A Comprehensive Study of Road Traffic Accidents: Hotspot Analysis and Severity Prediction Using Machine Learning

Utkarsh Gupta
Dept. of Computer Science and Engineering, PES University
Bengaluru, India
utkarsh348@gmail.com

Varun MK
Dept. of Computer Science and Engineering, PES University
Bengaluru, India
mkvarun2001@gmail.com

Gowri Srinivasa
PES Center for Pattern Recognition
Dept. of Computer Science and Engineering
PES University, Bengaluru, India
gsrinivasa@pes.edu

***Abstract*—This study analyses road traffic accident data recorded over a period of time to gain insights to the underlying pain points in the infrastructure and policies. Such insight allows us to focus our efforts in the right direction to make the lives of people safer. The data includes various geographical and meteorological factors affecting the severity of these accidents. We use Kernel density estimation (KDE) plots to analyse hotspots of accident-prone areas weighed against severity over years to understand the evolution of these dangerous zones. Furthermore, we use machine learning algorithms to predict the accident severity given certain parameters and to understand the factors that have a major influence on the severity of the accident. We have studied a publicly available dataset of road traffic accidents in the UK as a proof of concept of the pipeline to understand the underlying patterns of accidents occurring in a region of interest.**

***Index Terms*—hotspots, KDE, exploratory data analysis, SMOTE, machine learning, random forest, feature importance**

I. Introduction

A steep increase in population and motorisation has led to a rising trend in road accidents. Road traffic injury is the eighth leading cause of death of the global level and the leading cause of death for young people aged 15-29 years. One of the primary causes being the increase in road transport in comparison to less progress in other types of transportation systems and insufficient infrastructures in Iran, has significantly increased the urban pollution, road users wasted time and above all the damages caused by traffic accidents [1], [2].

Efforts taken to prevent road accidents face the imperative problem of answering the question of where to implement safety precautionary measures. We answer this question by identifying “Hot spots”. “Hot Spots” or “Black Spots” are locations identified by a high accident occurrence compared with the other locations [3].

This is why it is important for us to analyse where and when these accidents occur frequently. The locations, which are identified by a high accident occurrence compared with the other locations, are known as hotspots or black spots. Occurrences of accidents are not random in space and time. They depend on factors such as geometric design, severe weather conditions, time of the day, etc.

Plotting, spotting and cataloguing these hotspots helps us locate the most accident-prone areas and where the focused efforts, energy and resources of the concerned authorities are required to diagnose the core issues and help counter the said issues. Simply plotting the individual crash sites on a map does not work either as this does show a strong concentration of incidents at certain locations but is vague in the information it provides, giving no density information. KDE (Kernel Density Estimation) plots are employed and weighed against a severity index to understand the precarious nature of certain locations. They are also plotted over time to understand the evolution of these hotspots.

A number of factors affecting the severity of an accident are interrelated. This study also aims at employing various statistical ML models to predict this severity with the factors taken into consideration being vehicle and road conditions. The relationship between these parameters not only help predict severity but also give us information as to what exactly influences the severity of these accidents the most.

The models used on the data are Naive-Bayes, Logistic regression, AdaBoost, XGBoost and Random Forest classifiers to classify the data based on different levels of prediction.

II. Previous Work

A study of past work has shown that many traditional statistical methods have been used to detect hotspots [3], [4]. In these models, spatial characteristics of hotspots were modelled as a constant for a given period of time which is not true [5]. This dataset provides a means to analyse the surrounding environment, severe weather conditions which can further be used to calculate severity index or simply take these factors into account during model building which shows a significant improvement in ranking/detection of hotspots [4], [5]. GIS-based statistical analytic techniques like KDE (Kernel

978-1-6654-9291-1/22/$31.00 ©2022 IEEE

Density Estimation) to analyse hotspots according to time intervals and seasons with and without taking SI (Severity Index) into consideration. This paper was able to conclude that SI is strongly related to seasons, time of the day. Comap was used to visualise all these various results. It is another example of how time and space are taken into consideration. Their study concluded that traffic accidents occurred frequently twice in a day and they were able to pinpoint the intersections [6]. Customised spatial weight matrices have been used to detect hotspots too. In this procedure, an Inverse Network Distance-Band Spatial Weights Matrix of Intersections (INDSWMI) is built to form the network of roads by taking road network constraints into consideration. Then the k-nearest Distance Band Spatial Weights Matrix Between Crash and Intersection (KDSWMCI) using the adjacency crashes for each intersection. These are then used to perform Intersection Hotspot Analysis (IHA) with the measure of statistic being Getis-Ord Gi* statistics and the accuracy was measured using Intersection Prediction Accuracy Index (IPAI) [7]. KDE has also been compared with other methods in some studies. In one study, KDE and Kriging were used for hotspot detection with the measure of accuracy being Prediction Accuracy Index (PAI).

It was found that the list of hotspots identified by the two methods were moderately different. It relates crashes as a function of potential variables to make its predictions [8]. Other unsupervised machine learning methods like k-means have also successfully profiled vehicle accident hotspots [9].

Statistical models which predict the severity of the accident exploit the relationship between the different parameters in question. Artificial Neural Networks (ANN) have also been used in this regard. As with any statistical model, they have a set of assumptions and will perform poorly in case these assumptions are violated. An ANN model trained to take the number of drivers and vehicles as input performed very well to predict the number of fatalities in accidents compared to previous models [10]. By performing feature analysis and ranking of variables in terms of efficiency for an ANN trained to predict the number of road accidents in Tehran-Qom freeway, it was shown that the mean daily traffic volume and mean speed of the agents have the greatest impact on these accidents [11].

Regression has also been employed to estimate quantities related to accidents. Injuries, fatalities and number of accidents. They also used Genetic Algorithm models for the same and concluded that ANN worked the best for their data [12], [13]. A study used data from New South Wales, Australia for an ordered logit model and ordered probit models to estimate the probability of injury and death given conditions like seating position, blood alcohol level, vehicle type, make of collision, etc [14]. But these models, with the exception of ANNs, have assumptions which do not hold in real life. This is why the Bayesian Network (BN) model was implemented since BNs are capable of making predictions without the need for presumptions and can also be used to make graphical representations of complex relationships. The factors which affected their classification the most were accident type, lighting and number of injuries [15], [16].

A study aimed to improve the power of the SVM model to predict crash severity, given vehicle and road condition attributes, by using 15 crash-related parameters in a Fuzzy C-Means (FCM) clustering algorithm [17]. It also explored 3 other models: FNN (Feed-Forward Neural Networks), FNN-FCM and SVM (Support Vector Machines).

Contributions of the present work:

- A detailed analysis of traffic accidents and their nature.
- Plotting accident hotspots to recognise smaller areas that prove to be relatively more dangerous.
- Classification of accidents into severity levels using ML models.
- Comparison of the performance of the ML models on the dataset and describing features with the highest impact based on the best performing model.
- Corroborating the findings from exploratory data analysis of the dataset and model predictions with findings from predecessor research.

This comprehensive look at traffic accident data serves as a great first step towards planning and shaping proposing action to fix the issues that plague a road transport network. The study is also presented in a way that makes it very suitable to translate to another dataset.

III. Data Description

As a proof of concept for a study of this nature, we use the publicly available dataset detailing road accidents that occurred in 2005-2017 in the UK [18]. The data shows us a lot of interesting information about the accident that we can choose to visualize such as the number of accidents, the conditions and status of the environment at the time, severity of these accidents etc. The dataset is separated into two different sets, one pertaining to accident information with 34 columns and 2047256 rows and the second set detailing data regarding the vehicles involved in accidents with 24 columns and 1488981 rows. To use the two different datasets in conjunction, we had to join them based on the accident

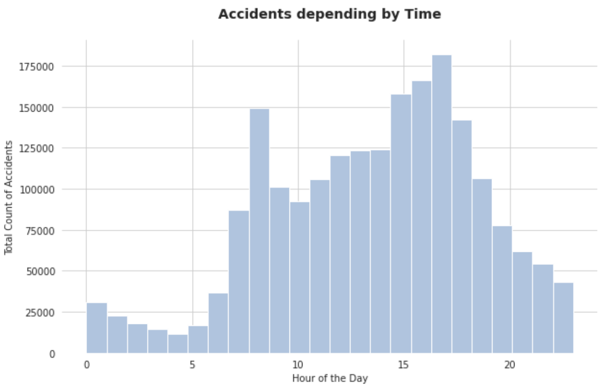

Fig. 1. A plot of number of accidents by the time of day.

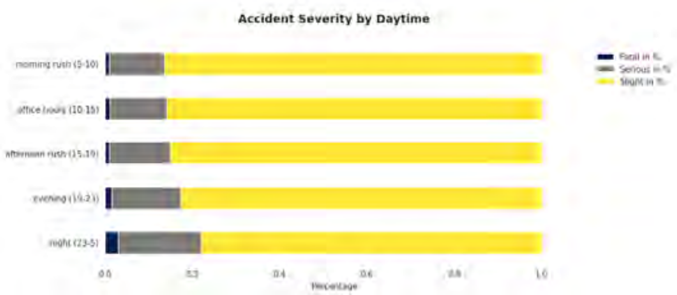

Fig. 2. Severity of accidents plotted against the time of day.

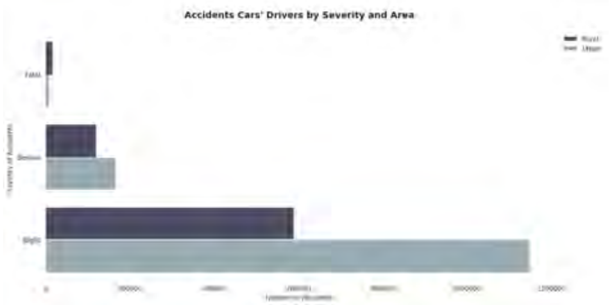

Fig. 3. Severity of accidents in rural and urban areas.

index to be sure we were in studying accidents for which data was recorded in both sets, thus maximising our knowledge of that particular accident. The accidents themselves are labelled as “Slight”, “Serious” or “Fatal”.

Fig. 1 shows a plot of the number of accidents occurring during the time of day by the hour. We can observe from the graph that the number of accidents are lower in number during the morning and evening, while the number peaks during the afternoon. This can be due to the fact that there are more vehicles on the road at this time. Fig. 2 illustrates a plot of the severity of accidents occurring during the day. The graph lends more insight as we can observe, the number of accidents that are serious or fatal in nature increase towards the night, arguably when the driving conditions are much worse. Fig. 3 shows the plot of the number of accidents occurring in rural and urban areas by severity. We note that the number of serious and slight accidents in urban areas are more, which may be due to the fact that there may be more sophisticated road networks and greater number of vehicles in urban areas. However, the number of fatal accidents occurring in rural areas is more than in urban areas. This tells us that the areas of the country we distinguish as rural need further planning and development policies.

IV. Kernel Density Estimation

KDE is a non-parametric spatial statistical tool that estimates probabilities, probability density functions for random variables. It is a solution to a smoothing problem to get inferences from the data. We aim to analyze and plot the hotspots of all such accidents so as to pinpoint what are the points of concern in the highway network. In the dataset, all road accidents are classified under three levels of severity: Slight, Serious and Fatal. It is vital that we do not treat all records the same but give more weight to the locations that proved to be more dangerous overall by forming a severity index [6]. This shortcoming is accentuated by the fact that the more severe accidents are vastly outnumbered by the minor, slight accidents. Here we choose to allot quartic weights to each severity level.

This can skew our understanding and analysis of the map when our focus is looking for hazardous locations. Another observation that we have noted is that there is an overall decrease in the total number of accidents every year.

This is good but it also means that plotting all accidents for a certain location can only give a part of the information. It becomes quite important to then use KDE to plot accident hotspots for certain periods of time to not just record all hazardous locations but also to take notice of how these hotspots evolved over time.

This temporal evolution of hotspots can be interpreted in three different ways corresponding to how they change over time [19].

- Type 1 - There is dissipation in hotspots at certain locations either due to changes in infrastructure or applications of new safety protocols.
- Type 2 - Indicates hotspots that have shown no change in their characteristics over time and continue to identify dangerous locations.
- Type 3 - Appearance of new hotspots or increase in density of ones present already, suggesting worsening conditions.

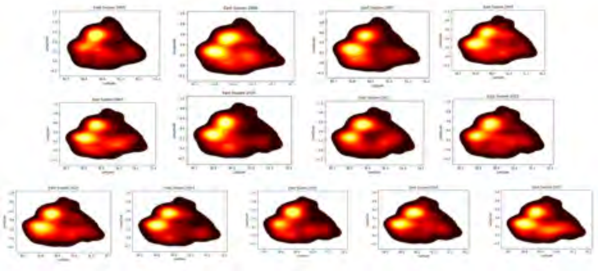

Fig. 4. Weighted KDE plots of accidents in East Sussex.

A. Case Study 1: East Sussex

Fig. 4 shows the weighted KDE plot for East Sussex. There are two important observations to make of the East Sussex study:

- There is an appearance of a prominent hotspot at the coordinates +0.3 longitude, +50.8 latitude, suggesting some dangerous change over time during the decade that needs to be dealt with.
- Another matter of concern is the worsening conditions at the coordinates +0.1 longitude, +51 latitude. This depicts

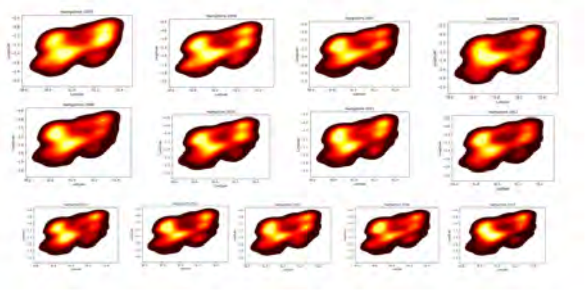

Fig. 5. Weighted KDE plots of accidents in Hampshire.

a location that could prove to be more hazardous in the future if not handled well.

B. Case Study 2: Hampshire

- Fig. 5 shows the weighted KDE plot for Hampshire. The bigger hotspot centered around -1.3 longitude, +50.9 latitude remains more or less the same size with the same intensity over the years. It is also the location with visibly the highest severity of accidents in the general area. This means that not much is being done about this issue or the tactics tried thus far have proved to be unsuccessful
- The hotspot centered around -1.0 longitude, +51.3 latitude has slowly dissipated over time from 2005 to 2017. This is good because if we need models to study and implement in other similar locations, it is helpful to understand where exactly certain measures work and the reason(s) they do.

V. Accident severity prediction

The next step in understanding this dataset is to predict the severity of the accidents taking place based on the parameters recorded. The datasets separates different records into different levels of severity - Slight, Serious, Fatal. We use classification models - a division of machine learning algorithms to sort different accident records into these levels and essentially predict the severity. We have compared the performance of five small data machine learning models - Naive-Bayes classifier, Logistic Regression, AdaBoost, XGBoost and Random Forest Classifier for the prediction task. In the final step, we analyze the results and infer the attributes in the dataset that have the most impact on damage due to accidents.

A. Data Preprocessing

A huge majority of the parameters that are a part of the dataset are categorical attributes which requires us to encode them. We use the ordinal encoder to encode each of these attributes. Another point of concern is that this dataset is extremely imbalanced i.e.as expected, the number of "Fatal" and "Serious" accidents are much less in number when compared to "Slight" accidents.

To tackle this issue, we use an imbalanced learning sampling method SMOTE (Synthetic Minority Oversampling Technique). This is a data augmentation technique wherein instead of creating copies of minority classes to oversample, we synthesize new records(being relatively similar) from the data. The drawback of this issue is that due to oversampling this way, training times are increased by a lot which makes it difficult to tune hyperparameters.

B. Naive Bayes Classifier

Categorical NB classifies discrete features which assume categorical distribution for each feature where the features have to be encoded using label encoding techniques such that each category would be mapped to a unique number. In Multinomial Naive Bayes, to estimate weights, only a single class c is used. On the other hand, in Complement Naive Bayes Algorithm, all training data of the classes except c class is used.

C. Logistic Regression

Logistic regression is a classification model built around the logistic function to predict the probability of the occurrence of a particular class. The model works best when the classes are linearly separable and particularly distinct in their characteristics. Similar to linear regression, the model is fit on the data using different kinds of optimisation algorithms, the decision for which can be made by considering the nature and size of the data. The model is implemented with cross validation to automatically select better hyperparameters along with the saga solver considering the presence of more than two classes and the size of the final dataset.

D. AdaBoost and XGBoost

AdaBoost or "Adaptive Boosting" is a boosting ensemble machine learning algorithm that trains multiple weak classifiers sequentially. It is an iterative method to chain these weak classifiers and correct the incorrect classifications made by the previous model. Weak classifiers are better than random classification but not accurate enough to be useful. Decision trees are usually the default stumps or weak classifiers at each step and the cumulative result is one of a much more accurate and strong model than any one individual model used. Weights dependent on the error rate are given to each classification based on if it was classified correctly or not, to mark these incorrect classifications for the next classifier to work on. We train XGBoost with default parameters using gbtree as the booster [20].

E. Random Forest

The Random Forest classifier model is also an ensemble machine learning model wherein instead of boosting, there are individual trees in this "forest", each acting as an independent predictor. A majority voting scheme is then applied to all the individual classifiers and based on this vote is the final classification allotted. An important point to note about the Random Forest model and its functionality is that each of these individual trees work independently and are not correlated in their working i.e, each tree is different and is fit on a unique set of attributes from the dataset.

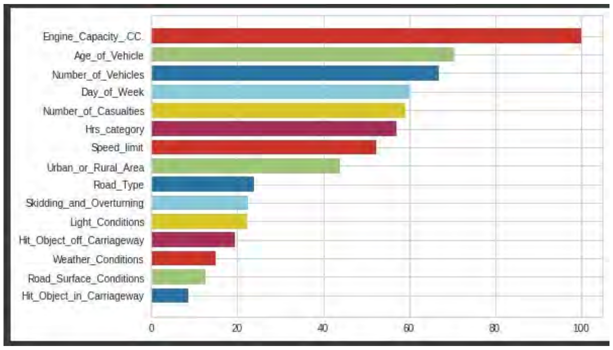

Fig. 6. Feature Importance from Random Forest model.

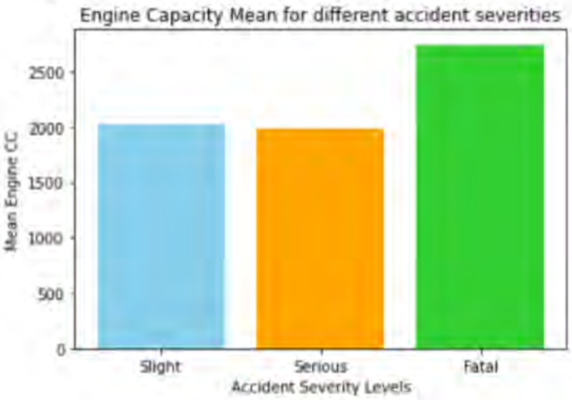

Fig. 7. A plot of Mean Engine CC vs accident severity.

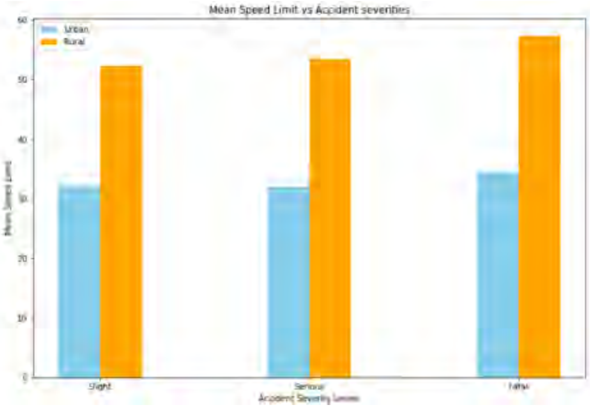

Fig. 8. A plot of mean speed limit vs accident severity in rural and urban areas.

TABLE I
PREDICTION METHODS AND THEIR ACCURACY

Prediction Method	***Accuracy***	***F1-Score***
Complement Naive Bayes	0.42	0.33
Bernoulli Naive Bayes	0.47	0.51
Categorical Naive Bayes	0.57	0.63
XGBoost without SMOTE	0.81	0.91
XGBoost with SMOTE	0.81	0.91
Random Forest without SMOTE	0.86	0.93
Random Forest with SMOTE	0.95	0.93

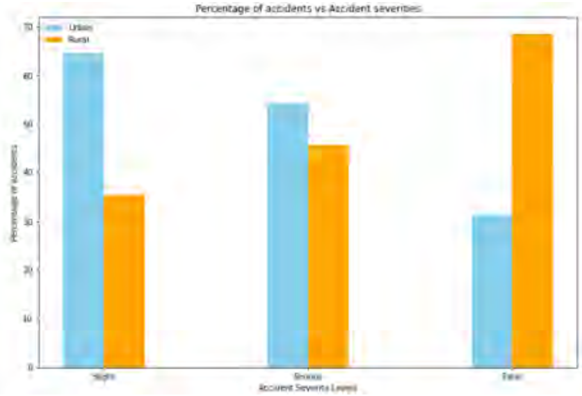

Fig. 9. Percentage of accidents vs accident severity.

VI. INSIGHTS FROM THE DATA

In Table 1, we chose accuracy and f1 score to evaluate the model over recall or precision. Recall would tell us how many accidents were correctly predicted, and precision tells us the ability to correctly identify the accidents(True Positives). We want to be precise but avoid false negatives, being cautious stands to our advantage. While accuracy would also satisfy these conditions, the f1 score gives the best balance for minority classes and acts as a balance between recall and precision.

Each of the models were tried with stratified values and with SMOTE to tackle the extreme imbalance in the data. The Naive-Bayes and Logistic regression models did not perform well on this dataset. Naive-Bayes, presumably, due to it assuming the independence of features in the dataset. From Table 1, we can observe logistic regression delivered 86% without and close to 49% with SMOTE while Naive Bayes gave a mean accuracy of 57%. Boosting models, AdaBoost specifically gave a decent accuracy on the stratified split but performed poorly while trying to classify the minority classes, namely “Serious” and “Fatal”. After being fit on the SMOTE balanced dataset, the average accuracy dropped but its performance on the more hazardous classes increased.

We infer that the accuracy of the XGBoost model from Table 1 is 70% which is the same as our Adaboost implementation.With a balanced dataset, we see a good improvement in accuracy for the minority classes as well as the overall accuracy, where random forest gives a final mean accuracy of 95%.

Feature Importance is a technique that is used to determine how important an input attribute is at determining/predicting the target attribute. Each input attribute is given a score based on its importance. We plot the feature importance graph for our Random Forest model which gives us the feature importance on how likely it is for an input attribute to affect the accident severity.

A. Feature importance

Fig. 6 shows the feature importance results calculated using the random forest model and we can observe that the engine capacity or CC of the vehicles involved seems to have the largest influence on the prediction, followed by the age of the vehicle, traffic density (number of vehicles, day of the week, etc.), the mean speed limit in the vicinity of the accident and the area type where the accident occurred. We have also observed there is a relation between the time of the day with the severity of the accident. This is consistent with the findings reported in literature [21].

Mean engine capacity. Fig. 7 shows the plot of the mean engine cc against the severity of accidents. We observe that the vehicles involved in fatal accidents have engines with higher power output than vehicles in less severe incidents.

Mean speed limit. Fig. 8 shows the plot of the mean speed limit in the area of the accident against the severity and area type of the accident - rural or urban. We notice a consistent pattern that the severity of accidents occurring increases as the speed limit increases, with the speed limit being much higher in rural areas. The relationship between speed limits and the severity of injuries in accidents was studied by Doecke *et al.* [22]; they came to the conclusion that a positive exponential relationship between fatality rates and speed limits was observed.

Rural and urban area accidents. Fig. 9 shows the plot of the percentage of accidents of a certain severity occurring in rural and urban areas. We observe that while the percentage of less serious accidents occurring in urban areas is more, the share of fatal accidents occurring rural areas is much higher. This insight allows initial plans for infrastructure development policies to take place more easily. Brown *et al.* [23] discuss related data extensively in their study wherein they acknowledge the falling rates of accidents in all areas but the accidents in rural areas remain higher than urban areas.

VII. Conclusion

This study presents a pipeline to analyze road accident data to recognize the most dangerous and concerning areas in the road network and to recognise the factors that can lead to severe and fatal accidents. We explore computational techniques that allow us to gain an insight into the issues that can be addressed by the respective personnel to improve the general safety of people on the street. Finally, a corroboration of some of our observations with available literature lends credence to the analysis pipeline. We can replicate this study to any region of interest to gain insights and drive projects to strengthen the infrastructure and safety of road transport.

Reproducible research: To facilitate reproducing the results in this study, all the code and data used to obtain the results reported in this study are available online [24].

References

[1] K. Zimmerman, D. Jinadasa, B. Maegga, and A. Guerrero, "Road traffic injury on rural roads in tanzania: measuring the effectiveness of a road safety program," *Traffic injury prevention*, vol. 16, no. 5, pp. 456–460, 2015.

[2] I. Bargegol, V. N. M. Gilani, M. Ghasedi, and M. Ghorbanzadeh, "Delay modeling of un-signalized roundabouts using neural network and regression," *Computational Research Progress in Applied Science & Engineering*, vol. 2, no. 1, pp. 28–34, 2016.

[3] I. Thomas, "Spatial data aggregation: exploratory analysis of road accidents," *Accident Analysis & Prevention*, vol. 28, no. 2, pp. 251–264, 1996.

[4] J. Choudhary, A. Ohri, and B. Kumar, "Spatial and statistical analysis of road accidents hot spots using gis," in *3rd Conference of Transportation Research Group of India (3rd CTRG)*, 2015.

[5] V. Prasannakumar, H. Vijith, R. Charutha, and N. Geetha, "Spatio-temporal clustering of road accidents: Gis based analysis and assessment," *Procedia-social and behavioral sciences*, vol. 21, pp. 317–325, 2011.

[6] K. G. Le, P. Liu, and L.-T. Lin, "Determining the road traffic accident hotspots using gis-based temporal-spatial statistical analytic techniques in hanoi, vietnam," *Geo-spatial Information Science*, vol. 23, no. 2, pp. 153–164, 2020.

[7] Z. Zhang, Y. Ming, and G. Song, "A new approach to identifying crash hotspot intersections (chis) using spatial weights matrices," *Applied Sciences*, vol. 10, no. 5, p. 1625, 2020.

[8] L. Thakali, T. J. Kwon, and L. Fu, "Identification of crash hotspots using kernel density estimation and kriging methods: a comparison," *Journal of Modern Transportation*, vol. 23, no. 2, pp. 93–106, 2015.

[9] C. Sinclair and S. Das, "Traffic accidents analytics in uk urban areas using k-means clustering for geospatial mapping," in *2021 International Conference on Sustainable Energy and Future Electric Transportation (SEFET)*. IEEE, 2021, pp. 1–7.

[10] O. F. Cansiz, "Improvements in estimating a fatal accidents model formed by an artificial neural network," *Simulation*, vol. 87, no. 6, pp. 512–522, 2011.

[11] F. Bagheri Khalili, A. Sheikholeslami, and A. Mahmoudabadi, "Variable efficiency appraisal in freeway accidents using artificial neural networks—case study," in *CICTP 2012: Multimodal Transportation Systems—Convenient, Safe, Cost-Effective, Efficient*, 2012, pp. 2657–2664.

[12] A. A. E. Doğan and A. P. ANgüngör, "Estimating road accidents of turkey based on regression analysis and artificial neural network approach," *Advances in transportation studies*, vol. 16, pp. 11–22, 2008.

[13] A. P. Akgüngör and E. Doğan, "An artificial intelligent approach to traffic accident estimation: Model development and application," *Transport*, vol. 24, no. 2, pp. 135–142, 2009.

[14] S. Rajalin and H. Summala, "What surviving drivers learn from a fatal road accident," *Accident Analysis & Prevention*, vol. 29, no. 3, pp. 277–283, 1997.

[15] J. De Oña, R. O. Mujalli, and F. J. Calvo, "Analysis of traffic accident injury severity on spanish rural highways using bayesian networks," *Accident Analysis & Prevention*, vol. 43, no. 1, pp. 402–411, 2011.

[16] M. Simoncic, "A bayesian network model of two-car accidents," *Journal of transportation and Statistics*, vol. 7, no. 2/3, pp. 13–25, 2004.

[17] "Uk road safety: Traffic accidents and vehicles," https://www.kaggle.com/datasets/tsiaras/uk-road-safety-accidents-and-vehicles.

[18] K. Assi, S. M. Rahman, U. Mansoor, and N. Ratrout, "Predicting crash injury severity with machine learning algorithm synergized with clustering technique: A promising protocol," *International journal of environmental research and public health*, vol. 17, no. 15, p. 5497, 2020.

[19] M. Bíl, R. Andrášik, and J. Sedoník, "A detailed spatiotemporal analysis of traffic crash hotspots," *Applied geography*, vol. 107, pp. 82–90, 2019.

[20] "A gentle introduction to xgboost for applied machine learning," https://www.machinelearningmastery.com.

[21] S. M. Rifaat and H. C. Chin, "Accident severity analysis using ordered probit model," *Journal of advanced transportation*, vol. 41, no. 1, pp. 91–114, 2007.

[22] S. D. Doecke, C. N. Kloeden, J. K. Dutschke, and M. R. J. Baldock, "Safe speed limits for a safe system: The relationship between speed limit and fatal crash rate for different crash types," *Traffic Injury Prevention*, vol. 19, no. 4, pp. 404–408, 2018, pMID: 29323934. [Online]. Available: https://doi.org/10.1080/15389588.2017.1422601

[23] L. H. Brown, A. Khanna, and R. C. Hunt, "Rural vs urban motor vehicle crash death rates: 20 years of fars data," *Prehospital Emergency Care*, vol. 4, no. 1, pp. 7–13, 2000. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1090312700700669

[24] "A comprehensive study of road traffic accidents: Hotspot analysis and severity prediction using machine learning [source code]," https://github.com/VarunMK/UK-Road-Traffic-Hotspot-Analysis-Repo.