JCBI
Journal of Computing & Biomedical Informatics
ISSN: 2710 - 1606
Volume 06 Issue 02
2024

*Review Article*
https://doi.org/10.56979/602/2024

# A Comprehensive Analysis of Machine Learning and Deep Learning Approaches for Road Accident Prediction

**Hamid Ghous[1], Mubasher H. Malik[1*], Salman Qadri[2], Amna Atiq[3], and Syed Ali Nawaz [4]**

[1]Department of Computer Science, Institute of Southern Punjab, Multan, Pakistan.
[2]Department of Computer Science, Muhammad Nawaz Sharif University of Agriculture, Multan, Pakistan.
[3]Department of Information Technology, Institute of Southern Punjab, Multan, Pakistan.
[4]Department of Information Technology, The Islamia University of Bahawalpur, Bahawalpur, Pakistan.
[*]Corresponding Author: Mubasher H. Malik. Email: mubasher@isp.edu.pk

**Received:** December 21, 2023 **Accepted:** February 11, 2024 **Published:** March 01, 2024

**Abstract:** Road accidents pose a substantial concern for all individuals. On a daily basis, a substantial number of precious lives are tragically lost due to vehicular collisions. The research conducted on road accident detection and prevention involves the utilization of various datasets to predict potential scenarios that may result in road accidents. Nevertheless, a significant obstacle that arises during the development of computer-vision models aimed at identifying traffic attributes in road accidents is the scarcity of available datasets. The dataset lacks diverse factors contributing to road accidents, which could lead to leveraging for training deep-learning models on a broad scale. This work seeks to give an in-depth examination of road accident databases, implementing machine-learning techniques, and use of Deep Learning (DL) algorithms on road accident datasets. This research examines the methods of Machine Learning (ML) and DL algorithms that are utilized in the process of creating road accident projections, as well as their relevance to the data sets that are being taken into consideration.

**Keywords:** Road Accidents; Machine Learning; Deep Learning; Automobile Incidents; Accident Detection System.

## 1. Introduction

The occurrence of road traffic accidents poses a significant risk to both human life and the safety of the surrounding living environment. Unfortunately, road accidents have increased worldwide, with highway accidents having become especially prevalent in recent years. Traffic accidents lead to several losses, mostly the loss of human life, damage to property, and financial loss.

In numerous developing nations with economic constraints, such as Libya, Pakistan, and Afghanistan, automobile collisions and road accidents have emerged as primary contributors to mortality rates, surpassing the toll inflicted by armed conflicts by a factor of 2x8. According to cited sources [1], it has been observed that on a daily basis, vehicular accidents result in the unfortunate loss of life for approximately 3,700 individuals worldwide [2]. This phenomenon has a significant effect on society and the economy of each affected country, primarily due to the considerable impact of human lives lost and injuries incurred.

Researchers and contributors have utilized a range of datasets to forecast road accidents. By identifying the various factors that contribute to the occurrence of traffic collisions, data analysis enables the eradication of significant hazards that result in accidents. The appropriate choice of a data analysis approach is crucial for identifying the factors contributing to accidents in a particular geographic region and accurately predicting future accidents' likelihood. Prior studies have highlighted different transportation and traffic conditions as factors associated with the occurrence of accidents. Instances of these identifiers include factors such as congested roadways, psychological parameters [3], road types, road conditions [4] and road segments [5]. Apart from these identifiers the condition of the automobile, rules and regulations according to the government and even the environmental and atmospheric characteristics play a vital role.

Furthermore, the available large-scale datasets suffer from limitations such as being privately maintained, outdated, or lacking crucial contextual information like environmental triggers (such as weather conditions et. c. Road accident records often include a range of inherent and contextual characteristics, confined to location, time, atmospheric conditions, time of hour etc.

Forecasting future outcomes, especially in a context where autonomous vehicles are prevalent, can provide considerable challenges. Loop-based or other static traffic analyzers have been extensively employed in detector-based methodologies for unwanted traffic detection [6] over an extended period of time.

The collection of datasets and the analysis of traffic accidents have presented numerous obstacles. The examination of road accidents datasets is primarily influenced by traditional approaches to statistical analysis centered around logistic regression, MLR, and other statistical methods as generalized ordered logistic regression. [7] This sector of road accident detection is also using more data-driven approaches like neural networks, ML, and DL.

The objective of this work is to present a comprehensive examination of the many data sources of automobile incidents. The implementation of ML techniques on these data sources, and the utilization of deep-learning methods on road accident datasets. This research presents not only a comprehensive analysis of the ML and DL techniques that were used in the construction of automobile collision prediction models but also an evaluation of how compatible those algorithms are with the particular data sets that are the focus of this research.

The datasets referenced in this research will provide valuable assistance in the development of automated accident detection systems that utilize DL and computer vision techniques. Highway management systems can employ the utilization of these datasets for the purpose of identifying traffic accidents through the use of live video surveillance systems installed strategically across the metropolitan area, including important highways and highways. The technology will provide government officials and authorities the ability to monitor traffic accidents in real-time as they occur.

## 2. Data Sources of Road Accidents

Research datasets can exhibit diverse places and encompass a wide array of information, including practical information and raw data. The datasets covering highway crashes and road accidents are typically subject to protection by authorities or made available online through the efforts of researchers and individuals who have conducted experiments and collected data from various sources. Multiple sources of road accident statistics exist, with one particularly crucial among them being the data provided by Government organizations. Governmental departments are responsible for generating, collecting, and making the data sets accessible to the public. Government agencies encompass law enforcement, highway authorities and relevant agencies. Data collection is conducted by obtaining Initial Information Records from the Police department. The other form of data collection is done by accessing databases of Insurance companies and referring to healthcare records. The primary devices employed for data collecting include automobile speed detection sensors and cameras. The radars and signals breaking cameras and the saved data they hold is also used for the cause. The previously mentioned data are employed to inform the development of public policies in various countries pertaining to road safety initiatives and transportation infrastructure improvements i.e. urban roads [8]. An additional source of information is the open-access database, often funded by the public as a whole, and available to all individuals without any limitations [9] [10]. Publicly available data refers to data that is attainable to the population as a whole through specialized websites that provide search and discovery connections, such as web-based applications

Numerous scholars have conducted studies regarding the identification of road accidents by utilizing data obtained from various public and governmental sites and resources. This paper provides a comprehensive review of the datasets utilized by numerous researchers, including their respective sources, methodologies and data attributes.

The traffic data utilized by Zhao [11] encompassed information on traffic accidents. The study also associated traffic flow data upstream and downstream within a specific temporal scope. The analysis utilized sample data consisting of traffic-related crash dataset and associated "traffic flow state data" collected from interstate highway I5 in California, USA, spanning from mile 495.493 to mile 539.045. This system may be obtained at no cost through the URL(Universal Resource Locator) provided in Table 1.

In a study carried out by Mohsin [12], the primary objective was to assess the significant consequences of traffic accidents. The study utilized the FARAS dataset, which encompasses information pertaining to microcontrollers, the Global Positioning System, and a range of sensors. The dataset is used for accident detection and evaluation of the external characteristics of the involved automobiles. The FARAS dataset comprises standardized descriptions of each reported fatal crash. NHTSA provides this dataset and is available on the government's website. This data specifically pertains to the occurrence of fatal injuries resulting from motor vehicle traffic crashes. The dataset can be accessed through the hyperlink provided in Table 1.

A vision-based approach was reported by Agarwal which made use of the traffic-Net dataset provided by the Bing Image Search API [13]. The dataset obtained from Traffic-Net comprises a total of 4,400 photographs capturing traffic scenes, with each of the four groups containing approximately 1,100 images. However, the availability of road accident image databases specifically tailored to Indian traffic patterns is limited. Two distinct datasets were employed for the purpose of image classification, one being a standard set of images while the other was acquired by web surfing. The Traffic-Net dataset comprises a collection of traffic photographs, which have been gathered with the purpose of facilitating the training of vision-based models.

The dataset that was employed by Yeole [14] comprised information gathered by the Pimpri Chinchwad Municipal Corporation (PCMC) between 2014 and 2019. A total of 887 significant and trivial incidents have been documented within the specified context. The data were gathered between 2009 and 2018. The data was collected through two expressways located in Japan, namely the Toumei Expressway and Syutokou Expressway. Data was obtained at regular intervals of 5 minutes, specifically at the segment level for each part of the highway. The data was processed through attribute selection and deep neural networks along with K-means was employed.

Menendez conducted research utilizing a sensor-based dataset to generate simulations resembling real-life scenarios [15]. This was accomplished through the utilization of an open urban driving simulator known as CARLA [16]. The CARLA platform has been meticulously designed to facilitate the advancement, instruction, and evaluation of self-driving systems specifically tailored for urban environments. The provision of open digital assets by CARLA is made possible by its freely accessible code and procedures. Urban layouts are one aspect of these publicly accessible web pages. The interface further includes elements like infrastructures, and automobiles. These assets have been specifically developed for CARLA and are available for unrestricted use. Menendez presented a methodology centered on ML for predicting and detecting accidents. This approach exclusively considers the data obtained from the sensors installed in the vehicle. In addition, the dataset encompassed various circumstances and collisions, making up a total of 6,308,617 records that were utilized for experimental implementation and further study.

In 2018, Contreras devised the Maximum Sensitivity Neural Network, utilizing a dataset sourced from the INEGI database [17]. The database maintained by the National Institute of Statistics and Geography (INEGI) includes records of ground transportation accidents occurring in both urban and suburban regions. The present study examines the various categories of accidents as documented in the INEGI database. The following incidents are commonly observed in transportation accidents: passenger crashes, collisions with an animal, collisions with a cyclist, collision with a railroad, and crash with a motorcycle. The first case involves a collision with a pedestrian, resulting in the individual being run over. The second incident involves a collision with a motor vehicle. All the elements from these use-case scenarios were incorporated in this dataset, with every single one of them having a separate class.

The Decision Tree (DT) approach was utilized in the analysis of accident data from Historical North Dakota State, as documented[17]. The accident and incident database offer comprehensive information pertaining to accidents, including details regarding the place, time, environmental factors, and circumstances surrounding the occurrence of these incidents. The database known as the highway-rail crossing inventory provides comprehensive information regarding the multiple attributes associated with each junction. A novel dataset was constructed by employing the HRGC. The database featured comprehensive historical collision data pertaining to (HRGCs) in the state of North Dakota, spanning the years 1996 to 2014. During the specified time frame, North Dakota housed a total of 5,713 (HRGC) facilities, out of which 354 possessed documented actual accident reports.

In a study, Kezebou conducted the data collection focused on Highway incident detection (HWID12) [19]. The Highway Data Set (HWID12) is comprised of Eleven distinct categories for highway incidents. This dataset holds a special class for representative traffics samples that are regarded as being undesirable. In contrast to currently available datasets that include a multitude of random behaviors, the proposed data set (HWID12) has been purposefully curated to cater to the specific objective of identifying instances through the utilization of cutting-edge behavior identification technology. Consequently, HWID12 exclusively comprises data pertaining to incidents across various categories.

**Table 1.** Data Sources highlighting their respective usage and significant features

| Dataset | Source | Features |
|---|---|---|
| Traffic flow state data 495.493–539.045 California, USA [11] | URL: http://pems.dot.ca.gov/. | The information includes the 495.493-539.045-mile stretch of California's Interstate 5 used for the study. Within a specific time frame, the data stores upstream and downstream traffic flow information. |
| FARAS dataset [12] | URL: https://catalog.data.gov /dataset | It provides details regarding microcontrollers, the GPS, and sensors. The tangible elements of the driving vehicle are being described here. |
| Traffic-Net Dataset Bing Image Search API [13] | URL: https://github.com/Olafenwa Moses /Traffic-Net | Open-source software and interfaces make up CARLA, which may be used to access digital assets like city plans, structures, automobiles, and autonomous driving infrastructure. |
| CARLA [16] | URL: https://carla.org/ | CARLA is comprised of interfaces and open-source software that facilitate access to digital assets such as autonomous urban transportation systems, infrastructures, and automobiles. |
| INEGI database [17] | https://en.www.inegi.org.mx /programas/accidentes/#Tabular_data | Accident involving a passenger automobile, bicycle, railroad, motorcycle, pedestrian, etc. |
| North Dakota State accident data [18] | Key-search: Highway-Rail Grade Crossing Accident Data. URL: https://data.transportation.gov/Railroads | The time, date, location, and atmospheric conditions at the time of incidents were all included in the data set. The database of highway-rail crossings includes information about the exact spot and traffic at each crossing. |
| HWID12 (Highway Incidents Detection) [19] | url: https://www.kaggle.com/datasets | There are 11 distinct types of highway incidents, and a 12th kind for inadequate traffic samples. |

| | | | | |
|---|---|---|---|---|
| | | /landrykezebou/hwid12-highway-incidents-detection-dataset | | |

2.1. Metadata

The metadata pertaining to the aforementioned data search has been provided to offer thorough details regarding the datasets and methodologies employed in order to get the desired outcomes. The metadata contains descriptive information about its contents. Furthermore, this will enhance the accessibility, verification, and reusability of the aforementioned information.

**Table 2.** Data Sources of Road Accidents and Their Features

| Dataset | Preprocess | Methods | Novelty | Results |
|---|---|---|---|---|
| Traffic flow state data 495.493–539.045 California, USA [12] | RF, case-control sample analysis, multiparameter fusion clustering analysis | Bayesian network model | physical importance of explanatory factors for all sorts of vehicle safety data | accuracy = 84.9%, Crash prediction 60.8%, 3.non-crash accuracy = 92.3% |
| FARAS dataset [13] | PARAM (feature selection) =MaxDepth, K, Batch Size | ML classifiers K-NN Random Forest, and Gaussian Mixture Model | Improving the accuracy with severity. a fusion model by an ensemble of the KNN and RF | RF and KNN better than GMM. Correlation Coefficient=0.17 Mean Absolute Error=0.29 |
| Traffic-Net Dataset Bing Image Search API [14] | Noisy/mislabeled images removal. Image Augmentation Fixed Image Resolution. | CNN, ResNet50, DenseNet, EfficientNet-B1 | Specific data about a certain region of India-Mumbai. | Accuracy=84%. |
| Mixed traffic flow dataset from Pimpri Chinchwad Municipal Corporation (PCMC) [15] | 70% for preparation. 15% for Training. 15% for Testing. 15% for Validation. | ANN | New model with ANN and Comparison with Multiple Linear Regression | Multiple regressions Model=88% ANN Model performance= 93% |
| Carla [16] | Data simulation, Sensor connectivity- LiDAR, IMU Measurement, GNSS | ML Classifiers: LR, Linear Discriminant Analysis, NB, KNN, SVM, NNET, Ada Boost, Gradient Boosting, XGBoost | ML based system for collision prediction and detection. | Collision detection with 99% accuracy Risk level accuracy =43%. |

| | | | | |
|---|---|---|---|---|
| INEGI database [17] | Data Preprocessing, weight Adjustments | ANN | Maximum Sensitivity Neural Network based tool | Desity + Schedule GE=0.001597<br>Desity + Cause GE= 0.0011402<br>Schedule + Cause GE=0.0045598 |
| Collision data of North Dakota State [18] | Probability and decision profit values adjustments | DT | Evaluation of HRGC crashes | Accuracy event class=84.1%<br>Accuracy nonevent class = 77.2% |
| HWID12 (Highway Incidents Detection) dataset [19] | Modelling and dataset comparisons | 3D convolutional networks | New dataset and Performance benchmarking | accuracy of 88% and 98% for 12 and 2 class recognition |
| Dataset from Shanghai Public Security Bureau [20] | Data analysis and processed, R-tool. | RF, K-fold cross validation | FLD algorithm and other models' comparison | FLD = 92% |
| Federal Railway Administration's (FRA) data [21] | Resampling, Bagging | Random Forest Decision Trees | Comparison of two methods on 2 different datasets. | Accuracy of RF is 95.2%<br>Accuracy of decision tree is 77.7% |
| Dataset Addis Ababa, Ethiopia [22] | Classification | Hybrid K-means and Random Forest | hybrid K-mean, RF techniques were devised | accuracy of 99.86% |
| Dataset 5000 pieces Wuhan Transportation Authority [23] | Data Cleaning and Characteristics labeling | SVM, GA, SSA,PSO, GWO, | Practicable model for traffic forecast problems | SSA-Support Vector Machine > GWO- SVM, GA-SVM and PSO- SVM |
| 5793 Road accidents dataset from Abu Dhabi (2008-2013) [24] | Data Cleaning, deletion of invariant columns & categorization | ANN | 2 different classification ways were used. | Prediction=81.6%<br>ROC curve (AUC)= 1.00 |
| Dataset section 160 to 166 mile of I-15 | Classification | CNN | TAP-CNN | Prediction Accuracy=78.5% |

| | | | | |
|---|---|---|---|---|
| highway in the US. [25] | | | | |
| Traffic control stations data northern California. [26] | -Not available | LSTM, CNN, Deep Traffic Flow CN, LSTM Traffic Flow | Decentralized DL-based method | Successful congestion prediction. |
| S4A, InSync, BlueMAC, NOAA, Data from Orlando, Florida [27] | (SMOTE) | LSTM-CNN, XGBoost, Bayesian LR, LSTM, | A real-time artery collision risk forecasting model. | Sensitivity=88% AUC value =0.93 |
| Twin Cities Metro Data, Minnesota [28] | SMOTE-ENN removal | LSTM, RNN, CNN | Loop sensor data with deep abstraction using LSTM | TP=0.71 FP=0.25 |
| Dataset with 6,750,072 observations Source: Multiple [29] | Data augmentation to 2-D matrix | DCGAN LR, SVM ,AAN, CCN | DCGAN using unbiased data. | Sensitivity=0,88 and specificity = 0.907 |
| Office of Highway Safety Planning (OHSP) 2010-2016. [30] | SMOTE | SVM, AdaBoost,NB,RF, an d LR | Comparison of ML Algorithms | RF accuracy=75.5% better for predicting |
| Emirate of Abu Dhabi 2008-2013 [31] | Classification Screening of Data | NB, DT, SVM, Multi-Layer Perceptron | Model using 4 classification methods | DT (TS,10-fold-crossV, TS) was 81%, 73.62%, and 88.08%. |

2.2. Machine Learning based Road Accident Prediction Datasets

The assessment of accident data by employing ML techniques across various scenarios allows to pinpoint important factors that influence how severe injuries are. In addition to this, it helps in the process of choosing appropriate input data for the generation of predictive models. Models are constructed by utilizing accident data records, enabling an understanding of many attributes such as drivers' conduct, roadway conditions, lighting, atmospheric circumstances, and other relevant factors. This feature enables users to calculate safety measures that are beneficial in preventing fatalities. Various ML methods are employed for the purpose of identifying and forecasting the incidence of road accidents. A few of these algorithms are further elaborated on in the subsequent sections.

Mohsin [12] conducted an investigation into the significant consequences of traffic accidents, with a specific emphasis on analyzing the FARAS dataset. The computer vision-based ML classifiers that had the highest accuracy were K-Nearest Neighbors (KNN), Random Forest (RF), and (GMM). Notably, KNN and RF exhibited particularly promising results in the context of accident detection. When comparing the performance of the Gaussian Mixture Model (GMM) with Random Forest (RF) and K-Nearest Neighbors (KNN), it was observed that the last two methods yielded superior results. The work done includes a fusion model that combines the K-Nearest Neighbors (KNN) and Random Forest (RF) algorithms. Subsequently, a test was conducted to showcase the consequences of this fusion approach.

The traffic data preparation procedure in [20]utilized a dataset sourced from the Shanghai Public Security Bureau. Several studies have employed K-Cross Validation and Random Forest for model

prediction. Upon comparing the outcomes of several methodologies, such as FLD, RF, and Bagging decision tree with the KAPPA variables, it was evident that the FLD algorithm exhibited superior efficacy in crash prediction. The proper outcomes consist of a total of 95 cases. Contrary to the anticipated figure of 1.00092, as shown by the examination results, only a single occurrence of minor accidents was detected.

Zhou in 2020 have created a predictive model for traffic crashes occurring near rail-grade crossings instead of decision trees [21]. This study's primary data sources were the accident database and inventory data of the FRA pertaining to highway-rail crossings. The decision tree model's accuracy was 77.7%, but the random forest model achieved an accuracy of 95.2%.

An insightful framework was put forth by Yassin in 2020 that helps in determining the attributes that contribute to the seriousness of traffic accidents [22]. A hybrid approach combining the K-means and Random Forest (RF) methodologies was developed in order to identify the key features that exhibit the highest predicting capability for road accidents. The research's collection of information contained a total of 5000 road traffic accidents, sourced from FTP archives spanning the years 2011 to 2018 in Addis Abeba, Ethiopia. The evaluation and findings indicated that, out of the several categorization methodologies, the proposed methodology demonstrated a 99.86% level of accuracy.

**Table 3.** Road Accident Prediction Datasets using ML

| **Dataset Name** | **Data Source** | **Machine learning Method** | **Results** |
|---|---|---|---|
| FARAS dataset [12] | URL : https://catalog.data.gov /dataset | ML classifiers K-NN Random Forest, and Gaussian Mixture Model | RF and KNN better than GMM. Correlation Coefficient=0.17 Mean Absolute Error=0.29 |
| Shanghai Public Security Bureau database [20] | Dataset from Shanghai Public Security Bureau | Random Forest, K-fold cross validation | FLD = 92% |
| Federal Railway Administration's (FRA) data [21] | Data from the database of FRA available online. | Random Forest Decision Trees | Accuracy of RF is 95.2% Accuracy of decision tree is 77.7% |
| Dataset Addis Ababa, Ethiopia [22] | 5000 road traffic accidents compiled from FTP archives from year 2011 till 2018 | Hybrid K-means and Random Forest | accuracy of 99.86% |

2.3. Deep Learning Datasets for Traffic Collision Forecasting

DL has been extensively employed in the realm of road accident prediction, enabling proactive identification of potential incidents prior to their actual occurrence. Numerous DL techniques, such as Convolutional Neural Networks (CNN), Artificial Neural Networks (ANN), Long Short-Term Memory Networks (LSTM), (ARIMA), Recurrent Neural Networks (RNN), Deep Convolutional Generative Adversarial Networks (DCGAN), Support Vector Machines (SVM), and others, continue to be employed for the prediction and detection of accidents that occur on roads and highways. DL approaches require a robust dataset that is unbiased and encompasses diverse properties suitable for training a model.

Yeole conducted a study utilizing the dataset from the Pimpri Chinchwad Municipal Corporation (PCMC) spanning the years 2014 to 2019 [14]. A total of 887 significant and minor accidents have been documented within the specified context. Data was distributed in three categories for model development and evaluation70% of the data was put to use for developing models and training. 15% for the assessment of performance. The model's efficacy was assessed using 15%. The model performance was 88%, whereas the ANN model performance was 93%.

In Zhong's suggestion for a viable model addressing challenges in traffic forecasting, a total of ten independent parameters spanning both internal and external influences were selected. [23] The dataset includes around 5000 variables and is comprised of information regarding road accidents that were gathered through accident tracking database maintained by the Wuhan Transportation Authority. The methodologies employed in this study encompassed Grey Wolf Optimization (GWO), Support Vector Machines (SVM), Genetic Algorithm (GA), SSA, and PSO. The computational findings indicate that the SSA-SVM demonstrates effective predictive performance when compared to the GWO-SVM, GA-SVM, and PSO-SVM.

The proposal by Alkheder [24] was that a classifier has to be utilized to forecast collisions, of which 5973 traffic collisions took place in Abu Dhabi. The dataset encompassed the time period spanning from 2008 to 2013. The ANN classifier was developed utilizing the WEKA software. Weka is a software application utilized for the purpose of data mining. The model's general prediction performance for training set was 81.6%. The model generated 74.6% for the testing set of data. An AUC value of 0.5 was obtained for a test that is fully random, whereas a value of 1.00 was observed for a flawless test.

In 2017, Wenqi et al. proposed a novel model (TAP-CNN) for forecasting road accidents. This model takes into account the various elements that have an impact on traffic accidents [25]. The approach utilized in this study was CNN, and the results demonstrated a prediction frequency of 78.5%. The dataset utilized in this study was obtained from the United States portion of the I-15 motorway, spanning a distance of around 160 to 166 miles.

**Table 4.** Deep Learning Datasets for Traffic Collision Datasets

| **Dataset Name** | **Data Source** | **Deep learning Method** | **Results** |
|---|---|---|---|
| Pimpri Chinchwad Municipal Corporation (PCMC) Dataset [14] | Mixed traffic flow dataset from online portal ofPCMC | Artificial Neural Network | Multiple regressions Model=88%<br>ANN Model performance= 93% |
| Wuhan Transportation Authority [23] | Dataset 5000 pieces from WTA<br>Wuhan Transportation Authority | SVM, GA, SSA,PSO, GWO, | SSA-Support Vector Machine > GWO-SVM, GA-SVM and PSO- SVM |
| Road accidents dataset from Abu Dhabi [24] | 5793 Road accidents dataset from Abu Dhabi (2008-2013) | Artificial Neural Network | Prediction=81.6%<br>ROC curve (AUC)= 1.00 |
| I-15 highway in the US [25] | Dataset section 160 to 166 mile of I-15 Highway data | Convolution Neural Network | Prediction Accuracy=78.5% |

**3. Road Accident Forecasting Datasets using Machine Learning and Deep Learning Algorithms**

Various subcategories of computer vision-based algorithms are currently being employed in order to achieve enhanced accuracy and authenticity in models. Various subcategories, such as Long Short-Term Memory networks (LSTMs), AdaBoost, DCGAN, XGBoost, and MLR, are employed to enhance accuracy in ML tasks. LSTMs, in particular, are widely utilized for the purpose of acquiring, analyzing, and categorizing sequential data. Some common uses of Long Short-Term Memory (LSTM) include analyzing emotions, linguistic simulation, recognition of speech, and visual evaluation.

As a DL methodology, Fouladgar suggested the application of an actual transportation dataset obtained from the Caltrans Performance Measurement System (PeMS) in the state of California. Several DL-based techniques, including LSTM, CNN, and RMSE, were utilized in the development of the decentralized approach. According to the findings, predictions of congestion turned out to be accurate.

LSTM-CNNs were used to create a real-time roadway intersection collision prediction model [27]. CNN extracted the time-invariant features, while LSTM captured the long-term dependencies. Five

prominent models, including XGBoost, BLR, and LSTM, were created with the purpose of assessing and comparing their performance to that of the designed model. The dataset comprises 38 segments of arterials and 21 intersections, spanning a distance of one and a half miles. The greatest sensitivity of 88% was attained by utilizing data from four urban arterials situated in Orlando, Florida, together with information from S4A, InSync, BlueMAC, and NOA. Furthermore, in comparison to alternative methodologies such as LSTM, CNN, XGBoost, and BLR. The suggested strategy has an AUC value of 0.93, which was the highest of all of the methods.

XGBoost is a powerful machine-learning method that can improve data analysis and decision-making. XGBoost can be regarded as a software framework that implements gradient-boosting decision trees. Data scientists and researchers globally have utilized it for the purpose of enhancing their machine-learning models.

In their study, Mehrannia utilized data obtained from loop detector devices installed on the Twin Cities Metro in Minnesota motorways. This dataset was employed to develop a novel framework. The employment of a Long Short-Term Memory (LSTM) based architecture was the primary approach that was taken throughout the course of this research project's methodology. In addition to the aforementioned submethods, the Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM) were also employed. The findings of the study indicate that the devised methodology performed better than alternative methods under the same experimental conditions. The model had a true positive rate of 0.71 and False positive rate of 0.25.

The Deep Convolutional GAN (DCGAN) is an architectural design of a generative adversarial network. This approach employs several specific guidelines, notably: The suggested modification involves substituting pooling layers in the filter with convolutions and employing fractional-stride convolutions in the framework. The utilization of batch normalization in both the generator and the discriminator is employed.

The Deep Convolutional Generative Adversarial Network (DCGAN) model was suggested to be used by Cai in his research released in 2020 to fully apprehended the traffic data that contributes to accidents [29]. In the DCGAN framework, the input data was transformed into a two-dimensional matrix. Crash prediction algorithms utilized in this study encompassed the LR model, SVM, ANN, and CNN, each applied to a balanced collection of data. A comprehensive evaluation was conducted on all twelve models. The results suggested that DCGAN exhibited superior performance and a deeper understanding of accident prediction. The entire dataset, consisting of 6,750,072 observations, was utilized. Among these observations, there were 625 instances of collisions and 6,749,447 instances of non-collision situations.

The ML approach, known as AdaBoost, is used in the framework of an ensemble method. The prevailing estimator employed in conjunction with AdaBoost is decision trees characterized by a single level, indicating decision trees with a sole split. These arboreal entities are alternatively referred to as Decision Stumps.

AlMalook proposed many strategies, namely Support Vector Machines (SVM), AdaBoost, Naive Bayes (NB), Random Forest (RF), and Logistic Regression (LR), for the purpose of comparing models in order to categorize the degree of seriousness of crashes taking place. The Random Forest algorithm yielded a highest accuracy rate of 75.5%. The accuracy of logistic regression (LR) was reported to be 74.5% by other researchers. The numerical value displayed by the NB was 73.1%. The AdaBoost algorithm demonstrated a classification accuracy of 74.5%. The data included in this study was obtained from the Office of Highway Safety Planning (OHSP). The duration selected was spanning from 2010 to 2016.

MLR, also referred to as multiple regression, is a statistical approach employed for data analysis. The primary objective of this methodology is to forecast the probable result by considering several factors, hence establishing the correlation between multiple independent variables and a single dependent variable.

Taamneh made predictions on the severity of car crashes in Abu Dhabi [31]. The dataset utilized in this study spans from 2008 to 2013. The classification techniques implemented encompass decision trees, rule induction, naive Bayes, and multilayer perceptrons. In the case of the Decision Tree (DT) model, particularly when applied to the training set with a 10-fold crossV technique, the obtained accuracies for the J48 algorithm were 81%, 73.62%, and 88.08%. The accuracy scores for the PART algorithm were 81% for training. The 10-fold crossV had accuracy of 71.22% and test set had accuaracy of 82.18%.

Many other researches have been conducted for the prediction of severity of traffic crashes. The severity of these crashes with parameters like weather [32], road conditions that include use of both ML as well as DL. [33] [34].

**Table 5.** Datasets in prediction of Road Accidents using Fusion of Deep Learning and Machine learning algorithms

| Dataset | Method | Results |
|---|---|---|
| Traffic control stations data northern California. [26] | LSTM, CNN, Deep Traffic Flow CN, Traffic Flow | Successful congestion prediction. |
| S4A, InSync, BlueMAC, NOAA, Data from Orlando, Florida [27] | LSTM-CNN, XGBoost, Bayesian LR, LSTM, | Sensitivity=88% AUC value =0.93 |
| Twin Cities Metro Data, Minnesota [28] | LSTM, RNN, CNN | TP=0.71 FP=0.25 |
| Dataset with 6,750,072 observations. Source: Multiple [29] | DCGAN LR, SVM ,AAN, CCN | Sensitivity=0,88 and specificity = 0.907 |
| Office of Highway Safety Planning (OHSP) 2010-2016. [30] | SVM, AdaBoost, NB, RF, and LR | RF accuracy=75.5% better for predicting |
| Emirate of Abu Dhabi 2008-2013 [31] | NB, DT, SVM, Multi-Layer Perceptron | DT (TS,10-foldcrossV, TS) was 81%, 73.62%, and 88.08%. |

## 4. Discussion

The choice of source of data for research of automobile accidents is determined by the type of traffic problem that needs to be addressed. After determining the causes of traffic mishaps and the extent to which each element contributed to those mishaps, the investigative methods could finally be refined. In the future, the datasets that were discussed in this research could be utilized to train a variety of DL models. Over the course of the past few years, there has been an obvious lack of emphasis on the incorporation of spatial elements in data sets documenting occurrences of traffic on roads and highways. Hence, incorporating these attributes, training the datasets utilizing these constituents, and further developing them in subsequent iterations will yield significant benefits for both academic research and practical implementation.

## 5. Conclusions

The information collected from traffic collisions is a valuable resource that assists traffic safety programs with their analysis and preventative efforts. The use of these training programs often results in a decrease in the occurrence of automotive fatalities. They are utilized by a multitude of governing bodies, such as highway safety administrators tasked with educating the public and implementing efficacious preventive measures, authorities attempting to identify the culprits behind traffic accidents, and road safety researchers seeking access to reliable crash data. These entities are also responsible for formulating long-term decisions regarding traffic safety and nationwide strategic initiatives for road and traffic safety. In this paper, a brief overview of the traffic data sets, their sources, links, and methodology, as well as a discussion of how different DL branches have merged these datasets for enhanced crash detection and prediction, were included.

**References**


1. Agarwal, N. J. (2019). Intelligent System for Road Accident Detection in India using Deep Learning Models.
2. Agarwal, N. J. (2019). Intelligent System for Road Accident Detection in India using Deep Learning Models.
3. Ahmed, S. K.-K.-S. (2023). Road traffic accidental injuries and deaths: A neglected global health issue. . Health science reports, 6(5), (e1240.).
4. Alkheder, S. T. ((2017)). Severity prediction of traffic accident using an artificial neural network. Journal of Forecasting, 36(1), 100-108.
5. AlMamlook, R. E. ((2019, April). ). Comparison of machine learning algorithms for predicting traffic accident severity. In 2019 IEEE Jordan international joint conference on electrical engineering and information technology, 272-276.
6. Cai, Q. A.-A. ((2020).). Real-time crash prediction on expressways using deep generative models. . Transportation research part C: emerging technologies, , 117, 102697.
7. Chen, C. ((2017). ). Analysis and forecast of traffic accident big data. . In ITM Web of Conferences . EDP Sciences., Vol. 12, p. 04029.
8. Cong Chen, G. Z. (2016). Examining driver injury severity outcomes in rural non-interstate roadway crashes using a hierarchical ordered logit model. . Accident Analysis & Prevention, 79-87.
9. Contreras, E. T.-T. ((2018)). Prediction of car accidents using a maximum sensitivity neural network. In Smart Technology: First International Conference, MTYMEX 2017, Monterrey, Mexico, May 24-26, 2017, Proceedings Springer International Publishing., (pp. 86-95). .
10. Delen, D. S. (2006). Identifying significant predictors of injury severity in traffic accidents using a series of artificial neural networks. . Accident Analysis & Prevention, 38(3), 434-444.
11. Ditcharoen, A. C. (2018). Road traffic accidents severity factors: A review paper. International Conference on Business and Industrial Research (ICBIR), 5th, pp. 339-343.
12. Dosovitskiy, A. R. ( (2017, October). ). CARLA: An open urban driving simulator. In Conference on robot learning, PMLR., (pp. 1-16). .
13. Fouladgar, M. P. (2017, May). Scalable deep traffic flow neural networks for urban traffic congestion prediction. In 2017 International Joint Conference on Neural Networks (IJCNN) IEEE., (pp. 2251-2258).
14. Gianfranco, F. S. (2018). An accident prediction model for urban road networks. . Journal of Transportation Safety & Security, 10(4), , 387-405.
15. Kezebou, L. O. ((2022, May).). Highway accident detection and classification from live traffic surveillance cameras: a comprehensive dataset and video action recognition benchmarking. In Multimodal Image Exploitation and Learning 2022 SPIE, Vol. 12100,, pp. 240-250.
16. Kononen, D. W. (2011). Identification and validation of a logistic regression model for predicting serious injuries associated with motor vehicle crashes. . Accident Analysis & Prevention, 43(1), , 112-122.
17. Lee, J. Y. ((2019). ). Model evaluation for forecasting traffic accident severity in rainy seasons using machine learning algorithms. Seoul city study. Applied Sciences, 10(1), , 129.
18. Li, P. A.-A. ((2020). ). Real-time crash risk prediction on arterials based on LSTM-CNN. . Accident Analysis & Prevention,, 135, 105371.
19. Mehrannia, P. B.-B. (2023). Deep representation of imbalanced spatio-temporal traffic flow data for traffic accident detection. . IET Intelligent Transport Systems, 17(3), , 606-619.
20. Menendez, H. D. (2023, April). Detecting and Predicting Smart Car Collisions in Hybrid Environments from Sensor Data. . In WorldCist'23-11st World Conference on Information Systems and Technologies. Springer.
21. Mohammed, A. A. (2019). A review of traffic accidents and related practices worldwide. The Open Transportation Journal, 13(1).
22. Mohsin, A. R. (2023). Accident Detection and Classification using IoT Fusion-Enabled Framework with Machine Learning Classifiers. . 한국통신학회 학술대회논문집, , 1322-1324.
23. Niranga Amarasingha, S. D. (2014). Gender differences of young drivers on injury severity outcome of highway crashes. Journal of Safety Research, e1-120.
24. Rahim, M. A. (2021). A deep learning-based traffic crash severity prediction framework. . Accident Analysis & Prevention, , 154, 106090.
25. Sameen, M. I. (2017). Severity prediction of traffic accidents with recurrent neural networks. . Applied Sciences,, 7(6), 476.
26. Susanne Kaiser, G. F. (2016). Aggressive Behaviour in Road Traffic – Findings from Austria. Transportation Research Procedia(14), 4384- 4392. .
27. Taamneh, M. A. ((2017). ). Data-mining techniques for traffic accident modeling and prediction in the United Arab Emirates. Journal of Transportation Safety & Security, , 9(2), 146-166.
28. Wenqi, L. D. (2017, September). A model of traffic accident prediction based on convolutional neural network. 2nd IEEE international conference on intelligent transportation engineering (ICITE) IEEE, 2nd, (pp. 198-202). .
29. Yassin, S. S. ((2020).). Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. . SN Applied Sciences,, 2, , 1-13.

30. Yeole, M. J. ( (2022). ). Prediction of Road Accident Using Artificial Neural Network. . International Journal of Engineering Trends and Technology, 70(3),., 151-161.
31. Zhao, L. L. (2023). Highway Traffic Crash Risk Prediction Method considering Temporal Correlation Characteristics. Journal of Advanced Transportation.
32. Zheng, Z. L. ((2016).). Decision tree approach to accident prediction for highway–rail grade crossings: Empirical analysis. . Transportation Research Record, 2545(1), ., 115-122.
33. Zhong, W. &. (2023). Predicting Traffic Casualties Using Support Vector Machines with Heuristic Algorithms: . A Study Based on Collision Data of Urban Roads. Sustainability,, 15(4), , 2944.
34. Zhou, X. L. ((2020).). . Accident prediction accuracy assessment for highway-rail grade crossings using random forest algorithm compared with decision tree. Reliability Engineering & System Safety,.