recall	F1score
■ LR	0.872092	0.782812	0.810528
■ RF	0.8722	0.811617	0.83378
■ NB	0.75562	0.761564	0.758253
■ light- GBM	0.879715	0.814556	0.837982
■ DT	0.861004	0.800902	0.822562

Fig. 4: Comparison of Different Machine Learning Algorithms

This study demonstrated that machine learning models built from highly meaningful features using Michigan Traffic Crash dataset were able to achieve high recall and accuracy for classifying elder driving at risk of injury. After an accurate comparison between our algorithms, we noticed that RF and light-GBM achieved a higher efficiency of approximately 0.87. Fig. 5 shows the ROC performance of the Five Models.

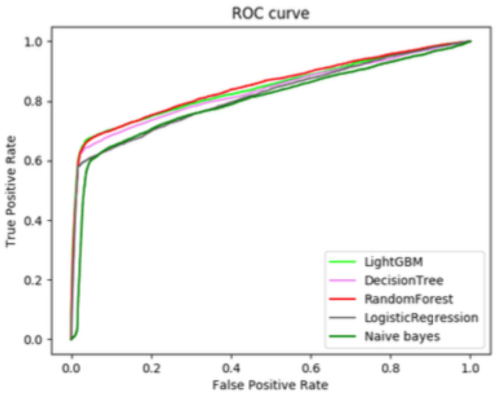
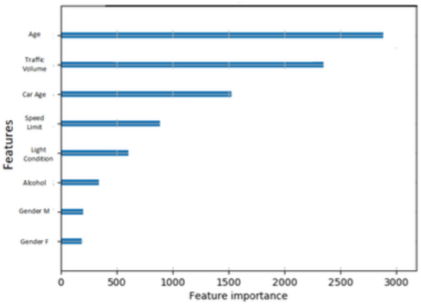


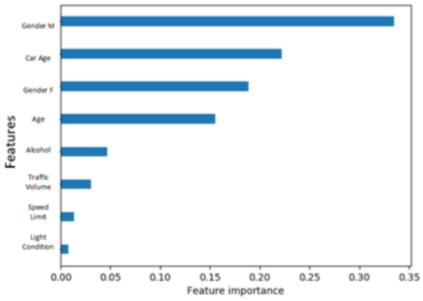
Fig. 5: ROC Curves of the Five Models

According to the three highest models' performance, the features importance rate are illustrated in Fig 6 (a,b,c), that shows the feature importance according to light-GBM, RF, and DT respectively. RF and DT have nearly agreed on the order of the features importance, while Light-GBM voted differently. According to Light-GBM Fig. 6 (a), the Age is the most important factor to predict the injury severity. The Ttraffic Volume is the second factor, while the Car Age is the third. Thus, a deep look at the data showed that the injury risk is 3% higher when using cars of 10 years or more age. RF and DT showed that genders are important factors. Returning back to the original data, car crash by male drivers have a higher probability of about 0.89 of the total recorded crash with the severe injury probability of 0.10. On the other hand, female drivers have recorded a lower number of crash but with a higher number of severe injury of 0.15. This gives a clue that female drivers are 5% more likely to be affected severely by car crash. The results also show that the age has been the most significant parameter in evaluating the level of severity associated with fixed object crashes among elderly drivers. Following these three contributing factors, traffic value, car age, and speed limit have been identified as the most important variables in the developed Light-GBM, respectively. This helps to identify gaps and improve public safety towards improving the overall highway safety situation of older drivers.



(a) Light-GBM.

In general, light-GBM classifier achieved the highest accuracy among all the classifiers. Therefore, there is no doubt that Random Forster algorithms has an advantage over the other three representative classifiers.



(b) Random Forest

5.0 CONCLUSION

In this paper, we have explored machine learning-based modeling techniques that are NB, DT, LR, Light-GBM, and RF, to investigate Michigan crash data. A pre-processing procedure is applied to overcome the missing data and to have the data normalized. The data was subsequently divided into a training set and a test set. SMOTE technique was used to produce a balanced training data.

The highest accuracy is achieved by light-GBM models with an accuracy of up to 87.97%. The recommended predictive model light-GBM can be used to identify the key factor causing crashes in injury severity for the Elderly. This model will help Michigan Traffic Agencies to be more proactive in combating high-risk of traffic injury to elder drivers. Results have also confirmed that the most importation features are the Age, and Traffic volume. Age and cars' Age have also shown great influences. In light of the results, the research recommends utilizing newer car models for elder drivers.

For future work, a comparative study which implements deep learning can be performed in transportation fields to calculate the risk score of traffic injury to elder driving for many other factors.

REFERENCES

[1] Dimitrijevic, Branislav, et al. Segment-Level Crash Risk Analysis for New Jersey Highways Using Advanced Data Modeling. No. CAIT-UTC-NC62. Rutgers University. Center for Advanced Infrastructure and Transportation, 2020.

[2] Li, Chang, et al. "Machine learning for text mining based on prediction of occupational accidents and safety risk calculation." Australian Journal of Engineering and Applied Science 13.6 (2020): 11-17.

[3] Bahrami, Javad, Viet B. Dang, Abubakr Abdulgadir, Khaled N. Khasawneh, Jens-Peter Kaps, and Kris Gaj. "Lightweight implementation of the lowmc block cipher protected against side-channel attacks." In *Proceedings of the 4th ACM Workshop on Attacks and Solutions in Hardware Security*, pp. 45-56. 2020.

[4] Chen, Lee, et al. "Machine learning established by using crowdsourced investigation vehicle data for forecast of expressway crash risk ." International Journal of Applied Science and Information Science 11.8 (2020): 356-363.

[5] Ahmadinejad, Farzad, Javad Bahrami, Mohammd Bagher Menhaj, and Saeed Shiry Ghidary. "Autonomous Flight of Quadcopters in the Presence of Ground Effect." In *2018 4th Iranian Conference on Signal Processing and Intelligent Systems (ICSPIS)*, pp. 217-223. IEEE, 2018.

[6] Zhang, Lixuan, et al. "Machine Learning Models established toward the Car Smash Injury Difficulty." European Journal of Applied Engineering and Basic Sciences 19.17 (2020): 4678-4685.

[7] Bozorgasl, Zavareh, and Mohammad J. Dehghani. "2-D DOA estimation in wireless location system via sparse representation." In *2014 4th International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 86-89. IEEE, 2014.

[8] Xiang, Zheng, et al. "Machine Learning process for injury severity prediction and Traffic accidents classification." International Journal of Management System and Applied Science 23.12 (2020): 997-1003.

[9] Amini, Mahyar, and Aryati Bakri. "Cloud computing adoption by SMEs in the Malaysia: A multi-perspective framework based on DOI theory and TOE framework." Journal of Information Technology & Information Systems Research (JITISR) 9.2 (2015): 121-135.

[10] Silva, Philippe Barbosa, Michelle Andrade, and Sara Ferreira. "Machine learning applied to road safety modeling: A systematic literature review." Journal of traffic and transportation engineering (English edition) 7.6 (2020): 775-790.

[11] Amini, Mahyar. "The factors that influence on adoption of cloud computing for small and medium enterprises." (2014).

[12] AlMamlook, Rabia Emhamed, et al. "Comparison of machine learning algorithms for predicting traffic accident severity." 2019 IEEE Jordan international joint conference on electrical engineering and information technology (JEEIT). IEEE, 2019.

[13] Amini, Mahyar, et al. "Development of an instrument for assessing the impact of environmental context on adoption of cloud computing for small and medium enterprises." Australian Journal of Basic and Applied Sciences (AJBAS) 8.10 (2014): 129-135.

[14] Rezapour, Mahdi, Amirarsalan Mehrara Molan, and Khaled Ksaibati. "Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and logistic regression models." International journal of transportation science and technology 9.2 (2020): 89-99.

[15] Amini, Mahyar, et al. "The role of top manager behaviours on adoption of cloud computing for small and medium enterprises." Australian Journal of Basic and Applied Sciences (AJBAS) 8.1 (2014): 490-498.

[16] Rahim, Md Adilur, and Hany M. Hassan. "A deep learning based traffic crash severity prediction framework." Accident Analysis & Prevention 154 (2021): 106090.

[17] Amini, Mahyar, and Nazli Sadat Safavi. "Critical success factors for ERP implementation." International Journal of Information Technology & Information Systems 5.15 (2013): 1-23.

[18] Siebert, Felix Wilhelm, and Hanhe Lin. "Detecting motorcycle helmet use with deep learning." Accident Analysis & Prevention 134 (2020): 105319.

[19] Amini, Mahyar, et al. "Agricultural development in IRAN base on cloud computing theory." International Journal of Engineering Research & Technology (IJERT) 2.6 (2013): 796-801.

[20] Yang, Yang, et al. "Identification of dynamic traffic crash risk for cross-area freeways based on statistical and machine learning methods." Physica A: Statistical Mechanics and Its Applications 595 (2022): 127083.

[21] Amini, Mahyar, et al. "Types of cloud computing (public and private) that transform the organization more effectively." International Journal of Engineering Research & Technology (IJERT) 2.5 (2013): 1263-1269.

[22] Fu, Yuchuan, et al. "A decision-making strategy for vehicle autonomous braking in emergency via deep reinforcement learning." IEEE transactions on vehicular technology 69.6 (2020): 5876-5888.

[23] Amini, Mahyar, and Nazli Sadat Safavi. "Cloud Computing Transform the Way of IT Delivers Services to the Organizations." International Journal of Innovation & Management Science Research 1.61 (2013): 1-5.

[24] Alkinani, Monagi H., Wazir Zada Khan, and Quratulain Arshad. "Detecting human driver inattentive and aggressive driving behavior using deep learning: Recent advances, requirements and open challenges." Ieee Access 8 (2020): 105008-105030.

[25] Amini, Mahyar, and Nazli Sadat Safavi. "A Dynamic SLA Aware Heuristic Solution For IaaS Cloud Placement Problem Without Migration." International Journal of Computer Science and Information Technologies 6.11 (2014): 25-30.

[26] Wahab, Lukuman, and Haobin Jiang. "Severity prediction of motorcycle crashes with machine learning methods." International journal of crashworthiness 25.5 (2020): 485-492.

[27] Amini, Mahyar, and Nazli Sadat Safavi. "A Dynamic SLA Aware Solution For IaaS Cloud Placement Problem Using Simulated Annealing." International Journal of Computer Science and Information Technologies 6.11 (2014): 52-57.

[28] Muhammad, Khan, et al. "Deep learning for safe autonomous driving: Current challenges and future directions." IEEE Transactions on Intelligent Transportation Systems 22.7 (2020): 4316-4336.

[29] Sadat Safavi, Nazli, et al. "An effective model for evaluating organizational risk and cost in ERP implementation by SME." IOSR Journal of Business and Management (IOSR-JBM) 10.6 (2013): 70-75.

[30] Cai, Qing, et al. "Applying a deep learning approach for transportation safety planning by using high-resolution transportation and land use data." Transportation research part A: policy and practice 127 (2019): 71-85.

[31] Sadat Safavi, Nazli, Nor Hidayati Zakaria, and Mahyar Amini. "The risk analysis of system selection and business process re-engineering towards the success of enterprise resource planning project for small and medium enterprise." World Applied Sciences Journal (WASJ) 31.9 (2014): 1669-1676.

[32] Wahab, Lukuman, and Haobin Jiang. "A comparative study on machine learning based algorithms for prediction of motorcycle crash severity." PLoS one 14.4 (2019): e0214966.

[33] Sadat Safavi, Nazli, Mahyar Amini, and Seyyed AmirAli Javadinia. "The determinant of adoption of enterprise resource planning for small and medium enterprises in Iran." International Journal of Advanced Research in IT and Engineering (IJARIE) 3.1 (2014): 1-8.

[34] Ziakopoulos, Apostolos, and George Yannis. "A review of spatial approaches in road safety." Accident Analysis & Prevention 135 (2020): 105323.

[35] Safavi, Nazli Sadat, et al. "An effective model for evaluating organizational risk and cost in ERP implementation by SME." IOSR Journal of Business and Management (IOSR-JBM) 10.6 (2013): 61-66.

[36] Mannering, Fred, et al. "Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis." Analytic methods in accident research 25 (2020): 100113.

[37] Khoshraftar, Alireza, et al. "Improving The CRM System In Healthcare Organization." International Journal of Computer Engineering & Sciences (IJCES) 1.2 (2011): 28-35.

[38] Mokhtarimousavi, Seyedmirsajad. "A time of day analysis of pedestrian-involved crashes in California: Investigation of injury severity, a logistic regression and machine learning approach using HSIS data." Institute of Transportation Engineers. ITE Journal 89.10 (2019): 25-33.

[39] Abdollahzadegan, A., Che Hussin, A. R., Moshfegh Gohary, M., & Amini, M. (2013). The organizational critical success factors for adopting cloud computing in SMEs. Journal of Information Systems Research and Innovation (JISRI), 4(1), 67-74.

[40] Silva, Philippe Barbosa, Michelle Andrade, and Sara Ferreira. "Machine learning applied to road safety modeling: A systematic literature review." Journal of traffic and transportation engineering (English edition) 7.6 (2020): 775-790.