Spatio-temporal Statistical Analysis and Deep Learning Techniques for Traffic Accidents Prediction

*Note: Sub-titles are not captured in Xplore and should not be used

Diena Al-dogom*‡, Nour Aburaed*, Mina Al-Saad*, Saeed Almansoori†
*College of Engineering and IT, University of Dubai
† Mohammed Bin Rashid Space Center (MBRSC)
Dubai, UAE
Email: ‡daldogom@ud.ac.ae

***Abstract*—Traffic accidents impose significant problems in our daily life due to the huge social, environmental, and economic expenses associated with them. The rapid development in data science, geographic data collection, and processing methods encourage researchers to evaluate, delineate traffic accident hotspots, and to effectively predict and estimate traffic accidents. In this study, Kaggle traffic accidents dataset that covers United Kingdom for the time period between 2012-2014 will be investigated. Our methodology consists of three main techniques. First, Morans I method of spatial autocorrelation, and Getis-Ord Gi* statistics will be used to examine and relate traffic accidents dataset in terms of spatial and temporal features. Second, weighted features will be used as inputs for Deep Feedforward Neural Network (DFFNN). Finally, the performance of the proposed DFFNN will be evaluated based on its accuracy, misclassification rate, precision, prevalence, histogram of errors, and confusion matrix. These evaluation metrics are then used as a comparison basis against the performance of Support Vector Machine (SVM). The results will focus on using spatial statistics techniques to effectively weight different features according to their contribution to traffic accidents. Consequently, the output of the DFFNN asserts the likelihood of accident occurrence given a certain location. Furthermore, it would be beneficial to investigate whether these accidents exhibit certain timely patterns, such as certain days or months where accidents potentially occur more frequently. The proposed method can be effectively used by different authorities to implement an improved planning and management approaches for traffic accident reduction. Moreover, it can identify and locate road risk segments where immediate action should be considered.**

***Index Terms*—traffic accidents, machine learning, decision trees, gradient boosting, GIS, spatial analysis, temporal analysis**

I. Introduction

With the rapid booming of urbanization accompanied with the increasing number of vehicles, traffic accidents have become a recurring problem that researchers continuously sought to solve. Companies progressively improve automobiles' safety measures in such a way that reduces accidents and the number of fatalities resulting from them. Unfortunately, accidents remain unavoidable. With the recent rise of the fields of big data and machine learning, they have been utilized to solve and boost the performance of many crucial areas and industries that seemed otherwise unsolvable or solved with unreliable accuracy, such as in health, education, and transportation. In the latter field, traffic accidents severity prediction is of particular interest for the scope of this paper. Severity is can be broadly categorized as fatal, serious, and slight. Therefore, it can be handled as a pattern recognition problem, where statistical techniques and machine learning algorithms can be used to predict the severity [1], [2]. This type of prediction is considered to be highly non-linear due to the amount of factors involved in the prediction, such as road type and surface, weather, and light conditions. Hence, the advantage of using machine learning methods as opposed to traditional statistical techniques is the ability to handle non-linear problems as well as obtaining a general solution that works for a wide variety of data. There have been several attempts in the literature to analyze traffic flow, accidents, and predict accident severity using machine learning paradigms and Geographic Information System (GIS) [1], [3]–[11]. The goal of these studies is to integrate GIS with spatial and temporal analysis in order to reduce accident fatalities by as much as possible. The accuracy of the used machine learning paradigm is crucial to the effectiveness of the study. Artificial Neural Networks (ANNs), which are a subset of machine learning, have proven high efficiency in solving prediction and classification of non-linear problems. In [2], the authors used a Recurrent Neural Network (RNN) to predict accident severity by using 1130 accident records gathered over the period of 2009-2015 in North-South Expressway (NSE), Malaysia. The authors also compare their proposed RNN model against Multilayer Perceptron (MLP) and Bayesian Logistic Regression (BLR). The validation accuracy of the RNN model was 71.77%, whereas the MLP and BLR models achieved 65.48% and 58.30%, respectively. A similar study was conducted in [12], where the authors developed a model for traffic accidents prediction based on Recurrent Neural Networks,

Identify applicable funding agency here. If none, delete this.

particularly Long-Short Term Memory (LSTM). The authors gathered traffic accidents data from Beijing between 2016-2017, and developed a model that learns connections between traffic accidents and their spatiotemporal patterns. The authors demonstrate that the accuracy of their model is better than others such as Decision Tree Regression (DTR) and Autoregressive Moving Average Model (ARMA). In [13], the authors combined supervised and unsupervised learning to improve prediction performance of traffic accidents data that was obtained from Knox County in Tennessee. The unsupervised part performs feature learning, and the supervised part fine-tunes the model to improve traffic accident prediction. The latter part also includes Multivariate Negative Binomial (MVNB) model as a regression layer in order to address heterogeneity issues. The authors state that their model provides 84.58% over deep learning models that do not use regression layers, and 158.27% over Support Vector Machine (SVM) model. In [14], the authors introduced intersection accident prediction model based on Backpropagation Neural Network, which establishes a relationship between accident forms and their factors. The considered accident forms are single vehicle accident, rear-end accident, front collision accident, side collision accident, scratch accident, and the considered factors are the import traffic volume of intersection; the location of intersection, the type of intersection, the grade of intersection, and traffic control mode. The network was trained using 197 intersection data, and the accuracy reached up to 89%. All of these studies established the effectiveness of machine learning and particularly ANNs for predicting accidents severity. However, according to [15], there are issues to be addressed, such as improving the prediction accuracy. In this paper.....

II. Accidents Dataset

The United Kingdom's (UK) most comprehensive traffic accidents data, which was accumulate by UK government and is publicly available on Kaggle [16] , inspired this research project. The dataset contains 1.6 million records ranging from the years 2000 up to 2016 of traffic flow data. Traffic accidents are only available for the years 2012-2014. Therefore, for this research project, only these years are considered. There are several features available in the dataset. The ones that are relevant to this research are longitude, latitude, severity, month, hour, road type, speed limit, light conditions, weather conditions, road surface conditions, area (urban or rural), and year. It is important to note that each data type for these features is different, and this is something to consider during the data analysis stage.

III. Methodology

A. Data Analysis and Pre-processing

In order to ensure that the dataset is meaningful for extracting relevant information, certain pre-pocessing steps need to be performed. The original dataset contained missing values, as well as ambiguous cases recorded as "others". Since the dataset is considerably large, all of these cases, which amount to a few hundreds only, have been omitted. Each data type has been handled accordingly. Hour, month, year, speed limit, and number of casualties are integers. Longitude and latitude are of type double, but they are involved in the spatial analysis only and not in severity classification. Similarly, month and year are also not used for machine learning severity classification; they are used for temporal analysis only. Furthermore, two important distinctions need to be made; categorical and ordinal data. Severity is an ordinal feature, and the values have natural order. For instance, severity 1, 2, and 3 mean fatal, serious, and slight, respectively. On the other hand, categorical features, such as weather conditions, road surface, light conditions, and road type, have no natural ordered relationship between their values. Thus, this type was handled by using one-hot encoding (add table to demonstrate?). To prevent colinearity, the first column of every nominal category is removed. Area type (urban/rural) can be handled either as ordinal or categorical because it has two types only. In order to prevent further colinearity, the heat map is Fig. 2 is observed. Multicolinearty exists between weather and road surface, as well as between area and speed limit. Furthermore, negative colinearity can be observed in the heat map between daylight and darkness with street lights unlit, as well as between wet/damp road surface and fine weather conditions without high winds. This negative colinearity entails that when one of these two features increases, the other decreases, which is a logical occurrence that meets the expectations. One of the important steps performed in this research work is classifying the data according to severity. Therefore, it is worth mentioning that data imbalance exists between severity classes. According to Fig. 1, 84% of the records are of severity of type 3. While types 2 and 1 takes up 14% and 1%, respectively. This data imbalance can affect the performance of machine learning algorithms. This can be handled by producing more data of the minority types, such that the percentages are closer to each other. SMOTE method is used to increase the amount of data for both types 1 and 2. However, this method reduced the accuracy of the classification down to the half. Therefore, as an alternative solution, the number of samples for severity 2 and severity 3 were cut down in order to match the number of samples for severity 1. The total amount of data remaining is 15531,which is enough to meaningful train a machine learning algorithm.

B. Machine Learning Framework

There are two machine learning techniques utilized in order to discover the most prominent features affecting the accidents' severity. The first one is Extreme Gradient Boosting (XGBoost) library, which has recently dominated machine learning research problems as an effective classification method. It is an implementation of gradient boosted decision trees, firstly created by Tianqi Chen and Carlos Guestrin [17], followed by several other contributions from many developers. XGBoost supports three forms of gradient boosting, which are Gradient Boosting, Stochastic Gradient Boosting, and Regularized Gradient Boosting. In this paper, Gradient Boosting is utilized. XGBoost is sparse-aware, which means it automatically handles missing values. Additionally, it supports

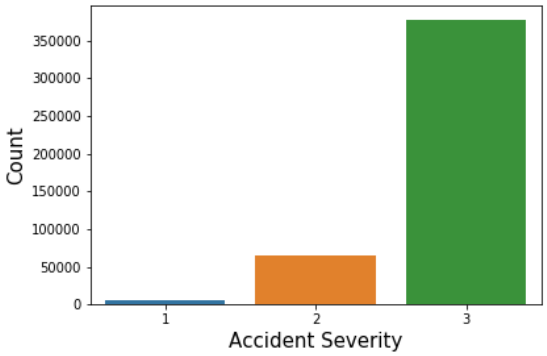

Fig. 1. Data imbalance between severity types.

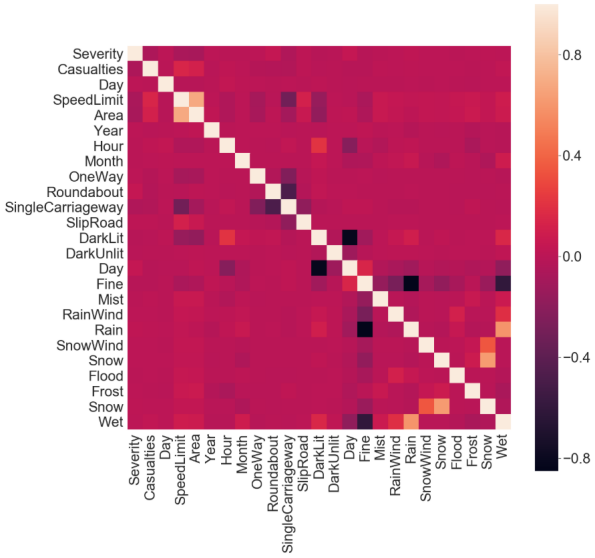

Fig. 2. Heatmap that shows correlations between the different features that will be used to classify accident severity.

parallel tree construction. Boosting is an ensemble learning method. That is, new learning models are added that predict residual errors of prior models until no improvements can be made. Eventually, all the models are cascaded together. XGBoost has proven computational efficiency among all other gradient boosting implementations. The second method is Extra Trees Classifier. This method was initially proposed by Geurts et al [18]. It is a tree-based ensemble supervised learning method that is suitable for classification and regression problems. It is also known as Extremely Randomized Trees because the features and node splits in the tree are chosen at random. This randomness reduces the risk of overfitting and makes the model less computationally extensive compared to Random Forest. Furthermore, Extra Trees Classifier does not use bootstrapping, which means it samples observations without replacement. It has produced state-of-the-art results in many complex problems. For full explanation and analysis about this methodology, the reader is referred to the original paper [18]. Extra Trees Classifier was used to determine the most prominent features affecting accident severity for the years 2012, 2013, and 2014 separately. According to Figure 3, "Hour" was the feature with the highest weight for all years. Similarly, XGBoost was used to extract feature weights for each year separately. As seen in Figure 3, the most important feature for the years 2013 and 2014 is Area, whereas 2012 shows that the most important one is Speed Limit. This kind of confusion is expected to happen, since it has been already determined that high correlation exists between these two features.

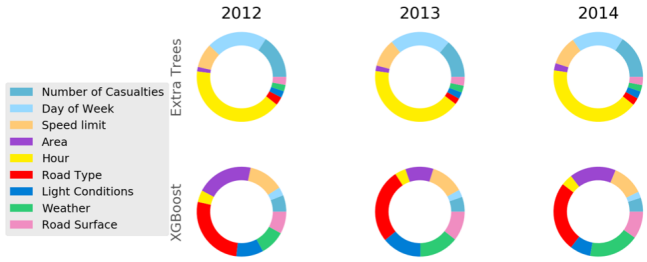

Fig. 3. Weights distribution according to Extra Trees Classifier in the upper column, and XGBoost in the lower column, for the years 2012 - 2014.

C. Spatial Autocorrelations: Moran's I Method

In GIS, spatial autocorrelation is used to measure the degree of similarity between object and other surrounding objects through space. These statistics are used for the assessment of spatial data clustering either locally by using individual features, or globally by using the entire study area. There are two common types of spatial auto correlation analyses, one of them is local spatial autocorrelation and the other is the global spatial autocorrelation [19], [20]. Moran's I method, is one of the earliest statistical methods used to measure the spatial autocorrelation, that is used for determining the overall autocorrelation of the whole data set using both feature locations and feature attributes simultaneously. For a defined set of spatial features and their related values, it assesses whether the spatial pattern is clustered, dispersed, or random. Moran's I can be expressed as shown in Eq .1 [21]

$$I = \frac{n}{s_0} \frac{\sum_{i}^{n} \sum_{j}^{n} w_{i,j}(x_i - \overline{x})(x_j - \overline{x})}{\sum_{i}^{n}(x_i - \overline{x})^2}, \tag{1}$$

where n is the total number of features, x_{ij} is the attribute value for variable x at location i and j, respectively. $\overline{x}$ is the mean value of x, w_{ij} weight matrix element between i and j, n is the total number of spatial features, and s_0 is the aggregation of all the elements in the weight matrix as shown in Eq.2

$$s_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \tag{2}$$

According to Eq. 1, Moran's I values may vary between -1 and 1, where positive values of index represent the spatial

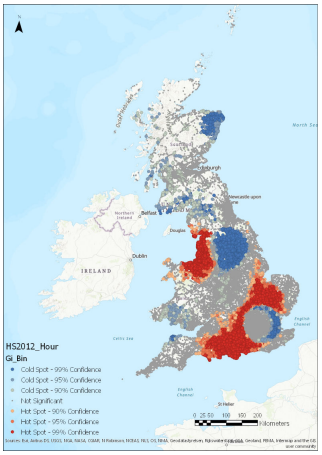

Fig. 4. Hotspots maps 2012.

clustering of similar values, negative values show the spatial dispersion, while the zero or close to zero values indicate random distribution pattern.

D. Hotspot Analysis

Hotspot analysis is considered as an effective tool for the road and transportation authority to improve accident reduction and prevention strategies [22]–[25]. There are various spatial analysis techniques in GIS to determine the traffic accident hotspots. These techniques use statistical analysis to identify the locations of significant hotspots and coldspots in the data set. Three main processes are considered for the estimation of hotspots analysis, which are collect event function, Getis-Ord GI* statistics for clusters mapping, and density estimation using kernel density function(KDE). These analyses were performed using the spatial statistics tools in ArcGIS software. Getis-Ord GI* statistic is represented as Eq.3 [26]

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j} x_j}{\sum_{j=1}^{n} x_j} \tag{3}$$

IV. Results

Spatio-temporal analyses of the traffic accidents were performed in order to predict the severity degree of the accidents during 2012, 2013, and 2014, respectively. Two hotspot maps were produced for each year for both XGBoost and ETC methods, these maps were generated by taking into consideration the features with high weights, and this process was repeated for each algorithm and year as shown in fig. 4. For Extreme Tree Classifier, The attribute with higher weight is the hour attribute for three different years, and the road type feature was considered for 2012 and 2013 using XGboost, where as for 2014 the area feature corresponding to the higher weight than other features for the same algorithm.

The predicted severity of the traffic accidents as shown in Fig. for each year and for each method. For the accuracy assessment the confusion matrix for XGBoost and ETC methods had been calculated by comparing predicted severity with ground truth severity, XGBoost show better results than ETC as shown in Table I

TABLE I
ACCURACY ASSESSMENT FOR XGBOOST AND ETC DURING 2012-2014

Year	XGBoost	ETC
2012	99.20%	95%
2013	100%	95.40%
2014	100%	97.40%

References

[1] M. I. Sameen and B. Pradhan, "Assessment of the effects of expressway geometric design features on the frequency of accident crash rates using high-resolution laser scanning data and gis," *Geomatics, Natural Hazards and Risk*, vol. 8, no. 2, pp. 733–747, 2017.

[2] M. I. Sameen and B. Pradhan, "Severity prediction of traffic accidents with recurrent neural networks," *Applied Sciences*, vol. 7, no. 6, 2017.

[3] Y. Lv, Y. Duan, W. Kang, Z. Li, and F. Wang, "Traffic flow prediction with big data: A deep learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, pp. 865–873, April 2015.

[4] S. H.-A. Hashmienejad and S. M. H. Hasheminejad, "Traffic accident severity prediction using a novel multi-objective genetic algorithm," *International Journal of Crashworthiness*, vol. 22, no. 4, pp. 425–440, 2017.

[5] Y. Duan, Y. Lv, Y.-L. Liu, and F.-Y. Wang, "An efficient realization of deep learning for traffic data imputation," *Transportation Research Part C: Emerging Technologies*, vol. 72, pp. 168 – 181, 2016.

[6] S. Alkheder, M. Taamneh, and S. Taamneh, "Severity prediction of traffic accident using an artificial neural network," *Journal of Forecasting*, vol. 36, pp. 100–108, apr 2016.

[7] R. Yu, Y. Li, C. Shahabi, U. Demiryurek, and Y. Liu, *Deep Learning: A Generic Approach for Extreme Condition Traffic Forecasting*, pp. 777–785. 06 2017.

[8] Q. Zeng and H. Huang, "A stable and optimized neural network model for crash injury severity prediction," *Accident Analysis & Prevention*, vol. 73, pp. 351 – 358, 2014.

[9] W. Huang, G. Song, H. Hong, and K. Xie, "Deep architecture for traffic flow prediction: Deep belief networks with multitask learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, pp. 2191–2201, Oct 2014.

[10] M. Fogue, P. Garrido, F. J. Martinez, J. Cano, C. T. Calafate, and P. Manzoni, "A system for automatic notification and severity estimation of automotive accidents," *IEEE Transactions on Mobile Computing*, vol. 13, pp. 948–963, May 2014.

[11] A. Najjar, S. Kaneko, and Y. Miyanaga, "Combining satellite imagery and open data to map road safety," in *AAAI*, 2017.

[12] H. Ren, Y. Song, J. Wang, Y. Hu, and J. Lei, "A deep learning approach to the citywide traffic accident risk prediction," in *21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 3346–3351, 2018.

[13] C. Dong, C. Shao, J. Li, and Z. Xiong, "An improved deep learning model for traffic crash prediction," *Journal of Advanced Transportation*, vol. 2018, 2018.

[14] L. Yuejing, Z. Xing-lin, Z. Haixia, and L. Ming, "Research on accident prediction of intersection and identification method of prominent accident form based on back propagation neural network," in *International Conference on Computer Application and System Modeling (ICCASM)*, pp. 434–438, 2010.

[15] M. I. Sameen, B. Pradhan, H. Z. M. Shafri, and H. B. Hamid, "Applications of deep learning in severity prediction of traffic accidents," in *GCEC 2017* (B. Pradhan, ed.), (Singapore), pp. 793–808, Springer Singapore, 2019.

[16] D. Fisher-Hickey, "1.6 million UK traffic accidents." Kaggle, 2017. (Accessed: 15 September 2015).

[17] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, (New York, NY, USA), pp. 785–794, ACM, 2016.

[18] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, pp. 3–42, Apr 2006.

[19] Abdulhafedh, "A novel hybrid method for measuring the spatial auto-correlation of vehicular crashes: Combining morans index and getis-ord gi* statistic," *Scientific Research Publishing*, vol. 7, no. 2, pp. 208–221, 2017.

[20] R. C. V. Prasannakumar, H. Vijith and . Geetha, "Spatio-temporal clustering of road accidents: Gis based analysis and assessment," *Procedia Social and Behavioral Sciences*, vol. 21, p. 317325, 2011.

[21] Sipos, "Spatial statistical analysis of the traffic accidents," *Periodica Polytechnica Transportation Engineering*, vol. 45, no. 2, pp. 101–105, 2017.

[22] Y. Li and C. Liang, "The analysis of spatial pattern and hotspots of aviation accident and ranking the potential risk airports based on gis platform," *Journal of Advanced Transportation*, vol. 2018, p. 12, 2018.

[23] G. A. Shafabakhsh, A. Famili, and M. S. Bahadori, "Gis-based spatial analysis of urban traffic accidents: Case study in mashhad, iran," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 4, no. 3, pp. 290 – 299, 2017.

[24] 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. International Conference: Spatial Thinking and Geographic Information Sciences 2011.

[25] 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.

[26] Z. Z. Z. Cheng and J. Lu, "Traffic crash evolution characteristic analysis and spatiotemporal hotspot identification of urban road intersections," *Sustainability*, vol. 11, no. 1(160), pp. 1–17, 2019.