Literature Review

Implementing Machine Learning to Predict Traffic Accidents

Total Word Count (excluding references section): 1937

1. Introduction

An increasing number of deaths highlights the severity of traffic-related deaths each year with estimations that 1.35 million people died in 2016 (WHO, 2018). Most traffic accidents involve vulnerable road users such as pedestrians, cyclists, or motorcyclists and research shows that stringent traffic policies are helping to reduce traffic accidents.

Machine Learning (ML) accident prediction has been researched for several years. The aim is to support traffic management to predict accident severity, duration, or probability of crashes on rural roads, city intersections, or highways. ML is a perfect tool for this research because it can help make sense of both classification and statistical problems. The main issue facing ML traffic accident prediction is correlating relations between crash predictors. Identification of the strength of a correlation can help drive regulation and traffic infrastructure, thus reducing traffic accidents. For this reason, transportation safety requires the identification of relationships between several variables used in crash occurrences (Das et al., 2021).

If made available to researchers, crash data rarely includes environmental factors that enable the identification of cause-and-effect relationships corresponding to accident rate probabilities. And, because traffic accidents are complex phenomena involving numerous variables, ML traditionally uses simple variables such as traffic flow, speed, and weather. Most times, researchers compensate the quality and availability of these data with surrogate values, which make it difficult to recommend one ML algorithm over another for all prediction models.

2. Literature Review Outline

2.1 Purpose

Throughout the literature, researchers often reference several common ML methods, including Neural Networks (NN), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Random Forest (RF), K-Nearest Neighbour (KNN), Decision Trees (DT) and Bayesian Networks. This review aims to understand which ML method researchers use often, the variables considered in research data, and the predictive strengths of the ML method.

2.2 Audience

The intended audience is any reader interested in understanding the current state of research for ML prediction models and classifiers.

2.3 Limitations

Preference is for peer-reviewed academic journals and a few industry conferences. The review lists any accuracy results of ML classifiers if the authors provide them. This review considers the main classifiers listed but does not consider ML history nor describe ML classifiers or clustering algorithms. The relationship between crash variables and outcomes is also not explored in this review.

2.4 Scope and structure

The review consulted academic journals, books, and industry conferences. Elsevier, Science Direct, Scopus, Google Scholar, Taylor and Francis Online, and Web of Science supplied search results. Search terms included “artificial intelligence”, “machine learning”, “traffic accident”, “accident prediction”, and “performance”. This review is structured mostly chronologically and concludes with a summary of salient points taken from the central literature and presents gaps or ideas for further research on this topic.

3. Literature Review

RQ1: **Which ML method is most relied on in literature for accident prediction?**

RQ2: **What variables receive the most attention in ML prediction models?**

Literature by Lv et al. (2009) presents the first time a KNN method was applied to real-time traffic data. They used six broad variables for research: traffic flow measures representing changes between normal and hazardous traffic conditions, traffic volume, speed, vehicle occupancy, road condition, and time. The time variable was added because several other crash variables depended on it. The results showed KNN has an 80% prediction rate, with performance increasing in KNN models as the number of variables increases. Sawalha and Sayed (2006) state that increasing the number of variables in accident prediction models leads to over-fitting issues, which is not acknowledged in this literature.

Jadaan et al. (2014) explored ANN simulations to determine their suitability in predicting traffic accidents in Jordan, Egypt. They used five data variables: number of vehicles, population, length of roads and gross domestic product. Their results showed that the ANN models have a high coefficient of 0.992, closely reflecting the actual traffic accident data. The authors, unfortunately, do not offer an in-depth analysis of the ANN data, for example, the importance ranking of variables like “Number of Registered Vehicles” or “The gross domestic product”. The lack of analysis is expected because ANN is often considered a black box that does not report the correlation between explanatory variables and outcome.

Later, Najafi et al. (2021) also used pattern recognition ANN to provide a traffic accident prediction model. The downside to their research is the limited dataset available to the research, which was for twelve months—the average period in this review is eleven years. Their inputs included time of day, human characteristics, environment, and accident types. Because of the limited dataset, they analysed 632 incidents, of which 106 were injury or fatal accidents. Another limitation of this research is the small number of training epochs used (92). Despite the limitations, their results showed ANN produced high coefficients of 0.989, in line with the findings of Jadaan et al. (2014).

In research by Kwon et al. (2015), DT (often used in data mining) outperforms the Naïve Bayes algorithm if dependencies between variables are paramount. They reached this conclusion by looking at over 1.3 million collision reports over six years using 26 crash variables. Critical variables included human behaviour, drug use, gender, driver age, vehicle age, road surface, time, area population, weather and type of collision. They initially grouped these variables into five categories, resulting in skewed results because of incorrect classification. The authors resolved the bias (which affected the DT models) by differentiating the data sets and merging standard variables. A change in the clustering approach led to the improved predictive performance of DT.

Iranitalab and Khattak (2017) showed that the Nearest Neighbour Classification (NNC) outperforms RF (73.95% and 45.75%, respectively). SVM outperforms Multinomial Logit (MNL) (30.29% and 20.09%, respectively). This conclusion is based on an analysis of 49000 crashes in Nebraska over three years, comparing the performance of NNC to SVM, RF, and MNL for crash severity prediction. They used over ten variables: road surface, number of lanes, alcohol use, age, gender, area population, lighting conditions, weather, and the presence of debris or obstructions. Like Kwon et al. (2015), they encountered bias in the data and split it into homogenous clusters, allowing NNC to score the best overall, followed by RF and SVM. The researchers did not use ANN because a significant portion of the binary sample set contained “0” values, affecting ANN prediction and performance. Also, they did not utilise the ordered probit model (considered the most used statistical model). If the authors included the cost of crashes in their ML models—which they did not—it would have affected their conclusion about NNC’s overall performance.

In contrast, Hamad et al. (2020) later show that SVM is a better prediction algorithm than NCC. They used a dataset of 140000 incidents from Houston, Texas, spanning nine years, with over 50 variables to measure the effective classification performance of ANN, SVM, Gaussian Process Regression (GPR), Regression Decision Tree and Ensemble Tree. Some variables included GPS location, road name, type of vehicles involved, weather condition and date and time of an incident. The results demonstrated that Regression Trees had the lowest training time, followed by (on average) Ensemble Trees, ANN, SVM, and GPR. But, in terms of accuracy, SVM outperformed all other algorithms, followed by ANN, Ensemble Trees, GPR and Regression Trees.

When considering RF accuracy, research by Zhang et al. (2018) concluded that RF produced the best prediction accuracy (53.9%) in overall and severe crashes, compared to DT (50.7%), SVM (52.6%), ordered probit (44%), KNN (52.9%) and MNL (50.9%). This is

strange because the accuracy values do not align with those found in Iranitalab and Khattak (2017) for each respective ML method. The authors state (regarding RF) that the “predictive performance of RF is not very robust and varies when applied on different data”. Data used in this research related to 5538 crashes over three years on freeway diverge areas in Florida and used ten variables: ramp type, number of lanes, deceleration length, road surface, speed, light, weather and area surrounds, alcohol use and crash type. The authors conclude that higher prediction accuracy does not imply a more accurate variable importance estimation for any ML method.

Deep Learning (DL) fares reasonably well for this topic, as Theofilatos et al. (2019) researched. They compared the performance of ML and DL models for real-time crash prediction using data for Attica Tollway in Greece. Five years’ worth of real-time traffic and weather data was linked, covering 284 crashes. They compared the DL model to KNN, Naïve Bayes, DT, RF, and SVM. Included in the data were traffic flow, vehicle occupancy, speed, and length, rainfall, temperature, humidity, radiation, wind speed and direction. Results showed that using DL for prediction delivers well-balanced metrics. It surprised the researchers that Naïve Bayes performed very well, despite being “far less complex than the other models”.

Erzurum Cicek and Kamisli Ozturk (2021) introduced one of the first studies to consider accident prediction using One-Class Classification (OCC). OCC is unique because it distinguishes if new data is an outlier, works well in unbalanced or skewed data and can work only with a single data set. They used RF to identify significant variables that contribute to crashes. Data was sourced for 8380 accidents over nine years to compare OCC’s performance against SVM, RF, and KNN. Over 35 variables were applied to the data, including injury severity, number of accidents, accident risk, frequency and occurrence, number of fatalities, weather, road condition, and human factors. In the results, OCC got an F-score of 0.6, well above SVM (0.525), RF (0.552) and KNN (0.486), showing OCC to be an excellent contender for accident prediction, especially if datasets contain only crash data.

Fiorentini et al. (2022) are the first to introduce regression-based models to a road network under study in Italy. They used quantitative means to evaluate four ML accident prediction models for Italian roads. ML methods used were Classification and Regression Tree (CART), Boosted Regression Trees, RF and SVM. They used data spanning eight years, covered 5802 crashes, and used the following variables: traffic flow, intersection, and driveway density, and area location. Data over-fitting was identified as an issue during the training

phase for the RF model. The SVM model was the most stable during training and testing and, as a result, outperformed the other ML methods used.

4. Conclusion

Traffic accident prediction using ML is a classification or a regression problem involving various variables needed to predict a crash's occurrence accurately. Variable selection is a significant point mentioned in modern literature because it is no trivial task to recognise crash variables' impact and magnitude (Fiorentini et al., 2022). Zhang et al. (2018) state that an ML method's estimations of importance on exploratory variables vary, while Hala et al. (2018) state that different ML methods place importance on different factors.

From this literature review, an answer to RQ1 is now provided (ordered by accuracy): SVM, NNC, RF, DT, KNN, and Naïve Bayes are most relied on. Contrasting the results with research by Fiorentini et al. (2022) reveals that SVM, RF, ANN, KNN, and Naïve Bayes make up over 50% of broader research (Figure 2). For RQ2, this literature review identified the top five variables in ML accident prediction models (in order of impact): weather (7%), speed (4%), road surface (4%), traffic flow (4%) and human factors (4%). Judging these results with broader research (**Error! Reference source not found.**) from Silva et al. (2020) shows speed (4.55%), age (4.55%), traffic volume (4.55%), weather (3.64%) and heavy vehicles (3.64%).

A gap in the presented literature is

Further research is to

References

Das, S., Kong, X. & Tsapakis, I. (2021). Hit and run crash analysis using association rules mining. *Journal of Transportation Safety & Security*, 13 (2):123-142.

Erzurum Cicek, Z.I. & Kamisli Ozturk, Z. (2021). Prediction of fatal traffic accidents using one-class SVMs: a case study in Eskisehir, Turkey. *International Journal of Crashworthiness*, 1-11. DOI: https://doi.org/10.1080/13588265.2021.1959168

Fiorentini, N., Leandri, P & Losa, M. (2022) Defining machine learning algorithms as accident prediction models for Italian two-lane rural, suburban, and urban roads. *International journal of injury control and safety promotion*, pp.1-13. DOI: https://doi.org/10.1080/17457300.2022.2075397

Hala, H., Anass, C., Rajaa, B., Youssef, B. & Garza-Reyes, J. (2021). Machine learning techniques for forecasting the traffic accident severity. *In 2021 International Conference on Digital Age & Technological Advances for Sustainable Development (ICDATA)*. 47-52. Marrakech, Morocco. 29-30 June 2021. IEEE. DOI: https://doi.org/10.1109/ICDATA52997.2021.00018

Hamad, K., Khalil, M.A. & Alozi, A.R. (2020). Predicting freeway incident duration using machine learning. *International Journal of Intelligent Transportation Systems Research*, 18 (2):367-380. DOI: https://doi.org/10.1007/s13177-019-00205-1

Iranitalab, A. & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention*, 108:27-36. DOI: https://doi.org/10.1016/j.aap.2017.08.008

Jadaan, K.S., Al-Fayyad, M. & Gammoh, H.F. (2014). Prediction of road traffic accidents in Jordan using artificial neural network (ANN). *Journal of Traffic and Logistics Engineering*, 2 (2). DOI: https://doi.org/10.12720/jtle.2.2.92-94

Kwon, O.H., Rhee, W. & Yoon, Y. (2015). Application of classification algorithms for analysis of road safety risk factor dependencies. *Accident Analysis & Prevention*, 75:1-15. DOI: https://doi.org/10.1016/j.aap.2014.11.005

Lv, Y., Tang, S. & Zhao, H. (2009). 'Real-time highway traffic accident prediction based on the k-nearest neighbor method'. *2009 international conference on measuring technology and mechatronics automation Vol. 3:547-550.* Zhangjiajie, China. 12-12 April 2009. IEEE.

Martensen, H., Diependaele, K., Daniels, S., Van den Berghe, W., Papadimitriou, E., Yannis, G., Van Schagen, I., Weijermars, W., Wijnen, W., Filtness, A. & Talbot, R. (2019). The European road safety decision support system on risks and measures. *Accident Analysis & Prevention*, 125:344-351. DOI: https://doi.org/10.1016/j.aap.2018.08.005

Najafi Moghaddam Gilani, V., Hosseinian, S.M., Ghasedi, M. & Nikookar, M. (2021). Data-driven urban traffic accident analysis and prediction using logit and machine learning-based pattern recognition models. *Mathematical problems in engineering, 2021*. DOI: https://doi.org/10.1155/2021/9974219

Sawalha, Z. & Sayed, T. (2006). Traffic accident modeling: some statistical issues. *Canadian journal of civil engineering*, 33 (9):1115-1124. DOI: https://doi.org/10.1139/l06-056

Silva, P.B., Andrade, M. & Ferreira, S. (2020). Machine learning applied to road safety modeling: A systematic literature review. *Journal of traffic and transportation engineering (English edition)*, 7 (6):775-790. DOI: https://doi.org/10.1016/j.jtte.2020.07.004

Theofilatos, A., Chen, C. & Antoniou, C. (2019). Comparing machine learning and deep learning methods for real-time crash prediction. *Transportation research record*, 2673 (8):169-178. DOI: https://doi.org/10.1177%2F0361198119841571

WHO, (2018). Global Status Report on Road Safety 2018. Geneva, Switzerland.

Zhang, J., Li, Z., Pu, Z. and Xu, C. (2018). Comparing prediction performance for crash injury severity among various machine learning and statistical methods. *IEEE Access*, 6:60079-60087. DOI: https://doi.org/10.1109/ACCESS.2018.2874979

Appendix

2.5 ML method variables

Figure 1 Summarised from Table 3. Main contributing factors for crash occurrence or severity. (Silva et al., 2022)

posted speed limit	4.55%
age	4.55%
traffic volume	4.55%
weather conditions	3.64%
percentage of heavy vehicles	3.64%
segment length	2.73%
traffic volume	2.73%
seat belt use	2.73%
road surface type	1.82%
median barrier	1.82%
horizontal curvature	1.82%
number of lanes	1.82%
number of injuries	1.82%
pavement surface	1.82%
roadway width	1.82%
bus stop	1.82%
sex	1.82%
cause of crash	1.82%
type of vehicle	1.82%
sex	1.82%
annual precipitation	1.82%
point of impact	1.82%
type of collision	1.82%
volume of traffic	0.91%
time of the crash	0.91%
maximum damage to vehicle	0.91%
movement prior to collision	0.91%
type of traffic violation	0.91%
collision of type “sideswipe”	0.91%
lighting	0.91%
number of involved vehicles	0.91%
skid resistance	0.91%
collision partner	0.91%
type of crash	0.91%
collision type	0.91%
use of alcohol or drugs	0.91%
contributing circumstances	0.91%
lightning	0.91%
day of the week	0.91%
location type	0.91%
age of vehicle	0.91%
segment length	0.91%
demographic characteristics of driver	0.91%
time	0.91%
existence of intersections	0.91%
land use	0.91%
vehicles involved	0.91%
type of highway	0.91%
alcohol or drugs use	0.91%
lane width	0.91%
existence of the right and left shoulders	0.91%
vehicle/driver action	0.91%
gender	0.91%
weather	0.91%
rollover	0.91%
light condition	0.91%

safety devices	0.91%
presence of parking area	0.91%
cause of collision	0.91%
segment in military area	0.91%
horizontal signaling	0.91%
average annual daily traffic (AADT)	0.91%
shoulder width	0.91%
precipitation	0.91%
type of crash	0.91%
road separation	0.91%
vertical curvature	0.91%
road shoulder	0.91%
road surface condition	0.91%

2.6 ML methods referenced in literature

Figure 2 Summarised from Table1. Road Safety-related Analyses by MLAs (Fiorentini et al., 2022)

SVM	15.5340%
RF	9.71%
CART	9.7087%
ANN	9.7087%
LR	7.7670%
KNN	6.7961%
NB	2.91%
NBR	2.91%
OPM	2.91%
MNL	2.91%
RENB	2.91%
DNN	1.9417%
MARS	1.9417%
BRT	0.9709%
LSTM-RNN	0.9709%
RMNL	0.9709%
XGBoost	0.9709%
OLM	0.97%
BBN	0.97%
GEP	0.97%
RBF-ANN	0.97%
LSTM-CNN	0.9709%
CNN	0.9709%
NLM	0.97%
SVM-KNN ensemble	0.97%
M5T	0.9709%
PART	0.97%
ERT	0.97%
DJ	0.9709%
BQR	0.97%
SVM ensemble	0.97%
SVM-MKL ensemble	0.97%
SVM-MKL	0.97%
CLM	0.97%
EGB	0.97%
BRNN	0.97%