Stacking Ensemble Learning For Traffic Accident Severity Prediction

Hazhir Salari[1] [0000−0001−8611−3652]
and Seyed Amin Hosseini Seno[2] [0000−0002−0838−1800]

[1] Ferdowsi University of Mashhad, Mashhad, Iran
hazhir.salari@gmail.com
hazhir.salari@um.ac.ir
[2] Ferdowsi University of Mashhad, Mashhad, Iran
hosseini@um.ac.ir

Abstract. In the present day, people rely more heavily on transportation systems than ever before. Analysis of past accident data reveals that transportation systems consistently pose a threat to human life and property. In cases of severe accidents, it is necessary to alert an emergency center near the accident site, in addition to the police center, during the vital time period, in order to save lives and minimize casualties. Obviously, sending alert to the emergency center is not required for non-severe accidents. This article aims to identify key features and create a stacked ensemble learning model, utilizing two models - LightGBM and XGBoost, to identify severe accidents. Based on the evaluation findings, the proposed model outperformed recent works, obtaining higher levels of accuracy (85.75%), precision (86.89%), recall (84.22%), and f1-score (85.54%).

Keywords: ITS · Road Safety · ML · Ensemble Learning · Stacking.

1 INTRODUCTION

Due to the ever-increasing human need for transportation and vital dependence on vehicles in meeting human needs, the number of vehicles increases every day, which has caused an increase in road accidents and sometimes causes irreparable injuries and damages [1].

Road accidents are usually investigated for two reasons: the first reason is that accidents are a significant risk to human life or endangering human physical health. according to the official statistics of world health organization (WHO) the number of annual deaths caused by road accidents exceeds 1.3 million people, and also, as a result of the accidents, around 20 to 50 million people are injured or disabled every year. On the other hand, the leading cause of death of children and young people between the ages of 5 and 29 is accidents [2]. Financial loss is the second reason for investigating accidents. In developed countries, 2 percent, in developing countries, 1.5 percent, and in underdeveloped countries, 1 percent of gross national product (GNP) are paid to financial damages caused by accidents. For a better understanding, 68 billion dollars is the share of underdeveloped countries for accident damages, which is a staggering cost [3]. Therefore, anticipating traffic accidents and preventing them are important steps for car safety and it is very important to provide a basic solution for this problem. Intelligent transportation system (ITS) is based on the IoT, which represents the convergence of the transportation system, communication, and computer science, which is used as a solution for road safety and road management [4–6].

In order to solve the concern of accidents and sharing information among vehicles, ITS offers the concept of vehicular ad-hoc network (VANET) and internet of vehicles (IoV) [7, 8]. VANET is a special type of wireless network that enables vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication. In addition to improving traffic flow, VANET has provided robust solutions for road and vehicle safety [9]. One the other hand, the goal of IoV is to facilitate communication with various networks beyond the two connections that were specified for VANET [10].

By examining the accidents that have occurred, it was observed that accidents do not happen by chance, but these accidents have patterns that can be analyzed, predicted, and prevented from occurring according to these patterns [3]. While it is crucial to forecast the occurrence of accidents, accurately predicting the severity is also essential for prompt medical care of those who are injured. Furthermore, forecasting the severity of the accident can aid in managing traffic as well [11]. Machine learning (ML) is a member of the artificial intelligence (AI) family that enables a system to have the ability to learn and add knowledge and experience using the least amount of human intervention [12]. ML provides useful tools that are efficient for data generation, communication, and data sharing in ITS [13]. Machine learning can be used in various parts of ITS such as safety applications, traffic management, information applications, routing, mobility management, security, etc [14].

Some works have conducted research in the field of accident severity prediction based on a single model, while others have designed and evaluated ensemble learning models for this purpose.

1.1 Single Model Learning:

Single model learning encompasses research endeavors that have solely employed a single ML model, such as decision trees (DT), support vector machine (SVM), backpropagation (BP), or other models, which does not fall under the category of ensemble learning (EL) [15].

1.2 Ensemble Learning:

EL aims to create a novel model by combining two or more ML models, with the expectation that it will outperform the individual models used in the ensemble learning [16]. In addition, ensemble learning itself has learning frameworks such as bagging, boosting and stacking [17].

LightGBM and XGBoost: Gradient Boosting Decision Tree (GBDT) is an ensemble learning model with a boosting approach that builds a classifier with better performance by combining several decision trees. Both LightGBM and XGBoost utilize the GBDT algorithm, which is widely employed in various research and practical applications. However, each of these models has unique strengths that make them more suitable in different circumstances [18, 19].

1.3 Contributions

The contributions introduced in this article can be described as follow:

- Present a new ensemble learning model to predict accident severity.
- Comparing the proposed model with recent works in this field.

1.4 Paper Structure

section 2 of this study will present a literature review of the research conducted in this field. Following that, in Section 3, we will introduce the dataset used in our study. In Section 4, we will describe our proposed system, and in Section 5, we will analyze the evaluation of our model and compare it with other researches. Finally, we will discuss our conclusions in section 6.

2 Literature Review

Based on the algorithms used in recent works, we have divided the articles into two categories as follows:

- Based on EL.
- Based on single model learning.

Ensemble Learning category: The aim of [20] was not only to predict the accident severity but also to develop a complete framework for road safety. In this study, 9 different predictions were developed in the form of a directed acyclic graph (DAG), the output result of some predictions is used as input for the next prediction, and for these predictions, Logistic Regression (LR), DT, LightGBM, Random Forest (RF), Convolutional Neural Network (CNN), Long-short Term Memory (LSTM), and Multilayer Perceptron (MLP) models were implemented, and in the end, decision tree-based algorithms performed better than other algorithms. In this study, a recommender system was also developed for different conditions driving conditions. Research [3] developed an accident prediction model based on ensemble learning using the combination of RF and CNN models. In this study, two approaches were used to compare the presented model with other models. In the first approach, using the entire data set features for training and testing, while in the second approach, the number of 20 effective features were calculated using the random forest feature importance. The proposed method of this study is depicted in Fig. 1. As a challenge and

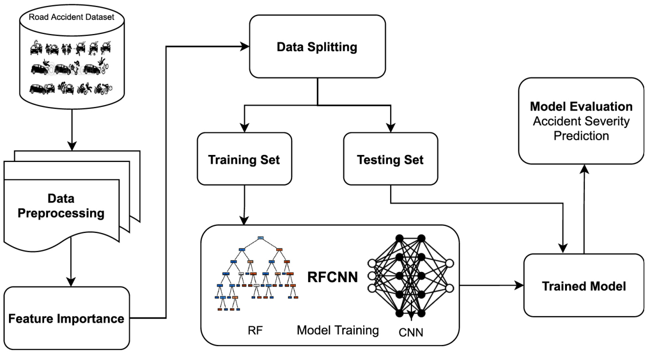

Fig. 1. RFCNN Workflow [3]

problem that can be stated in the mentioned research is that combining two models increases the complexity compared to when only one model is used.

Single Model Learning category: To study the 2016 accidents in Italy, researchers examined three methods: all accidents, non-severe accidents, and severe accidents. The aim was to identify the factors that contributed the most to the accident severity using LR, due to the significant financial damage caused by accidents in the country [21]. The study [22] viewed vehicles as part of the IoT in transportation. The paper explored various accident prediction techniques, including DT, RF, Artificial Neural Network (ANN), k-nearest neighbors (KNN), and LR, to find a proactive and highly accurate method using crowdsourced information to predict accidents before they happen. After comparisons, the decision tree was identified as the most accurate method for accident prediction. Researchers in [23] used traffic accident data from the National Police Agency of Taiwan to study accidents occurring within a 50-meter radius of intersections. While predicting low-risk accidents showed relatively good accuracy, the highly skewed distribution of the three-category data resulted in a lack of training data for medium and high-risk intersections, leading to lower prediction accuracy. In this scenario, the MLP and Deep Neural Network (DNN) models showed the best performance. The UK accidents dataset was used in [24] to apply a CNN model, which can automatically extract features from the large volume of data collected in the VANET to predict traffic accidents. Finally, the results of simulations and experiments have been compared with the BP model, which showed its superiority according to the evaluation metrics (loss and accuracy). In the data set reviewed by [25], due to the fact that the number of fatal accidents in the available data set was small, fatal accidents were merged with injury accidents and to determine the effect of each independent variable on the dependent variable (accident severity) used logistic regression for accidents and finally using forward stepwise 11 variables were selected as influential variables and after training the model it was found that the developed ANN model had better results. In [26], to de-correlate a large amount of data in VANET, they used the principal component analysis (PCA) algorithm and used the BP algorithm to train the prediction model. Also, they used the 2016 UK car accident dataset. Finally, they compared their proposed method with the basic SVM and basic BP models, and they were able to show that their proposed method performed better. The aim of [27] was to design a more comprehensive accident prediction system based on Hidden Markov Model (HMM) which considers several parameters related to the accident at the same time. This paper proposed an Accident Prediction System (APS) for VANET in urban environments, where crash risk is a latent variable that can be observed using several observations such as speed, weather conditions, crash location, traffic congestion, and driver fatigue.

3 Dataset Description

As previously stated, analyzing the collected accident data revealed that accidents are not random occurrences, but instead have identifiable patterns that can be utilized to construct models for predicting accidents and their severity. This section will now present the dataset employed in this study. Several factors such as changing the driving lane, fatigue or lack of alertness of the driver, weather condition, vehicle speed, day of the week, type of road, urban or rural area, etc. have an effect on the accident occurrence and on the accident severity, which endangers the safety of passengers [12, 26, 27]. This research utilizes the accident data of France, consisting of four distinct data sets, as illustrated in the Fig. 2 [28, 29].

As can be seen in Fig. 2, this data set consists of 4 separate data sets, each of which is briefly mentioned below. Users data set: the information of users who were present in an accident is included in this data set, and accident_id can be used to identify users present in an accident. Vehicles dataset: Information related to the vehicles involved in an accident and details about the vehicle's maneuver at the time of the accident are included in this dataset. Characteristics data set: Information about the time of the accident and environmental conditions can be found in this data set. Places dataset: information related to the characteristics of roads and streets were placed.

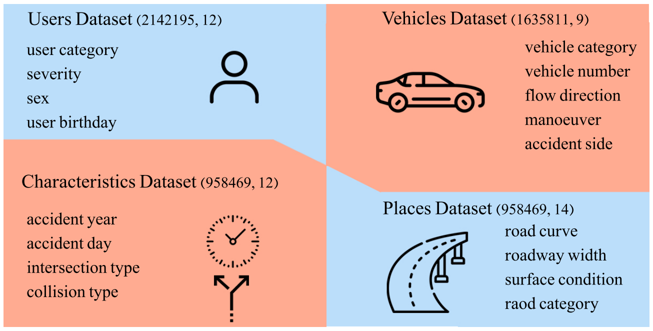

Fig. 2. France Accidents Dataset Description

Before incorporating this dataset into ML models, the four distinct data sets must be combined and subjected to data pre-processing procedures in order to create a refined dataset. The subsequent section will delve into these procedures in detail.

4 Proposed System

This section is divided into two main subsections of data preprocessing and the ML model explanation. A summary of the proposed system is illustrated in Fig. 4.

4.1 Data Cleaning

The conversion of low-quality data into high-quality data using various methods is crucial for achieving more reliable results and making precise decisions [30]. For the preprocessing of the data set used in this study, the following steps have been performed in order:

- **Data Merging**: To determine whether an accident is severe or not, those users and those vehicles present have selected which had higher severity in the users dataset and vehicles dataset, respectively. Next, according to the accident_id column, we merged these 4 datasets, and in the end, the obtained dataset had 958,462 rows.
- **Missing Values:** To address the issue of missing values in this dataset, we first removed any columns with more than 10 percent missing values. For the remaining columns with missing values, we implemented different strategies: we used the LightGBM model to predict the missing values for some columns, while for others that exhibited high variability in the values, we opted to fill in the missing values by the most frequent values according to the dispersion of the data.
- **Categorical Data:** Given the large number of columns with categorical values in this dataset, we began by replacing the values in some of these columns, such as department, destination, street number, and others, with their respective values frequency. Next, we applied label encoding to convert all categorical columns into numerical values.
- **Data Scaling:** In order to improve the performance of learning models and reduce the impact of outlier data, it is necessary that the input data to the learning models is standardized.
- **Class Imbalance:** Class imbalance means that in the classification problem, one of the classes of the dependent variable in the dataset has a higher frequency (for example, the frequency of 1s is more than 0s), considering that the ML models mostly consider data classes distribution uniformly by default. In, these cases, a step should be added to the data pre-processing to solve the challenge of class imbalance in order to avoid model bias usnig Synthetic Minority Oversampling Technique (SMOTE) library.

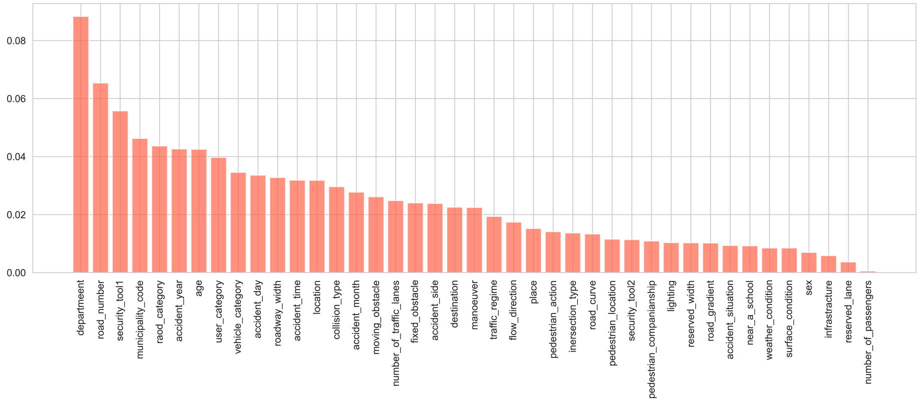

Fig. 3. Random Forest Feature Importance

- **Feature Selection:** Considering the large number of columns in this data set, it has been tried to select the most relevant columns using random forest feature importance in this study. Due to the selection of relevant columns, the model can perform faster and focus more on relevant features in the learning stages. Finally, 30 features were selected.

4.2 Model Explanation

As discussed in the section 2, different models have been presented to predict the severity of accidents. Now, in this section, we are trying to develop an stacking ensemble learning model so that we can use the advantages of different models in predicting the accidents severity. Stacking is a two-layer model, in the first layer, two or more basic learning models perform the prediction process separately. Finally, the output of the basic models is given to the meta-learning model as new data to predict the final value (severity). In our proposed model, LightGBM and XGBoost models are considered as basic learning models, and LightGBM is considered as a meta-learning model.

In order for the model to have its best performance, we have adjusted a number of parameters of each of these models separately using the GridSearchCV library. The input parameters and the best parameters can be seen in the table 1.

Table 1. Base Models Hyperparameter Tuning Values.

Base Model	Hyperparameter	Hyperparameter Values	Best Hyperparameter Value
XGBoost	number of estimators	[50, 100, 150, 200]	150
LightGBM	number of estimators	[50, 100, 150, 200]	200
XGBoost	maximum trees depth	[10, 15, 20, 25, 30]	10
LightGBM	maximum trees depth	[10, 15, 20, 25, 30]	20

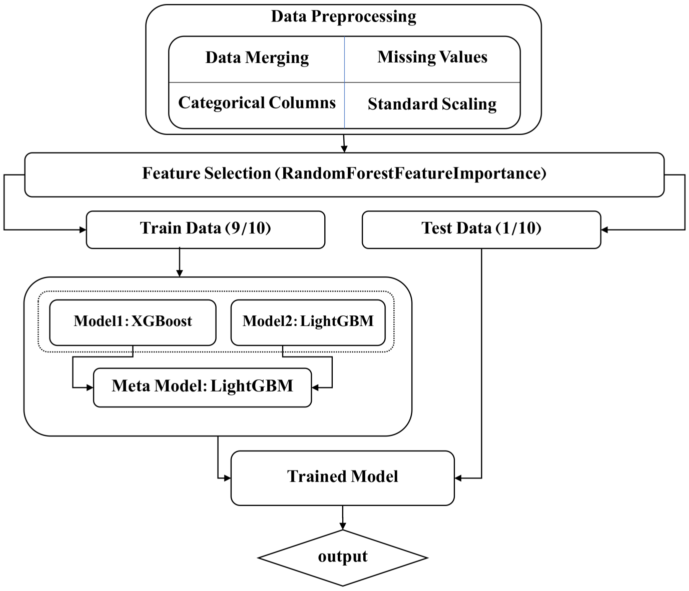

Fig. 4. proposed-system

5 Performance Evaluation

This study utilized the Sickit-Learn, LightGBM, and XGBoost libraries to develop the model, and the model was run on a computer equipped with an Intel Core i7 10750H processor and 16GB of RAM.

In order to evaluate the proposed system with the work done in this context, we use 4 criteria: accuracy, precision, recall and f1-score, the formula of each of which is as follows:

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{1}$$

$$Precision = \frac{TP}{(TP + FP)} \tag{2}$$

$$Recall = \frac{TP}{(TP + FN)} \tag{3}$$

$$F1 - Score = 2 * \frac{(Precision \times Recall)}{(Precision + Recall)} \tag{4}$$

The comparison of the evaluation results between the model proposed in this research and the model presented in [20], as demonstrated in Table 2, reveals an enhancement in the evaluation metrics. One of the factors behind this enhancement is the utilization of a EL model that integrates various basic models,

Table 2. Evaluaion Results.

works	Year	Dataset	Accuracy	Precision	Recall	F1-Score
[11]	2022	France Accidents, 2005-2018	82.5	81.9	76.7	79.9
Our-Work	2023	France Accidents, 2005-2018	85.75	86.89	84.22	85.54
[20]	2021	France Accidents, 2005-2017	-	73	69	71
Our-Work	2023	France Accidents, 2005-2017	-	86.81	84.01	85.39

which are reputed for their exceptional performance on extensive datasets in the domain of classification problems. Additionally, the meta model employed in this study is distinct from [20], and it has resulted in a marginal improvement in the performance.

This research has a strong pre-processing stage, where missing values are predicted using a machine learning model. Another benefit is the use of an ensemble learning model instead of a single model, which utilizes base models known for their excellent performance in classification problems. Table 2 shows the evaluation results of the proposed model in this research compared to the one in research [11].

6 Conclusion

This study aimed to enhance the prediction of accident severity by using an ensemble learning model in order to determine whether it is necessary to dispatch an emergency team to save the lives of vehicle occupants within critical time. Furthermore, it is possible to suggest novel models that enhance the efficacy of the current models and assess them based on various factors such as the computational complexity of the model, its speed, and memory requirements.

References

1. Shendekar, S., Thorat, S., Rojatkar, D.: Traffic accident prediction techniques in vehicular ad-hoc network: A survey. In: 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI). pp. 652–656. IEEE (2021)
2. Road traffic injuries. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries, Last accessed 2 Feb 2023
3. Manzoor, M., Umer, M., Sadiq, S., Ishaq, A., Ullah, S., Madni, H.A., Bisogni, C.: Rfcnn: Traffic accident severity prediction based on decision level fusion of machine and deep learning model. IEEE Access **9**, 128359–128371 (2021). https://doi.org/10.1109/ACCESS.2021.3112546
4. Alvi, U., Khattak, M.A.K., Shabir, B., Malik, A.W., Muhammad, S.R.: A comprehensive study on iot based accident detection systems for smart vehicles. IEEE Access **8**, 122480–122497 (2020). https://doi.org/10.1109/ACCESS.2020.3006887
5. Gholamhosseinian, A., Seitz, J.: Safety-centric vehicle classification using vehicular networks. Procedia Computer Science **191**, 238–245 (2021). https://doi.org/10.1016/j.procs.2021.07.030, https://doi.org/10.1016/j.procs.2021.07.030

6. Vizzari, D., Bahrani, N., Fulco, G.: Coexistence of energy harvesting roads and intelligent transportation systems (its). Infrastructures **8**(1), 14 (2023)
7. Bibi, R., Saeed, Y., Zeb, A., Ghazal, T.M., Rahman, T., Said, R.A., Abbas, S., Ahmad, M., Khan, M.A.: Edge ai-based automated detection and classification of road anomalies in vanet using deep learning. Computational intelligence and neuroscience **2021**, 1–16 (2021)
8. Rani, P., Sharma, R.: Intelligent transportation system for internet of vehicles based vehicular networks for smart cities. Computers and Electrical Engineering **105**, 108543 (2023). https://doi.org/https://doi.org/10.1016/j.compeleceng.2022.108543, https://www.sciencedirect.com/science/article/pii/S0045790622007583
9. Sheikh, M., Liang, J., Wang, W.: Security and privacy in vehicular ad hoc network and vehicle cloud computing: A survey. Wireless Communications and Mobile Computing **2020**, 1–25 (01 2020). https://doi.org/10.1155/2020/5129620
10. Ali, E.S., Hasan, M.K., Hassan, R., Saeed, R.A., Hassan, M.B., Islam, S., Nafi, N.S., Bevinakoppa, S.: Machine learning technologies for secure vehicular communication in internet of vehicles: recent advances and applications. Security and Communication Networks **2021**, 1–23 (2021)
11. Tamim Kashifi, M., Ahmad, I.: Efficient histogram-based gradient boosting approach for accident severity prediction with multisource data. Transportation research record **2676**(6), 236–258 (2022)
12. Hossain, M.A., Noor, R.M., Yau, K.L.A., Azzuhri, S.R., Zaba, M.R., Ahmedy, I.: Comprehensive survey of machine learning approaches in cognitive radio-based vehicular ad hoc networks. IEEE Access **8**, 78054–78108 (2020)
13. Gillani, M., Niaz, H.A., Tayyab, M.: Role of machine learning in wsn and vanets. International Journal of Electrical and Computer Engineering Research **1**(1), 15–20 (2021)
14. Mchergui, A., Moulahi, T., Zeadally, S.: Survey on artificial intelligence (ai) techniques for vehicular ad-hoc networks (vanets). Vehicular Communications **34**, 100403 (2022)
15. Dietterich, T.G., et al.: Ensemble learning. The handbook of brain theory and neural networks **2**(1), 110–125 (2002)
16. Pintelas, P., Livieris, I.E.: Special issue on ensemble learning and applications (2020)
17. Zhang, Z., Yang, W., Wushour, S.: Traffic accident prediction based on lstm-gbrt model. Journal of Control Science and Engineering **2020**, 1–10 (2020)
18. Zhang, D., Gong, Y.: The comparison of lightgbm and xgboost coupling factor analysis and prediagnosis of acute liver failure. IEEE Access **8**, 220990–221003 (2020)
19. Liang, W., Luo, S., Zhao, G., Wu, H.: Predicting hard rock pillar stability using gbdt, xgboost, and lightgbm algorithms. Mathematics **8**(5), 765 (2020)
20. Boujemaa, K.S., Berrada, I., Fardousse, K., Naggar, O., Bourzeix, F.: Toward road safety recommender systems: Formal concepts and technical basics. IEEE Transactions on Intelligent Transportation Systems **23**(6), 5211–5230 (2021)
21. Eboli, L., Forciniti, C., Mazzulla, G.: Factors influencing accident severity: an analysis by road accident type. Transportation research procedia **47**, 449–456 (2020)
22. Mohanta, B.K., Jena, D., Mohapatra, N., Ramasubbareddy, S., Rawal, B.S.: Machine learning based accident prediction in secure iot enable transportation system. Journal of Intelligent & Fuzzy Systems **42**(2), 713–725 (2022)
23. Lin, D.J., Chen, M.Y., Chiang, H.S., Sharma, P.K.: Intelligent traffic accident prediction model for internet of vehicles with deep learning approach. IEEE Transactions on Intelligent Transportation Systems **23**(3), 2340–2349 (2021)

24. Zhao, H., Cheng, H., Mao, T., He, C.: Research on traffic accident prediction model based on convolutional neural networks in vanet. In: 2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD). pp. 79–84. IEEE (2019)
25. Najafi Moghaddam Gilani, V., Hosseinian, S.M., Ghasedi, M., Nikookar, M.: Data-driven urban traffic accident analysis and prediction using logit and machine learning-based pattern recognition models. Mathematical problems in engineering **2021**, 1–11 (2021)
26. Zhao, H., Mao, T., Yu, H., Zhang, M., Zhu, H.: A driving risk prediction algorithm based on pca-bp neural network in vehicular communication. In: 2018 10th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC). vol. 2, pp. 164–169. IEEE (2018)
27. Aung, N., Zhang, W., Dhelim, S., Ai, Y.: Accident prediction system based on hidden markov model for vehicular ad-hoc network in urban environments. Information **9**(12), 311 (2018)
28. Annual databases of road traffic injury accidents - years from 2005 to 2021. https://www.data.gouv.fr, Last accessed 2 Feb 2023
29. Accidents in france from 2005 to 2016. https://www.kaggle.com/datasets, Last accessed 2 Feb 2023
30. Porwal, S., Vora, D.: A comparative analysis of data cleaning approaches to dirty data. International Journal of Computer Applications **62**(17), 30–34 (2013)