OPEN ACCESS JOURNALS
Indiana Journal of Multidisciplinary Research
Abbreviated Tittle- Ind J Multi Res
ISSN *(Online)*- 2583-3820
Journal Homepage Link- https://indianapublications.com/Journals/IJMR
DOI: https://doi.org/10.5281/zenodo.21610673

INDIANA PUBLICATIONS
PRODUCTIVE AND QUALITY RESEARCH

Review Articles

Volume-06|Issue-04|2026

Machine Learning Framework for Road Accident Prediction

Spoorthi M[1]

[1]Spoorthi M, Assistant Professor, Information Science and Engineering, Dayananda Sagar College of Engineering, Bangalore 560078.

Article History
Received: 05.05.2026
Accepted: 22.06.2026
Published: 25.07.2026

Citation
Spoorthi, M. (2026) Machine Learning Framework for Road Accident Prediction. *Indiana Journal of Multidisciplinary Research, 6*(4), 452-456.

Abstract: Road accidents are a major public safety issue because they kill 17 people every hour in India. We need to find good ways to stop accidents right away because each death is linked to many injuries of varying severity. Accurate forecasting is essential for transportation safety management due to the concerning increase in daily incidents attributed to the rapid rise in vehicle numbers. This study develops an accident prediction model utilising data mining techniques, particularly the K-Nearest Neighbours (KNN) Classifier and Support Vector Machines (SVM), and ensembles both results to increase the accuracy. The study takes into account the environment, the condition of the roads, and the amount of traffic. By spotting trends and predicting how often accidents will happen, this method may help government agencies, transportation departments, and NGOs focus their safety measures. Contractors, public works agencies, and automakers may also use the results to make safety improvements and better infrastructure design. The proposed forecasting technique illustrates how machine learning methodologies can improve road safety by reducing accident frequencies.

Keywords: Accident prediction, Machine learning, Data mining, SVM, KNN, Road accidents, Pattern of data.

INTRODUCTION

Concerns about how traffic accidents affect public health are very important because they are still one of the leading causes of death and injury around the world. In India, where more than 100,000 people were hurt in traffic accidents in 2017 and 25,859 people died, this is especially true. Not only do road accidents kill people, but they also cost a lot of money in medical bills, fixing damaged infrastructure, and lost productivity. About 1.17 million people die in traffic accidents around the world every year, and another 10 million are left unable to work.

In the past few years, there has been a disturbing rise in the number of accidents involving vehicles in India. According to the Indian National Police's Traffic Corps, the number of reported accidents went up slowly from 2014 to 2015, reaching 98,970. In 2016, the number of reported accidents reached 105,374. Before this spike, there were dips, like a high of 117,949 cases in 2012 and a low of 100,106 cases in 2013. We need to find out right away what is causing the number of accidents to go up.

Traffic accidents happen because of human mistakes, faulty cars, old road systems, and natural disasters. The most common cause is human error. About 75% of toll road accidents in India are caused by problems with the driver, according to research. Some examples are problems with driving, unpredictable reactions to environmental cues, differences in how people see danger, and not being able to drive well enough. Toll roads have the highest accident rates of any type of highway because they are more likely to have drunk drivers and a wider range of driver behaviours.

Because their accident databases aren't good enough, developing countries have a hard time coming up with good ways to stop accidents. Some places in the world don't keep full records or don't collect data at all. This gap makes it hard to figure out how well safety measures work and to correctly find the factors that are most dangerous. If we really want to make roads safer, especially those with tolls, we need to come up with a way to predict traffic road accidents (TRAs).

The goal of a TRA prediction model is to give objective estimates of how often accidents will happen without falling into regression-to-the-mean bias and other statistical traps. Taxi companies, lawmakers, and other government agencies may use this model to see how safety measures on the road affect the number of accidents. This allows you to prioritize highway maintenance projects, spend money more efficiently, and design safety programs that target specific areas one by one. India and other developing countries could learn a lot from these models about how to make toll roads safe.

Predicting accidents has long been difficult to do using traditional statistical methods due to the various problems they present. Models of this kind fail to depict real-world accidents as they are not able to capture nonlinear interactions of accident factors. Scientists are increasingly employing advanced data mining and

***Corresponding Author:** Spoorthi M

machine learning techniques to overcome these obstacles.

An integrated TRA prediction model was developed using ANNs and SVMs. Models can use **R-Miner** to mine data from toll road operators to detect patterns in accident frequency. Machine learning algorithms can find patterns in high-dimensional data that traditional methods miss, which can help make better predictions that can be used on a larger scale.

The main goal of this study is to learn more about the reasons behind toll road accidents in India so that we can make policies and procedures that are safer. This study establishes a foundation for the integration of predictive analytics into traffic management to enhance road infrastructure design in developing nations, reduce accident rates, and minimise casualties.

1.1 Contribution

- Proposed a novel methodology using machine learning to predict accidents.
- Tuned a machine learning model to increase the model performance.

LITERATURE SURVEY

The academic community has recently focused on predicting traffic accidents, employing diverse statistical, machine learning, and deep learning methodologies. Park, Kim, and Ha (2016) were the first to use big data analytics to predict highway accidents using data from Vehicle Detection Systems (VDS). To address the challenges posed by unforeseen accidents in high-velocity traffic, their research illustrated the utilisation of data gathered from extensive sensors to predict the likelihood of road accidents. Support Vector Machines (SVMs) were shown to be useful for forecasting traffic deaths by Li, Yang, Wang, Chen, and Yao (2016), around the same time. In their study, they emphasised SVM's ability to depict nonlinear relationships among variables such as traffic, road conditions, and driver psychology.

From **2018–2022**, many studies were conducted on forest-based ensemble models, namely Random Forests. Taamneh and Taamneh used accident data from Abu Dhabi in 2019 to test how well Random Forests predicted the severity of car accidents. Their study showed that it could classify crashes better than other models by using factors like time, cause, driver demographics, and road conditions.

Yan and Shen (2022) developed a model (**BO-RF**) in the form of a Bayesian-optimized Random Forest model. The model makes predictions related to the severity of accidents that take place on urban streets. The researchers, using partial dependency graphs, demonstrated which variables made accidents worse (e.g., light, vehicle speed) and were accurate in over 80% of their predictions.

As people began using open data platforms, researchers began testing algorithms on shared data. Parveen, Shaik, Raheem, and Raju (2023) applied different models for classifying the severity of road accidents, such as Random Forest, XGBoost, Linear Regression, and Support Vector Machines. They acknowledged that class imbalance—a surplus of major accidents than fatal ones—posed a critical obstacle that required careful preprocessing. Researchers Prajapati et al. (2023) experimented with the use of Decision Trees, Random Forest, and SVM on Kaggle datasets and found them appropriate for accident hotspot detection. They did, however, mention that datasets curated like this one can't capture the complexity of actual road traffic situations in their full scope.

Recent studies also focus on graph-based methodologies and deep learning techniques. Han, Yang, and Chen (2021) used features like collision speed, accident site, and road information in a Random Forest model to enhance it and obtain more accurate predictions on Chinese datasets. Udupi and Jose (2024) presented a deep learning categorisation model that surpassed traditional machine learning methods in capturing more complex accident patterns. Khanum, Kulkarni, Garg, and Faheem (2024) used machine learning techniques for predictive severity analysis to meet urgent requirements for localised accident modelling specific to Indian roadways. GNNs and other advanced designs have come into the picture during this time. Nippani, Li, Ju, Koutsopoulos, and Zhang employed nine million accident data records from the United States to create a GNN-based framework. This framework does a good job of modelling the spatial relationship between road networks, and it also helps to make better predictions.

Along with these changes in methods, thorough evaluations have helped to make the field's progress more solid. Behboudi, Moosavi, and Ramnath (2024) put together the results of about two hundred studies that came out in the last five years. Their analysis brought up several important issues, such as data imbalance, regional specificity, and the need for explainable AI in accident modelling. These results highlight the shift from statistical approaches to interpretable ML and advanced DL models in traffic accident prediction, with a focus on regional applications such as toll roads and highways in India.

MATERIALS AND METHODS

The proposed methodology is provided in the figure below. It has steps such as collecting the database used for training and testing the model. The obtained database is preprocessed to remove noise and missing values. Features are extracted from the dataset. Once the features are extracted, they are used to train and test the two classification models: Naïve Bayes and KNN classifiers. Lastly, performance analysis is carried out on the testing

***Corresponding Author:** Spoorthi M

data to determine the performance of the trained model. The figure shows the stated steps.

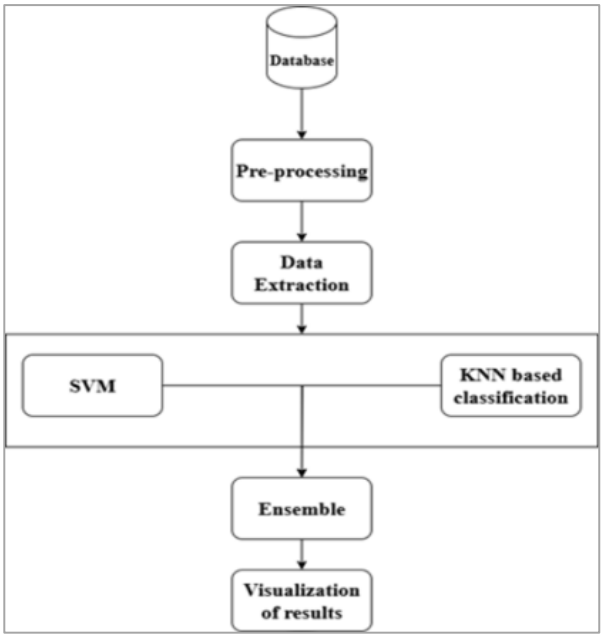

Figure 1: Proposed model

The foundation for constructing dependable predictive systems is data preparation, which is the starting point for every model development process. Raw datasets are likely to have discrepancies, null values, or ambiguous representations, tending to decrease the correctness of the model if not handled properly. Accordingly, the data should be significantly preprocessed before being fed to an algorithm for computation.

The first step involves handling missing values. Examples with a lack of evidence for the relevant signature characteristics are excluded because missing information may introduce noise and produce biased results. For datasets where missing values (such as accident and fatality) are sensitive features, per-entity imputation techniques, such as using the mean or most frequent category to fill in missing values, are not reliable, and dropping partial examples is more suitable. This allows the final dataset to contain only relevant and stable data.

Then, the dataset's numeric codes are converted to their nominal values using the user manual (data dictionary). For example, coded values may represent "Male/Female" ("1"/"2") or "No Injury/Injury/Fatal" ("0", "1", and "2"). The process of encoding not only makes linear representations understandable to humans, but it also aims to prevent algorithms from misinterpreting a code as a numeric category, which may lead to a misleading interpretation.

Once standardized, the fatality rate can be computed. Instead of just counting how many people died, the rate adjusts for size. To adjust for size, fatalities are divided by the total number of accidents or exposure. For example, fatalities per 100 accidents would be a rate. By normalizing the measure, it becomes more informative and can be compared across regions, times, or accident types.

To make things easy for the model to fit and to bring out patterns, the fatality rate has then been binned into two bins, namely **High** and **Low**. Binning refers to the act of converting continuous numerical values into discrete categories. In bins, the algorithms will focus on distinguishing between high-risk conditions and low-risk conditions. This is distinctly different from raw fatality rates, which might differ only slightly. An example of a cutoff can be determined and grouped as **"Low"** with a fatality rate < 0.05, and **"High"** otherwise.

Beyond these operations, the data is clean and invariant (consistent) for algorithmic analysis. The preprocessing helps the model learn from useful patterns and ignore missing values, ambiguous codes, or non-standard measures. In a nutshell, data preprocessing is the process through which raw accident data are formatted in an orderly manner for good prediction as well as interpretability.

Classifiers

Conceptual modeling is the activity of formally describing some aspects of the physical and social world around us for the purposes of understanding and communication. During this process, the knowledge items are determined, the associations between data objects are created, and the semantics (e.g., rules or constraints) are formulated. Through the formalization of the representation of data, knowledge modeling ensures that information is represented so that it can be efficiently stored, retrieved, or analyzed in ways that are useful for both operational and analytical purposes.

An effective data model is also necessary for enforcing business rules, regulatory compliance, and governmental standards. By modeling, the data can be standardized to the level of having consistent names, controlling default values and meanings, purposes, and security. This ensures the authenticity and trustworthiness of the content saved. Unlike process models, which focus on what transformations must be performed on the data before and after processes are executed, a data model focuses on what data is required by whom in support of specific business processes. **CXT30380**. However, unlike process models that concern themselves with what compositional structures correspond to potential navigational paths through an information space, a data model focuses instead on isolation and surface area requirements of relevant information characteristics. In this respect, the data model is analogous to a building blueprint—it specifies the structural form of relationships among data items that makes possible a unified and consistent conception of an information environment.

***Corresponding Author:** Spoorthi M

In the analysis of accidents and fatalities, knowledge modeling starts with the computation of a range of descriptive statistics calculated from the dataset. These metrics are intentionally chosen because they reflect most closely on fatal crashes, e.g., type of accident, weather, alignment, and severity. Once created, these features form the knowledge model by showing which variables influence what.

Machine learning algorithms like SVM and KNN search for sequences of interest in data encoded with such mapped representations. Taken together, these algorithms elucidate relationships among accident attributes to uncover patterns that can be employed for prediction and prevention strategies as well as policy-making.

Ensemble: To achieve higher accuracy, a soft voting-based ensemble technique is employed. This takes votes from both classifiers, and the class receiving the maximum votes is provided as the final decision.

RESULTS AND DISCUSSIONS

The system contains three core modules: Service Provider, View and Authorize Users, and Remote User. Data and Model Management is a module that allows administrators to manage the data and model management service. View and Authorize Users allow administrators to view, monitor, and authorize users. The Remote User module is for end users who use the system, providing registration and personal options. The system features include user registration, login, prediction of road accident status, profile management, and secure logout. The system further comprises a login screen for both existing and new users.

Figure 2: Login page for road accident prediction

In the next step, the user is allowed to provide details that are used to predict accidents, as shown in the figure below.

Figure 3: Details entered by the user for prediction

The Service Provider Module is an application that helps administrators create datasets, train predictive models, and monitor system performance. It provides the means to browse datasets, retrieve relevant accident data, and perform training and testing. The module also includes performance analysis and reporting tools such as simple bar charts and comprehensive accuracy reports. It is also capable of road accident prediction monitoring, where system administrators can view prediction results and remote users. In summary, the module offers extensive options for data storage and manipulation.

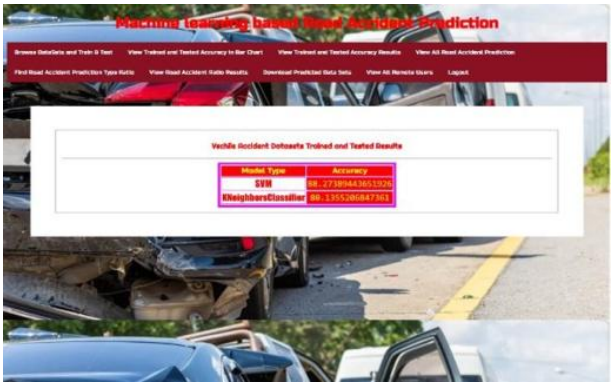

Figure 4: Accuracy achieved in predicting accidents

CONCLUSION

Road safety is a collective effort, and road authorities and car manufacturers must take responsibility for risk reduction. A machine learning model has been developed to predict accident probabilities based on factors like vehicle type, driver age, weather, and road structure using a Bangalore dataset. This model can help governments design safety measures and prevent accidents. Future refinements and scaling can consider more features and constraints. A smartphone app could guide drivers through safer routes and warn them of accident hotspots. Ride-hailing services like Uber and Ola could implement similar capabilities to improve safety. The model could also facilitate better road safety guidance, enhanced accident tracking, and quicker emergency response. The ensemble technique enhanced the performance of the model.

***Corresponding Author:** Spoorthi M

REFERENCES

1. Čubranić-Dobrodolac, M., Lipovac, K., Čičević, S., & Antić, B. (n.d.). *A model for traffic accidents prediction based on driver personality traits assessment*. Faculty of Transport and Traffic Engineering, University of Belgrade.
2. Shetty, P., P. C., S., Kashyap, S. V., & Madi, V. (n.d.). *Analysis of road accidents using data mining techniques*. Department of Computer Science and Engineering, National Institute of Engineering.
3. Ihueze, C. C., & Onwurah, U. O. (n.d.). *Road traffic accidents prediction modelling: An analysis of Anambra State, Nigeria*. Department of Industrial and Production Engineering, Nnamdi Azikiwe University.
4. Krishna, G. V. (n.d.). *An integrated approach for weather forecasting based on data mining and forecasting analysis*. Department of Computer Science and Engineering, GITAM University.
5. Hussain, S. M., & El-Mouadib, F. A. (n.d.). *Application of data mining techniques for analyzing road traffic incidents: A case study of Libya's road traffic*. University of Benghazi.
6. Kumar, S., & Toshniwal, D. (n.d.). *A data mining approach to characterize road accident locations*.
7. Greibe, P. (n.d.). *Accident prediction models for urban roads*. Danish Transport Research Institute.
8. Atnafu, B., & Kaur, G. (n.d.). *Survey on analysis and prediction of road traffic accident severity levels using data mining techniques in Maharashtra, India*. Symbiosis Institute of Technology.
9. Li, L., Shrestha, S., & Hu, G. (n.d.). *Analysis of road traffic fatal accidents using data mining techniques*. Department of Computer Science, Central Michigan University.
10. La Torre, F., Domenichini, L., Meocci, M., Graham, D., Karathodorou, N., Richter, T., Ruhl, S., Yannis, G., Dragomanovits, A., & Laiou, A. (n.d.). *Development of a transnational accident prediction model*.

***Corresponding Author:** Spoorthi M