

Data Mining Methods for Traffic Accident Severity Prediction

Qasem A. Al-Radaideh and Esraa J. Daoud

Abstract— The growth of the population volume and the number of vehicles on the road cause congestion (jam) in cities that is one of the main transportation issues. Congestion can lead to negative effects such as increasing accident risks due to the expansion in transportation systems. The smart city concept provides opportunities to handle urban problems, and also to improve the citizens' living environment. In recent years, road traffic accidents (RTAs) have become one of the largest national health issues in the world. Many factors (driver, environment, car, etc.) are related to traffic accidents, some of those factors are more important in determining the accident severity than others. The analytical data mining solutions can significantly be employed to determine and predict such influential factors among human, vehicle and environmental factors and thus to explain RTAs severity. In this research, three classification techniques were applied: Decision trees (Random Forest, Random Tree, J48/C4.5, and CART), ANN (back-propagation), and SVM (polynomial kernel) to detect the influential environmental features of RTAs that can be used to build the prediction model. These techniques were tested using a real dataset obtained from the Department for Transport of the United Kingdom. The experimental results showed that the highest accuracy value was 80.6% using Random Forest followed by 61.4% using ANN then by 54.8% using SVM. A decision system has been build using the model generated by the Random Forest technique that will help decision makers to enhance the decision making process by predicting the severity of the accident.

Keywords— Decision Making, Traffic Accidents Severity Prediction, Data Mining Methods, Knowledge based Systems.

I. INTRODUCTION

CITIES around the globe are growing steadily, and the common trends of the world are the urbanization process and making these cities “smarter” (Anastasi et al., 2013; Jensen et al., 2014). According to (Anastasi et al., 2013), around 60% of the population of Europe lives in cities, and to (Mulligan & Olsson, 2013) by 2050, 70% of the world's population will live in cities. Moreover, the Global Health Observatory (GHO) announced that around 6 out of every 10 people by 2030 and 7 out of 10 people by 2050 will live in an urban area (Madakam & Ramaswamy, 2015). As for Jordan, the report by the Jordanian Department of Statistics (2015) showed that 42.04% of the population lives in the Jordanian capital, Amman.

Qasem Al-Radaideh is with the Department of Computer Information Systems, Yarmouk University, Irbid, Jordan. (qasemr@yu.edu.jo).

Esraa Dauad is with Computer Center, AlAlbait University, Mafraq, Jordan. (esraa_jamil_cis@yahoo.com).

The urbanization process leads to upgrading people's standard of living, education, health services, and suitable transportation. Rapid urbanization produces significant challenges and issues that need to be addressed. The smart city concept provides opportunities to handle such challenges and urban problems, and also to improve the citizens' living environment (Yin et al., 2015).

The concept of smart cities has risen rapidly during the latest few years, defined as a consequence of evolution in information and communication technologies. Smart sites link a digital infrastructure with the physical elements of the city to improve performance, to achieve a high quality of life in cities, and to reduce the environmental impacts (Mulligan & Olsson, 2013; Abd-Elkawy, 2013).

Smart cities, which include city administration, transportation, education, healthcare, estate, and public safety, are IT complex systems of systems that are based on ICT (Information and Communication Technologies) and share information via communication networks. However, due to the fact that we deal with vast and heterogeneous resources and devices, the modeling of the various traffic patterns is introduced in such systems becomes extremely complex (Anastasi et al., 2013; Jensen et al., 2014; Madakam & Ramaswamy, 2015).

The growth of the population volume and the number of vehicles on the road cause congestion (set of vehicles moving nearby and slowly) in cities. Congestion is one of the main transportation issues in cities. It can lead to negative effects like air pollution, waste of time, money, and fuel, heart attack, emergency vehicle delay, and increasing accident risks due to the expansion in transportation systems (Zhang et al., 2011; Lécué et al., 2014).

For example, the growth of population and vehicles has been dramatic in Jordan, where the number of the Jordanian capital's population has increased to around 4 million in 2015, and the total annual population growth rate was about 5.3% between the years (2004 -2015), as the Jordanian Department of Statistics reported. Moreover, the vehicle ownership ratio rose to one vehicle for every 5 people in 2014, and the annual rate of increase in the number of vehicles has reached 7.78% within years (2005 -2014), as the Jordanian Traffic Department announced. Another example, at the beginning of 2010, the capital of China, Beijing, had 4 million vehicles and appended another 800 000 in the same year (Zhang et al., 2011).

As Lécué et al., (2014) cited that the urban traffic cost in the

USA has been (5.5) billion hours of travel delay in addition to 2.9 billion gallons of wasted fuel, all at the total price of \$121 billion. Furthermore, (Zhang et al., 2011) cited that the road traffic accidents that occurred in U.S. cities account for about 50%–60% of all congestion delays according to the Federal Highway Administration.

Mobility is one of the main dimensions of urban societies that include traffic system. Many traffic systems use ad-hoc sensors such as induction loops and cameras, but these sensors are restricted to monitor road locally where they are installed. Moreover, their installation and maintenance are very expensive. On the contrary, the GPS (embedded in the vehicle) monitors the entire road network virtually and requires low installation expenses.

More recently, social networks (e.g. Twitter and Facebook) have been considered as the social sensors and widely utilized as a source of information for the traffic events (real-world occurrence), such as traffic accidents and congestion or other events. Thus, with the assistance of social sensors, we can extract and detect which humans will present a certain event in near time and also can estimate traffic flow based on the social sensors (Anastasi et al., 2013; D'Andrea et al., 2015).

Status Update Message (SUM) refers to user message posted on social networks, and it may include geographic coordinates and current traffic state around the users while driving. The extracted events from social networks are employed with Intelligent Transportation Systems (ITSs). The ITS is an infrastructure that joins transport users, vehicles, and networks with Information and Communication Technologies (ICTs) and gives an ability of management and safety of transport networks. Furthermore, the ITS system provides real-time information like traffic congestion (Anastasi et al., 2013; D'Andrea et al., 2015).

In recent years, road traffic accidents (RTAs) have become one of the largest national health troubles in the world (Beshah et al., 2012; Liang, 2015). With the increasing number of vehicles in the traffic, RTAs have become a wide spreading and growing threat, causing the loss of human life (Shiau et al., 2015). Besides, the increased number of fatalities and injuries resulted from RTAs; the RTAs have a significant economic and social impact on the individuals and the governments (Yousif & AlRababaa, 2013; Devi et al., 2015).

According to the World Health Organization (WHO), there are more than 1.2 million people dying each year and more than 50 million ones who are injured worldwide. In addition, 518 billion dollars were the worldwide annual financial losses. The burden of the consequences of the accident has more impact on developing countries than in developed nations. For example, Jordan as a developing country has a high rate of RTAs between the years 1989- 2012, in which the fatalities were more than 13000 (Jadaan et al., 2014). Also, the annual rate of the increasing number of accidents has reached 3.2 % within years (2005 -2014) as the Jordanian Traffic Department reported.

According to the report of the Jordanian Traffic Department for the year 2014, there were about 102441 traffic accidents that caused about 688 fatal injuries, 2063 serious injuries, and

12727 slight injuries, and the proportion of fatal accidents in the capital Amman was 28.05% at an estimated cost of 239 million dinars yearly at the rate of 0.65 million dinars daily that is equivalent to 0.92% of the gross domestic product (GDP). In addition, there is approximately a traffic accident every 0.13 minute, an injured person every 35.54 minute and a dead person every 13 hours.

Many factors are related to traffic accidents, including infrastructure (environmental) such as (weather conditions and road signs), the vehicle itself (type and safety), the behavior of traffic user (driver, pedestrian, and passengers) and characteristics of the driver (age, using seatbelt and gender). Furthermore, some of those factors are more important in determining the accident severity than others. Thus, it is apparent that the analysis of the determinant factors of accident severity will help reveal more patterns and knowledge that can be used in prevention and safety strategies of traffic accidents (Kunt et al., 2011; Beshah et al., 2012).

The determinant factors for traffic accidents in Jordan reported by the Traffic Department (2014), where speed limits: 60 km/h, light conditions: daylight, road surface: dry, weather: clear, day of week: Thursday, time: 14:00-14:59, where the largest proportion of fatal accidents occurred at those factors with the approximate percentages of the total incidents that were 27.03%, 76.02%, 96.22%, 97.53%, 17.65%, and 8.28% respectively.

Data Mining in Traffic Accidents Area

With the technology revolution, data mining has developed as one of the major research domains in the recent decades for several reasons, such as the volume of data available for mining continues to grow at a tremendous rate in the large data storages which become seldom visited. As a result, the interpretation and making a decision based on such data exceed the human's ability. Data mining, known as the Knowledge Discovery in Databases (KDD) process, is the process of analyzing and categorizing data from many different dimensions and extracting useful and implicit knowledge and interesting patterns from the data which may contain a valuable decision, saving money or life protecting (Weiss & Davison, 2010; Han & Pei, 2011).

Currently, numerous systems of vehicular traffic safety have been developed in order to reduce the crashes. However, the abundant variables and complexity of relationships between the various transportation factors require analytical approaches, rather than traditional methods (Yousif & AlRababaa, 2013). In addition, the latest analytical data mining solution is significantly employed to determine and predict such influential factors among human, vehicle, and road (environmental) factors and thus to explain RTAs severity (Kunt et al., 2011; Yousif & AlRababaa, 2013; El Tayeb et al., 2015).

According to Jordan Traffic Institute reports (Masaeid, 2009; Obaidat, 2012), traffic accidents in Jordan are the main reason of fatality and the majority of errors that about 90% were caused by the drivers' mistakes and violation of safety rules such as speed of drivers. Nevertheless, 10% of the errors were caused by the deficiencies in the environmental factors.

In most of the RTAs, many of the deaths caused only due to the delay in the medical assist arrival. The ability to predict when and how RTAs will occur can help provide faster rescue operation. Moreover, descriptive analysis can guide to improve the road safety such as infrastructure design and human behavior by the targeted marketing campaigns (Beshah & Hill, 2010; Perone, 2015; Raut & Karmore, 2015).

As the number of vehicles on the road and the world population volume are increasing steadily and also with the static expansion of roadways, congestion became an increasingly global transportation issue that negatively affects all life dimensions like health and economy. In addition, the potential of a traffic accident may rise (Zhang et al., 2011; Lécué et al., 2014).

There are traditional ways to reduce road traffic accident. As a consequence, to reduce the traffic congestion, it is not always convenient or even needed to have the enormous expense such promoting public transport and increasing the trajectory capacity. Thus, the analytical and predictive techniques are required, such as data mining algorithms in order to make intelligent decisions to avoid further accidents, enhance transportation system and develop some intelligent traffic safety rules. This motivates us to use some classification data mining algorithms to predict and analyze road traffic accident (RTA) related to environmental factors.

The main objective of this research is to employ some classification techniques to detect the influential environmental features of RTAs that can be used to build predictive models to learn new knowledge patterns and thus to predict the severity of accidents from recorded traffic accident data. For this purpose, the accuracies of three classification algorithms that are Decision Trees, Artificial Neural Network (ANN), and Support Vector Machine (SVM) were investigated and compared to build the prediction model. The prediction model is then used to build a knowledge based prediction system.

II. RELATED WORK

As stated by several researchers, data mining techniques have a vast role in analyzing and predicting the severity of road accidents and in identifying the patterns of the components of accidents as spatial and non-spatial factors. In addition, the great potential of data mining prediction techniques plays a major role in preventing and controlling the problems of road accident safety.

In this section, we present some related works using data mining techniques to predict and analyze traffic congestion in urban areas, particularly smart cities, and also to predict and analyze the severity of RTAs.

A. Prediction and Analyzing Traffic Accidents Congestion

GPS-enabled vehicles are considered as mobile sensors that provide dynamic traffic information of a city's road network. So far, these sensors can be used for detecting potential traffic hotspots and jams (congestions) by analyzing the moving objects in trajectories and using data mining techniques. The following studies proposed mechanisms using GPS data.

Chang et al., (2010) presented a predictive approach for city's hotspot discovery. They applied three clustering algorithms: K-means, Agglomerative hierarchical clustering (AHC) and Density-based algorithm (DBSCAN) on GPS points using contextual information such as the weather condition. Then, they calculated a hotness index based on the spatial clusters and the distance between the driver's location and the cluster. The data collected in 2008 from June 25 to August 25 were given by five taxi drivers from the Taiwan Taxi Company; 2319 records were collected, but 487 records were eliminated due to the GPS readings for them were zeros or out of range.

Ong et al., (2011) applied flock mining (T-Flock algorithm) approach for traffic jams detection and classification on the road network using around 40,000 GPS-enabled cars in the Pisa city in Italy. The proposed method based on a combination of the data mining query language and T-Flock algorithm that was provided by MAtlas. The discovery of potential traffic congestions based on the speed of vehicles to filter and detect slow flock patterns regions, where a group of vehicles moves slowly together for a specified time, there are no percentage results.

Giannotti et al., (2011) designed an M-Atlas platform that provides the ability to integrate several mining algorithms such as clustering and classification techniques for the mobility knowledge discovery. They forecast the traffic related events as congestions (jams) possibility by predicting future regions of dense traffic and examining the variation of such areas over time on various trajectories, using a tree structure of frequent pattern (T-Patterns). The authors used two datasets of private onboard GPS cars in Italy cities. The first, Milano2007 dataset was about 17,000 cars with 200,000 travels over a week. The second, Pisa2010 dataset was about 40,000 cars with 1,500,000 trips over 5 weeks; there are no percentage results.

Liu et al., (2010) proposed a new clustering approach called mobility-based clustering that is based on the crowdedness (congestion) of regions to define traffic hotspots using vehicle speed information in the region. In this approach, low speed may indicate the high congestion and vice versa with taking the potential effects of other factors on vehicle speed into account, there are no percentage results.

Lécué et al., (2014) presented a Semantic Traffic Analytics and Reasoning for City (STAR-CITY) framework. STAR-CITY is a real-time city surveillance application for public transportation in Dublin and Ireland. Also, it can be applied to any city and contexts that use sensor data stream. STAR-CITY predicts congestion by considering all the traffic conditions like accidents and weather information. Moreover, STAR-CITY uses an apriori algorithm and produces the association rules between traffic snapshots. Besides, it analyzes historical and real-time data.

In the Jordanian study of RTA, Al-Zubi, (2010) developed a clustering approach and applied data visualization analysis of spatial data mining for Geographical Information System (GIS) of Amman city. The author clustered the accident datasets based on the spatial locations of accidents. Then

classified each spatial cluster based on the non-spatial attributes of the accidents. Accident hotspots, having the major role in defining reduction strategies of incidents, of the determined area were learned by identifying the geographical locations with high frequencies of accidents. Furthermore, the Minimum Bounding Rectangle (MBR) was applied for data visualization.

Social networks, such as Twitter and Facebook have been recently used as sources of information about traffic events, such traffic accidents, congestions and each user of these networks is considered as a sensor.

Anastasi et al., (2013) proposed a social-sensing approach named (SMARTY) that aims to develop an ICT Platform to innovate tools and services for mobility and transport in the smart city. The SMARTY paper monitors and analyzes user activities through online social networks. Also, it uses the real-time information extracted from, such networks like tweets and posts, to detect traffic events such as accidents and traffic congestion and also to suggest optimal tracks to the users. In addition, it uses data mining techniques for pre-processing and analyzing the heterogeneous collected data and then for the extraction of useful knowledge patterns from them. The authors built a labeled dataset consisting of 500 tweets that are classified in two SUMs (Status Update Messages) classes. They applied C4.5 classifier with 10-fold Cross-Validation and obtained a classification rate about 93.73%.

In other work for the detection of traffic jams using social sensors, D'Andrea et al., (2015) proposed a real-time surveillance framework for traffic event detection, such as congestion, from social media (Twitter stream analysis) by considering all Twitter users as traffic sensors. The authors applied text mining and classification techniques for fetching and classifying streams of tweets. Furthermore, their system is able to notify the presence of traffic events or not, and also to distinguish whether the traffic events are due to an external reason like a football match or not. They performed two experiments to classify each tweet in a class label, where the superiority was to SVMs among several classification models, namely C4.5, NB, PART, and KNN. In the first experiment, the achievable accuracy of SVM was 95.75% over 2-class balanced dataset consisted of 1330 tweets, while in the second experiment the accuracy of SVM was 88.89% over 3-class balanced dataset consisted of 999 tweets.

B. Prediction and Analyzing Traffic Accidents Severity

Classification techniques can be used for predicting and analyzing the severity and causes of road traffic accidents and can be employed for early notifying and alerting of accidents with the individual or hybrid model.

Pakgozar et al., (2011) employed Classification and Regression Trees (CART) and Multinomial Logistic Regression (MLR) approaches to get a descriptive analysis of human factor roles in the crash severity of 347285 accident records in Iran using SPSS based on Accident Severity variable (Fatal, Injury, No Injury). They found that the CART approach compared to MLR provided a higher accuracy and it is simpler to understand and interpret the results. The

accuracies obtained of CART and LR were 81% and 78.57% respectively.

Another geospatial analysis study by Effati & Sadeghi-Niaraki, (2015) presented an innovative knowledge discovery (OCART) approach based on a combination of ontological reasoning and (CART). In addition, they developed an ontology-driven geospatial framework to detect the crash severity through the proposed (OCART) method and applied a system prototype on a regional highway corridor to evaluate the performance of the proposed approach. According to their results, OCART improved the efficiency of CART that by discovering the new relationship between severity and factors of accidents.

The study by Beshah et al., (2013) was as an extension to their previous works in (Beshah et al., 2011) with the same dataset size 14,254 which consisted of 48 attributes. The authors utilized in their study Classification and Adaptive Regression Trees (CART), Random Forest, and TreeNet that added to the previous work to analyze accident data collected from the Traffic Office of Addis Ababa using a Salford Predictive Miners suite (SPM) tool. The experimental results showed that the TreeNet method compared to others provided the highest performance. The accuracies of 98.94 %, 86.59 %, and 84.5% were achieved for TreeNet, Random Forest, and CART respectively.

Another study by Beshah et al., (2012) utilized Classification and Adaptive Regression Trees (CART), Random Forest (RF), TreeNet and hybrid (combined models) ensemble approaches. As the results of their experiments, the ensemble technique performed better than other single classifiers in predicting injury severity. The overall accuracy of ensemble technique was 95.47%, TreeNet was 94.54%, CART was 93.52%, and Random Forest was 90.75%.

Beshah et al., (2012) applied data mining with the same dataset size, attributes number, city, and tool, which used in their other research (Beshah et al., 2011; Beshah et al., 2013), but with a difference in the empirical results, because the differences in class label values that were (fatal, injury or non-injury) in (Beshah et al. 2012), while in (Beshah et al. 2011; Beshah et al. 2013) were (injury and non-injury).

Krishnaveni & Hemalatha, (2011) conducted a perspective analysis of 34,575 traffic accident events in Hong Kong. They employed Naive Bayes, AdaBoostM1, J48, PART, and Random Forest classifiers to predict and detect the severity of injury and causes of accidents using WEKA tool. Moreover, they applied Genetic Algorithm for Feature selection to minimize dimensionality of the Accident dataset. According to the comparison results of classifiers, Random Forest outperforms all other algorithms. There are no percentage results.

Ibrahim et al., (2014) proposed a Real-time Transportation Data Mining (RTransDmin) method. RTransDmin had the ability to analyze real-time traffic data set and predict future useful information about traffic accidents. The decision tree ADTree and J48 algorithms were employed to build a classification model for a road accident dataset with 1385 records that was reported by the Department of Transport in

England using Weka and DTREG tools. DTREG developed confusion matrices with accuracy rate 87.2%, 85.9% for training and testing datasets respectively. The WEKA tool was used to develop a scatter plot for classification of objects.

In another study, Beshah and Hill, (2010) employed Naive Bayes, Decision Tree (J48), and K-Nearest Neighbors classifiers to develop a model that used to analyze and predict the role of road-related factors for traffic accident severity of 18,288 accidents in the Addis Ababa city with the accuracies that were 79.9967%, 80.221%, and 80.8281 % respectively. In addition, they used the PART algorithm to present the knowledge in the form of rules, with the accuracy of 79.94% using the WEKA tool.

Jadaan et al., (2014) employed Artificial Neural Network (ANN) approach to develop a prediction model for future road accidents in Jordan. MATLAB and SPSS statistical software were used, and four alternative models were developed with a different number of hidden layers. Model 4 was the best model that had the largest number of hidden layers and provided the highest coefficient ($R^2 = 0.992$). The model generated good results under Jordanian traffic conditions. Thus, it was found to be reliable to forecast future traffic incidents in Jordan.

Yousif & AlRababaa, (2013) implemented a Multilayer Perceptron (MLP) neural network technique to build a model to estimate and predicate the number of accidents over the time using the NeuroSolutions software, and Microsoft Office excels 2010. The back propagation learning algorithm (BP) and TanhAxon function were applied. They based on the datasets that were collected by the Jordan Traffic Institute and Traffic Department, which contained records of 10 years from 2002 to 2011. The achieved accuracy and call precision were 100% in identifying and classifying the accident type.

Shiau et al., (2015) presented various models to analyze and predict the causes of 2,471 traffic accidents in central Taiwan. The methods used were Fuzzy Robust Principal Component Analysis (FRPCA), Back-propagation Neural Network (BPNN), and Logistic Regression. They used Feature Selection that was a Recursive Feature Elimination (RFE). The experimental results displayed that the accuracy rate of the proposed FRPCA-BPNN 85.89% and FRPCA-LR 85.14%, combined with FRPCA, was higher than the rate of BPNN 84.37% and LR 85.06%.

Kunt et al., (2011) used an Artificial Neural Networks (ANN) with MLP, a genetic algorithm (GA), and a genetic algorithm combined with pattern search (PS) for forecasting the severity of 1000 recorded crashes in Tehran. Their experimental results using MATLAB software showed that the ANN achieved the highest accuracy of prediction with an R-value of around 0.87 then the combination of GA and PS with the of around 0.79 and GA of 0.79.

A novel framework based on the merging of computing and telecommunications was proposed by (Raut & Karmore, 2015) which was able to communicate anytime and anywhere. As in the proposed system, there is an onboard unit in each vehicle to sense the input signals from IR sensors and GPS devices to determine the position and speed of the motor and a control unit which uses ANN algorithm and fuzzy system. That could

help in decision making for real-time estimation of accident severity. Hence, the presented system could assist in providing immediate emergency and rescue assistance to save the victim life after accidents by reducing the response time of accident alert.

According to the authors' perspective, the proposed frameworks that can sense crashes notify them and estimate their severity. It is considered an intelligent, robust and cost effective system. But having such equipment of this system may be practical only in wealthy developed communities.

Olutayo & Eludire, (2014) applied ANN and Decision Trees techniques to learn new knowledge patterns hidden in a historical dataset related to traffic accidents of Nigeria. Their data were categorized into continuous and categorical data. ANN was used to analyze the continuous data and for the categorical data Decision Trees were used. They performed both Multilayer Perception (MLP) and Radical Basis Function Neural Networks (RBF) for ANN, whereas Id3 and Function Tree (FT) algorithms for decision trees using WEKA tool. According to their experimental results, the accuracy of the Id3 tree was better than FT, while RBF was better than MLP, which also observed that the decision tree outperformed ANN with a higher accuracy that was 77.70% in return 52.70%.

Effati et al., (2015) proposed a machine learning ML-based geospatial techniques. They integrated geospatial analysis, SVM and coactive Neuro-fuzzy inference system, developed by combining ANNs and fuzzy systems, (CANFIS) to discover spatial and non-spatial factors that involved in prediction severity of crash dataset stored over a 4-year in Iran. The authors used three software tools, Tanagra tool was used for SVM, the Neurodimension's NeuroSolutions software and MATLAB were used to train the CANFIS. SVM (with RBF kernel) outperforms CANFIS (with a TSK fuzzy, the DeltaBarDelta learning rule and three Gaussian functions) with accuracy was 85.49% against 76.44 %.

A comprehensive geospatial approach was presented by Effati et al., (2014) the proposed method based on combining of fuzzy and decision tree (CART) in (FCART) to model the crash severity against real data set 4,957 events from (2002 to 2005) in Iran using Inference Engine Tool and Spatial Analysis tool. The SVM, CART, FCART, and Bagged-FCART methods were tested with accuracies 66.43%, 71.576%, 76.821% and 79.12% respectively to uncover the role of spatial and non-spatial factors. The proposed approach demonstrated that the spatial factors had a substantial effect on the severity of the injury, as well as non-spatial factors.

Mohamed, (2014) employed a multi-class Support Vector Machine (SVM) technique with Gaussian Radial Basis function to predict the causes of road traffic accidents based on 1000 real crashes were taken from the police department in Dubai city. The accuracy achieved by the SVM model was greater than 75%. The author performed some preprocessing methods in addition to Gain Ratio Attribute selection method using WEKA.

The SVM (RBF) with 85.49 % accuracy in (Effati et al., 2015) contrasts with SVM 66.43% accuracy appeared in (Effati et al., 2014) and with SVM (GRB) accuracy that was

about 75% in (Mohamed, 2014). The variation may be due to the difference in used datasets, tools, SVM functions and the type of features whether spatial or nonspatial.

Such variation in tools and datasets caused to the diversity in the CART 71.576% accuracy in (Effati et al., 2014) that contrasted with the CART accuracy that resulted in each of (Pakgohar et al., 2011) and (Beshah et al., 2011; Beshah et al., 2012; Beshah et al., 2013) with 81%, 84.50%, 93.52%, 84.50% respectively.

Gakis et al., (2014) presented automatic incident detection (AID) system based on SVM and also based on data retrieved from inductive-loop detectors. The purpose of the proposed approach was to choose the most efficient feature and parameters of SVM that resulted in the most accurate solutions for road accident detection. The CvSVM class of OpenCV used to implement the proposed feature selection method and SVM in C++. Furthermore, the speed attributes are only used in incident detection. This approach achieved one of the best results (MTTD = 0:46 min, or 27:6 sec) among relevant research according to the comparison performed by the authors.

Perone, (2015) employed SVM, LR, Random Forests (RF), kNN, and Naive Bayes to build prediction models to assess the injury severity using about 20.798 accident records of the city of Porto Alegre/RS (Brazil). The author utilized the framework Django GIS Brasil from the previous work of him. In addition, he used Pandas library to get data analysis and scikit-learn framework to execute the pre-processing process and build the predictors. According to AUC results, the LR and SVM satisfied the best scores that were 0.94 then RF with 0.93, kNN with 0.90 and Naïve Bayes with 0.83 scores. According to average Precision/Recall/F1-score, the LR and SVM also were very similar and the best with 0.89% F1-score followed by RF 0.88%, kNN with 0.85%, and Naïve Bayes with 0.43%. The dataset used in this study lacked detailed information about vehicles, drivers, and victims; such information may have a significant effect in experimental analysis. Moreover, the author did not apply feature selection techniques.

Liang, (2015) proposed an intelligent transportation system (ITS) based on Internet of Things (IoT) platform and cloud computing. The author introduced an improved SVM model by Ant Colony Algorithm (ACA) for automatic accident detection. The real historical data of Qingdao collected from IoT platform were used by the SVM model to forecast possible future incidents and give alerts. Moreover, data analysis was performed using MATLAB.

III. THE METHODOLOGY

The main objective of the proposed methodology is to build the prediction classification rules of the best performing model (Decision Tree, ANN, or SVM). This section explains the proposed research methodology to compare the performance of the decision tree, SVM, and ANN models and to use the most accurate predictive model.

A. Overall Research Design

Figure (1) presents the phases of the methodology used in this research. In the first phase, the dataset is preprocessed to select the most influential attributes. Then the classification using the Decision Tree, SVM, and ANN methods were used to build the predictive models. In the evaluation phase, the results of the experiments for the three classifiers are evaluated based on four evaluation measures (Accuracy, Precision, Recall, F1-Measure). In the last phase, the model which achieves the best results is used to build the classification model and build the knowledge-based system that could be used to predict the severity of the accident.

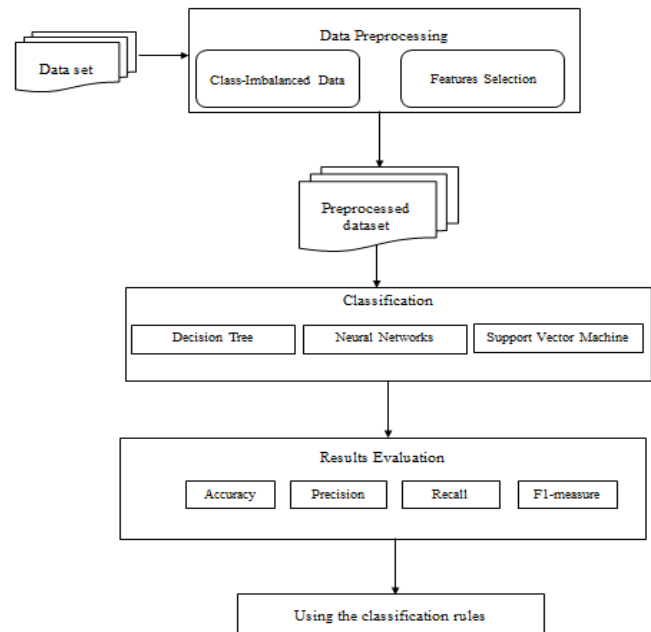


Fig. 1: The Design of the Methodology

B. Research Phases

The Dataset

The dataset used in this paper is the Road Accident and Safety data that is available at (<https://data.gov.uk/dataset/road-accidents-safety-data>) which was published by the Department for Transport of the United Kingdom in the year 2014. The dataset is related to environmental factors contains 49751 traffic accident records with 31 features in addition to 1 for the class label (Accident Severity). Table (1) shows dataset selected attributes. Table (2) shows the class label description.

Data Pre-Processing

Data pre-processing is an important stage for handling the data before using it in the data mining algorithms. This process involves various steps, including cleaning, normalization, feature selection, transformation. In this research, we need to apply Class-Imbalanced data solution and feature selection task on the dataset. The dataset used in the

research is preprocessed. The value -1 is exported for NULL or out of range values.

Table 1- Dataset Description

Feature Name	Feature Description and Values
Number of Vehicles	Number of Vehicles involved in an accident
Number of Casualties	Number of Casualties involved in an accident
Road Type	1: Roundabout 2: One-way street 3: Dual carriageway 4: Single carriageway 5: Slip Road
Speed limit	The Speed limitation of the road where the accident happened
Junction Control	1: Authorized person 2: Auto traffic signal 3: Stop sign 4: Give way or uncontrolled -1: Data missing or out of range
Light Conditions	1: Daylight 2: Darkness - lights lit 3: Darkness - lights unlit 4: Darkness - no lighting 5: Darkness - lighting unknown
Road Surface Conditions	1: Dry 2: Wet or damp 3: Snow 4: Frost or ice 5: Flood over 3cm. deep 6: Oil or diesel 7: Mud 8: Data missing or out of range
Weather Conditions	1: Fine no high winds 2: Raining no high winds 3: Snowing no high winds 4: Fine and high winds 5: Raining and high winds 6: Snowing and high winds 7: Fog or mist 8: Other 9: Unknown
Urban or Rural Area	1: Urban 2: Rural

Table 2- Class Label Description

Accident Severity/ label	Code	Number of instances
Fatal	1	429
Serious	2	5859
Slight	3	43463

- *Class-Imbalanced Data*

Class imbalance problem is a serious issue in the classification. It is caused by the skewed distribution of data between classes. Most of the classification algorithms focus on the main samples and ignore or misclassify minority samples. The minority samples are those that very important but rarely occurred. Sampling techniques are used to handle the imbalanced dataset problem; sampling techniques involve re-sampling the imbalanced dataset and also known as a preprocessing method. Resampling techniques can be achieved by under-sampling the majority class, over-sampling the minority class, or a hybrid of over and under sampling

approaches (Longadge & Dongre, 2013).

Our dataset is imbalanced where the major samples for the Slight class while the minority samples for the classes Fatal and Serious. In such situation, most of the classifiers are biased towards the major classes and hence provide poor classification rates on minor classes. In addition, it is also possible that classifiers predict everything as a major class and ignore the minor class, such in our case where the SVM and ANN classifiers predicted all classes as Slight classes and misclassified the Fatal and Serious classes.

To handle this issue and to improve the classification accuracy of class-imbalanced data, we used re-sampling techniques, namely Under-sampling, Oversampling, and hybrid sampling. Under-sampling randomly decreases the number of samples from the majority class while Oversampling increases the number of samples from the minority class until there is an equal number of the majority and minority samples and the hybrid sampling decreases majority class and increases minority class at the same time. We applied the down-Sample, Up-Sample, and SMOTE functions for Under-sampling, Oversampling, and hybrid sampling respectively using the (R) tool.

- *Feature Selection*

Feature selection, also known as attribute selection or variable selection, is a process of selecting a subset of relevant features for using in model construction. The used dataset contains 31 features, in addition to 1 for the class label.

We used Information Gain and Gain Ratio measures to rank the attributes and determine the most useful attribute, and accordingly, we determined different thresholds for the number of the most influential attributes to be used in the experiments. Then the used algorithms were applied to the dataset with these selected features, and the accuracies of them were compared and repeated this process with the multi thresholds to obtain the highest accuracies.

Classification Using Data Mining Algorithms

After preprocessing step, Data Mining algorithms are performed on the dataset to find the best one in the prediction of the traffic accident severity by comparing the accuracies between them.

Data mining has various tasks such as classification and prediction, clustering, association rule mining. Classification techniques classify data into the predefined class label. Data classification is a two-step process. The first step is the learning phase where training data are analyzed to build a model (classification rules) that describes a predefined set of classes. The second phase is the classification phase where the accuracy of the model is estimated using test data. If the accuracy is considered acceptable, the model (rules) can be used to class new unlabeled data instances and also the model can be used in the decision-making process. There are different techniques for data classification such as decision tree, neural networks, naïve Bayes classifiers, support vector

machine.

1) Decision Tree: Decision tree classifiers are one of the most popular and used classification techniques because the tree is constructed from the given data based on simple equations and uses the attribute selection measures such as a gain ratio measure, which ranks the attributes and determines the most useful attribute, and accordingly the researcher can realize the most efficient attributes on the predicted purpose (Han et al., 2011).

The decision tree is one of the main data mining technique that is used to build the classification model, it is a very practical method since it is relatively fast, does not require any domain knowledge or parameter setting, can deal multidimensional data, and can easily generate a set of simple classification rules that are interpretable and understandable for humans. In general, decision tree classifiers have good accuracy. Some of the decision tree classifiers are ID3, C4.5 / C5.0 / J48, CART, Random Tree, and other (Han et al., 2011).

2) Neural Networks: Artificial Neural Networks (ANN) is known as a powerful data modeling tool in prediction and classification. There are several kinds of ANN technique, but the most used are the back-propagation network. Advantages of ANN, however, include their flexibility when dealing with missing and noisy data. In addition to their ability to classify and deal with untrained and complex patterns, so they can be used with little knowledge of the relationships among attributes and classes. Neural networks are well suited for continuous-valued and are an inherently parallel technique that can be used to speed up the computation procedure. Furthermore, the ability to extract rules from trained neural networks makes neural networks very useful for classification and prediction in data mining. Moreover, they have been criticized for their poor interpretability (Han et al., 2011).

3) Support Vector Machine: A support vector machine SVM is a classification and prediction algorithm of both linear and nonlinear data that was suggested by Vapnik in 1960. Support vector machine has several advantages that distinguish them from other techniques. Although the training time of the SVMs can be extremely slow, they are highly accurate, owing to their ability to handle and model complex data structures. In addition, their perfect performance on data sets that have a vast number of features and also its ease of training (Han et al., 2011; Effati et al., 2014).

Performance measurement

In this step, the evaluation of the performance of the classification decision Tree, SVM, and ANN algorithms is performed and compared. Accuracy, precision, recall and F1-measure measures were used in the evaluation process.

C. Tool and implementation

For the implementation phase, R and WEKA toolkits were used to implement the preprocessing step and also to apply the Decision Tree (RandomForest, RandomTree, J48/C4.5, and CART), Artificial Neural Networks, and Support Vector

Machine algorithms to build the models.

IV. EXPERIMENTS AND RESULTS EVALUATION

This section presents and discusses the experiments and the results for the three different classifiers Decision tree (Random Forest, Random Tree, J48/C4.5, and CART), ANN (back-propagation), and SVM (polynomial kernel) with R and WEKA tools. Different comparisons and analysis were discussed in this section to see which of the three approaches provide better performance on prediction traffic accident severity. Accuracy, precision, recall and F1-measure measures were used in comparisons. Section 4.1 presents the dataset splitting method and Section 4.2 shows the result of the algorithms using R and WEKA tools. Section 4.3 presents the Discussion of the results. Section 4.4 presents examples of the prediction classification rules from the most accurate model.

A. Dataset Splitting

In section 3 we explained the preprocessing steps for our dataset, after finishing preprocessing we need to split this dataset into a Training dataset and Testing dataset, different ways are used to do This, this research used 10-fold Cross Validation and Holdout (training data 66% and testing data 34%) method during the experiments.

B. Experiments and Results

The following results of our 16 experiments are for the Decision Tree, SVM, and ANN experiments on a different number of features using WEKA tool, and Up-sample, Down-Sample, and hybrid techniques using R.

1) Decision Tree: The highest results of the used classifiers (CART, Random Forest, J48, and Random Tree) using Down-Sample, Up-Sample, Hybrid datasets were summarized in Tables (3), (4), and (5) respectively, while the overall highest results of the classifiers were summarized in Table (6).

Table 3-The highest performance results for Decision Trees with Down-Sample.

Classifier	Evaluation Method	Number of Features	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
CART	10-fold CV	6	49.18	48.2	49.2	48.4
RandomForest	10-fold CV	5	49.1	49.1	49.1	49.0
J48	Holdout	7	49.03	49.3	50.0	49.4
RandomTree	10-fold CV	4	48.87	47.3	48.9	47.3

Table 4-The highest performance results for Decision Trees with UpSample.

Classifier	Evaluation Method	Number of Features	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
RandomForest	Holdout	9	63.19	62.7	63.2	62.9
RandomTree	Holdout	9	63.0	62.6	63.0	62.7
CART	10-fold CV	9	62.48	62.0	62.5	62.1
J48	10-fold CV	9	62.22	61.7	62.2	61.8

Table 5-The highest performance results for Decision Trees with Hybrid.

Classifier	Evaluation Method	Number of Features	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
RandomForest	Holdout	9	80.65	81.4	80.6	80.1
RandomTree	Holdout	9	79.28	79.8	79.3	78.7
J48	10-fold CV	9	76.77	77.4	76.8	76.0
CART	10-fold CV	9	76.68	77.0	76.7	76.0

Table 6-The overall highest results of performance measures for Decision Trees.

Sampling	Classifier	Evaluation Method	Number of Features	Accuracy (%)	Pr. (%)	R (%)	F-M (%)
Under Sampling	CART	10-fold CV	6	49.18	48.2	49.2	48.4
Over-Sampling	Random Forest	Holdout	9	63.19	62.7	63.2	62.9
Hybrid	Random Forest	Holdout	9	80.65	81.4	80.6	80.1

2) **ANN**: The highest results of the ANN classifier using Down-Sample, Up-Sample, and Hybrid datasets were summarized in Tables (7).

Table 7- The overall highest results of performance measures for ANN.

Sampling	Evaluation Method	Number of Features	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
Under-Sampling	10-fold CV	7	48.09	48.4	48.1	48.0
Over-Sampling	Holdout	9	50.29	51.2	50.3	50.4
Hybrid	10-fold CV	9	61.44	59.7	61.4	59.0

3) **SVM**: The highest results of the SVM classifier using Down-Sample, Up-Sample, and Hybrid datasets were summarized in Tables (8).

Table 8- The overall highest results of performance measures for SVM.

Sampling	Evaluation Method	# Features	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)
Under-Sampling	Holdout	9	47.49	47.2	47.5	46.9
Oversampling	10-fold CV	7	46.63	45.2	46.6	44.2
Hybrid	10-fold CV	9	54.84	49.2	54.8	48.7

The overall highest results of the used classifiers were summarized in Table (9). The Decision Tree gets the highest performance than ANN and SVM.

Table 9- The overall highest results of performance measures for each classifier.

Classifier	Sampling	Eval Method	# Features	Acc (%)	Pr (%)	R (%)	F-M (%)
Random Forest	Hybrid	Holdout	9	80.65	81.4	80.6	80.1
ANN	Hybrid	10-fold CV	9	61.44	59.7	61.4	59.0
SVM	Hybrid	10-fold CV	9	54.84	49.2	54.8	48.7

4) Results before Sampling

The following results are for the Decision tree-based methods (Random Forest, Random Tree, J48/C4.5, and CART), SVM, and ANN experiments using the original dataset before sampling, where most of the classifiers are biased towards the major classes and hence provide poor classification rates on minor classes. In addition, it is also possible that classifiers predict everything as a major class and ignore the minor class; such in our case where the Fatal and Serious classes were misclassified, and all classes were classified as Slight.

Tables (10 - 15) show the bias and the poor classification rates for the classifiers using the original dataset. These results came in this way because the original dataset is an imbalanced dataset where the number of minor (Fatal and Serious) classes was 429 and 5859 respectively, while the number of the major (Slight) class was 43463.

Table 10- RandomForest classifier results using the original dataset with 10-fold CV.

Class	Precision (%)	Recall (%)	F-Measure (%)
Fatal	5.9	0.7	1.2
Serious	20.4	2.0	3.6
Slight	87.5	99.0	92.9
Average	78.9	86.7	81.6
Accuracy (%)	86.68		

Table 11- RandomTree classifier results using the original dataset with 10-fold CV.

Class	Precision (%)	Recall (%)	F-Measure (%)
Fatal	5.9	1.6	2.6
Serious	16.6	1.8	3.3
Slight	87.5	98.6	92.7
Average	78.5	86.4	81.4
Accuracy (%)	86.39		

Table 12- J48 classifier results using the original dataset with 10-fold CV.

Class	Precision (%)	Recall (%)	F-Measure (%)
Fatal	0.0	0.0	0.0
Serious	0.0	0.0	0.0
Slight	87.4	0.01	93.3
Average	76.3	81.5	81.5
Accuracy (%)	87.36		

Table 13- CART classifier results using the original dataset with 10-fold CV.

Class	Precision (%)	Recall (%)	F-Measure (%)
Fatal	0.0	0.0	0.0
Serious	0.0	0.0	0.0
Slight	87.4	0.01	93.3
Average	76.3	81.5	81.5
Accuracy (%)	87.36		

Table 14- ANN classifier results using the original dataset with 10-fold CV.

Class	Precision (%)	Recall (%)	F-Measure (%)
Fatal	0.0	0.0	0.0
Serious	0.0	27.3	0.1

Slight	87.4	0.01	93.2
Average	76.3	81.5	81.5
Accuracy (%)	87.351		

Table 15- SVM classifier results using the original dataset with 10-fold CV.

Class	Precision (%)	Recall (%)	F-Measure (%)
Fatal	0.0	0.0	0.0
Serious	0.0	0.0	0.0
Slight	87.4	0.01	93.3
Average	76.3	81.5	81.5
Accuracy (%)	87.36		

C. Results Discussion

As noticed in the results presented in the previous section, the results show that the highest Accuracy, Precision, Recall, and F-Measure were 80.650%, 0.814%, 0.806%, and 0.801% respectively to Decision Tree (Random Forest) followed by 61.445%, 0.597%, 0.614%, and 0.590% respectively to ANN then by 54.843%, 0.492%, 0.548%, and 0.487% respectively to SVM. All these results were achieved on the Hybrid sampling dataset and with Holdout method for Random Forest and 10-folds Cross Validation for each ANN and SVM.

The obtained results were agreed with the results of (Olutayo & Eludire, 2014) where the accuracy of the Decision Tree (ID3) with 77.70% outperformed the accuracy of the ANN that was 52.70%, and also agreed with (Effati et al., 2014) where the Decision Tree (CART) was more superior than SVM with the accuracy results is 71.576% and 66.43%. Furthermore, the outcomes contrast the results of (Perone, 2015) where the SVM outperformed the Decision Tree (Random Forest).

The accuracies on the Under-Sampling dataset for the Decision Trees Classifiers CART, Random Forest, J48, and Random Tree were 49.184%, 49.106%, 49.029%, and 48.873% respectively. As noticed, the CART classifier outperformed others.

The accuracies on the Oversampling dataset for the Decision Trees Classifiers Random Forest, Random Tree, CART, and J48 were 63.189%, 63.004%, 62.480%, and 62.222% respectively. It can be noticed that the Random Forest classifier outperformed others.

The accuracies on the Hybrid Sampling dataset for the Decision Trees Classifiers Random Forest, Random Tree, J48, and CART were 80.650%, 79.280%, 76.777%, and 76.684% respectively. It can be noticed that the Random Forest classifier outperformed others classifiers.

The results of the Decision Trees were agreed with the results of (Beshah et al., 2012) where the accuracy of the CART with 93.52% outperformed the accuracy of the Random Forest that was 90.75%, and also agreed with (Beshah et al., 2013) where the Random Forest was more superior than CART with the Accuracy results of 86.59% and 84.50%. Moreover, our outcomes agreed with the results of (Krishnaveni & Hemalatha, 2011) where the Random Forest was the most superior among used classifiers such as J48.

The highest results of Trees Classifiers were obtained on the Hybrid Sampling dataset, followed by Oversampling then by

Under-Sampling and the Random Forest classifier was the most accurate one.

The best result of ANN was on the Hybrid dataset, and the results were 61.445%, 50.293%, and 48.096% on Hybrid, Oversampling, and Under-Sampling datasets respectively.

The best result of SVM was on the Hybrid dataset, and the results were 54.843%, 47.489%, and 46.631% on Hybrid, Under-Sampling, and Over-Sampling datasets respectively.

The Accuracy rate on the original dataset of the used classifiers Decision tree (Random Forest, Random Tree, J48/C4.5, and CART), SVM, and ANN was high, and the Accuracies for classifiers were very close. SVM, ANN, J48, and CART achieved the same result, 87.361%, while 86.682% and 86.394% were for Random Forest and Random Tree respectively. In addition, the results of Precision, Recall, and F-Measure were poor for minor classes (Fatal and Serious) and bias to the major class (Slight) this due to the skewed distribution of data between classes on an imbalanced original dataset.

D. Generating the Decision Rules

To represent the knowledge and to identify significant and accurate rules, the PART algorithm was used. It adopts the separate-and-conquer strategy in that it builds a rule, eliminates the samples it covers and continues creating rules recursively for the remaining samples until no samples remain. This ensures that each sample of the training set is covered by at least one rule (Frank & Witten, 1998).

PART was run with the accuracy of 76.570% on the traffic accident dataset with 9 features on the Hybrid sampling dataset, and Cross Validation 10-folds were used. PART classifier generated 280 rules. A sample of the PART Prediction classification rules is shown in Table (16).

A. The Traffic Accident Prediction System in Action

To put the prediction model in action, a knowledge-based decision system is built. The PART algorithm was used to represent the knowledge and to identify significant rules. PART was run on the Traffic Accident dataset with different numbers of attributes. Rules were generated based on the following attributes: Urban or Rural Area, Speed limit, Light Conditions, and Number of Vehicles. The implementation of this system for Traffic Accident prediction is accomplished using JAVA language. Figure (2) shows sample screen shots for the Traffic Accident prediction system.

V. CONCLUSIONS AND FUTURE WORK

This paper provided a good review of literature in the field of Data Mining in Traffic Accidents area particularly, smart cities. In this research, three classification algorithms were implemented Decision tree (Random Forest, Random Tree, J48/C4.5, and CART), ANN (back-propagation), and SVM (polynomial kernel) to detect the influential environmental features of RTAs that can be used to build the prediction classification rules. These classifiers were trained and tested using the dataset was obtained from the Department for Transport of United Kingdom using WEKA tool. R tool was

used to apply sampling techniques to handle the imbalanced data problem of the used dataset. The experiment results show that the highest Accuracy, Precision, Recall, and F-Measure values were 80.650%, 0.814%, 0.806%, and 0.801% respectively to Decision Tree (Random Forest) followed by 61.445%, 0.597%, 0.614%, and 0.590% respectively to ANN then by 54.843%, 0.492%, 0.548%, and 0.487% respectively to SVM. All these results were achieved on the Hybrid sampling dataset and with Holdout method for Random Forest and 10-folds Cross Validation for each ANN and SVM.

The PART algorithm was used to present the knowledge in the form of rules. PART was run with the accuracy of 76.570% on the Traffic Accident dataset, and Cross Validation 10-folds were used. PART classifier generated 280 rules. Moreover, the JAVA language was used to build PART rules list for the prediction model. Rules were generated based on the following attributes: Urban or Rural Area, Speed limit, Light Conditions, and Number of Vehicles.

Due to the high rate of Road Traffic Accident in the Jordan and because that the RTAs in Jordan are the main reason of fatality, we are looking to collect traffic accident dataset and to detect the influential features of RTAs. Moreover, we are seeking to build a prediction Traffic Accident Severity system in the Jordan.

Table 16: Examples of PART Prediction classification rules.

Rule No	Classification Rules	# of Instances
1	IF Weather_Conditions <= 1 AND Speed_limit > 60 AND Number_of_Vehicles > 2 AND Road_Type > 4: THEN Fatal	82
2	IF Light_Conditions <= 6 AND Road_Type <= 3 AND Road_Surface_Conditions <= 2: THEN Fatal	59
3	IF Light_Conditions <= 6 AND Weather_Conditions <= 1 AND Number_of_Vehicles > 1 AND Urban_or_Rural_Area > 1 AND Speed_limit > 55 AND Number_of_Casualties <= 2: THEN Fatal	27
4	IF Junction_Control > 3: THEN Fatal	7
5	IF Road_Surface_Conditions <= 1 AND Road_Type == 3: THEN Serious	39
6	IF Speed_limit <= 30 AND Number_of_Vehicles > 2 AND Light_Conditions > 4 AND Weather_Conditions <= 2: THEN Serious	41
7	IF Junction_Control <= 1 AND Road_Surface_Conditions > 1 AND Speed_limit <= 55 AND Urban_or_Rural_Area == 1 AND Number_of_Vehicles <= 2 AND Weather_Conditions <= 1: THEN Serious	74
8	IF Speed_limit <= 40 AND Number_of_Vehicles == 2 AND Urban_or_Rural_Area == 1 AND Junction_Control > 2 AND Weather_Conditions <= 1 AND Light_Conditions <= 1 AND Number_of_Casualties <= 1 AND	1059

	Road_Surface_Conditions <= 1 AND Road_Type > 2: THEN Slight	
9	IF Speed_limit <= 30 AND Number_of_Vehicles == 3 AND Road_Type > 3 AND Light_Conditions <= 1 AND Junction_Control > 3 : THEN Slight	126
10	IF Speed_limit <= 30 AND Road_Surface_Conditions == 1 AND Number_of_Vehicles <= 2 AND Urban_or_Rural_Area == 1 AND Light_Conditions <= 4 AND Junction_Control > 2 : THEN Slight	677
11	IF Speed_limit <= 50 AND Number_of_Vehicles == 2 AND Urban_or_Rural_Area == 1 AND Road_Type <= 1 AND Weather_Conditions <= 1 AND Light_Conditions <= 1: THEN Slight	115

Prediction Traffic Accident Severity

Accident Severity : Fatal

Urban or Rural Area: Rural

Speed limit: Equal or More Than 70

Number of Vehicles: 5

Light Conditions: Daylight

Submit

Prediction Traffic Accident Severity

Accident Severity : Serious

Urban or Rural Area: Urban

Speed limit: 30

Number of Vehicles: 2

Light Conditions: Darkness - lights unit

Submit

Fig. 2: Screen shot for the prediction system.

REFERENCES

- [1] Abd-Elkawy, A.A.M., (2013) Application of Smart Community Indicators in New Egyptian Communities A Case Study: Smart Village, Greater Cairo, Egypt. *International Journal of Environmental Sciences*, 3(3), pp.143-154.
- [2] Al-Masaedi, H.R., (2009). Traffic accidents in Jordan. *Jordan Journal of Civil Engineering*, 3(4), pp.331-343.
- [3] Al-Zubi, A. S. A., (2010). Analysis of Vehicles Accidents in Amman City Using Spatial Data Mining and Visualization (Doctoral dissertation, The University Of Jordan).
- [4] Anastasi, G., Antonelli, M., Bechini, A., Brienza, S., D'Andrea, E., De Guglielmo, D., Ducange, P., Lazzarini, B., Marcelloni, F. and Segatori, A., (2013), Urban and Social sensing for sustainable mobility in smart cities. In *Sustainable Internet and ICT for Sustainability (SustainIT)*, (pp. 1-4). IEEE.

- [5] Beshah, T. and Hill, S., (2010), Mining Road Traffic Accident Data to Improve Safety: Role of Road-Related Factors on Accident Severity in Ethiopia. In *AAAI Spring Symposium: Artificial Intelligence for Development*.
- [6] Beshah, T., Ejigu, D., Abraham, A., Krömer, P., and Snásel, V., (2012). Knowledge discovery from road traffic accident data in Ethiopia: Data quality, ensembling, and trend analysis for improving road safety. *Neural Network World*, 22(3), p.215.
- [7] Beshah, T., Ejigu, D., Abraham, A., Snasel, V. and Kromer, P., (2011), Pattern recognition and knowledge discovery from road traffic accident data in Ethiopia: Implications for improving road safety. In *Information and Communication Technologies (WICT), 2011 World Congress on* (pp. 1241-1246). IEEE.
- [8] Beshah, T., Ejigu, D., Abraham, A., Snasel, V. and Kromer, P., (2013). Mining Pattern from Road Accident Data: Role of Road User's Behaviour and Implications for improving road safety. *International Journal of Tomography and Simulation*, 22(1), pp.73-86.
- [9] Chang, H.W., Tai, Y.C. and Hsu, J.Y.J., (2009). Context-aware taxi demand hotspots prediction. *International Journal of Business Intelligence and Data Mining*, 5(1), pp.3-18.
- [10] D'Andrea, E., Ducange, P., Lazzarini, B. and Marcelloni, F., (2015). Real-time detection of traffic from Twitter stream analysis. *Intelligent Transportation Systems, IEEE Transactions on*, 16(4), pp.2269-2283.
- [11] Devi, M.R.S., Kesavan, M.V.T. and Gayathri, M.V., (2015). Traffic Accident Classification and Automatic Notification Using GPS. *International Journal*, 13.
- [12] Effati, M., and Sadeghi-Niaraki, A., (2015). A semantic-based classification and regression tree approach for modeling complex spatial rules in motor vehicle crashes domain. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5(4), pp.181-194.
- [13] Effati, M., Rajabi, M.A., Hakimpour, F. and Shabani, S., (2014). Prediction of crash severity on two-lane, two-way roads based on fuzzy classification and regression tree using geospatial analysis. *Journal of Computing in Civil Engineering*, 29(6), p.04014099.
- [14] Effati, M., Thill, J.C. and Shabani, S., (2015). Geospatial and machine learning techniques for wicked social science problems: analysis of crash severity on a regional highway corridor. *Journal of Geographical Systems*, 17(2), pp.107-135.
- [15] El Tayeb, A. A., Pareek, V., Araar, A., (2015). Applying Association Rules Mining Algorithms for Traffic Accidents in Dubai. *International Journal of Soft Computing and Engineering (IJSC)*, 5(4).
- [16] Frank, E. and Witten, I.H., 1998, July. Generating accurate rule sets without global optimization. In *ICML* (Vol. 98, pp. 144-151).
- [17] Gakis, E., Kehagias, D. and Tzouvaras, D., (2014), October. Mining traffic data for road incidents detection. In *Intelligent Transportation Systems (ITSC), 2014 IEEE 17th International Conference on* (pp. 930-935). IEEE.
- [18] Giannotti, F., Nanni, M., Pedreschi, D., Pinelli, F., Renso, C., Rinzivillo, S., and Trasarti, R., (2011). Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal—The International Journal on Very Large Data Bases*, 20(5), pp.695-719.
- [19] http://census.dos.gov.jo/wp-content/uploads/sites/2/2016/02/Census_results_2016.pdf. Jordanian Department of Statistics.
- [20] <http://www.psd.gov.jo/images/traffic/docs/derasah2014.pdf>. Jordanian Traffic Department.
- [21] <https://data.gov.uk/dataset/road-accidents-safety-data>
- [22] Ibrahim, H. and Far, B.H., (2014), August. Data-oriented intelligent transportation systems. In *Information Reuse and Integration (IRI), IEEE 15th International Conference on* (pp. 322-329). IEEE.
- [23] Jadaan, K.S., Al-Fayyad, M. and Gammoh, H.F., (2014). Prediction of Road Traffic Accidents in Jordan using Artificial Neural Network (ANN). *Journal of Traffic and Logistics Engineering*, 2(2).
- [24] Jensen, M., Gutierrez, J. and Pedersen, J., (2014). Location Intelligence Application in Digital Data Activity Dimensioning in Smart Cities. *Procedia Computer Science*, 36, pp. 418-424.
- [25] Han, J. and Pei, J., (2011). Data Mining: Concepts and Techniques: Concepts and Techniques.
- [26] Krishnaveni, S., and Hemalatha, M., (2011). A perspective analysis of traffic accident using data mining techniques. *International Journal of Computer Applications*, 23(7), pp.40-48.
- [27] Kumar, A. and Kannathasan, N., (2011). A survey on data mining and pattern recognition techniques for soil data mining. *IJCSI International Journal of Computer Science Issues*, 8(3).
- [28] Kunt, M.M., Aghayan, I. and Noii, N., (2011). Prediction for traffic accident severity: comparing the artificial neural network, genetic algorithm, combined genetic algorithm and pattern search methods. *Transport*, 26(4), pp.353-366.
- [29] Lécué, F., Tallevi-Diotallevi, S., Hayes, J., Tucker, R., Bicer, V., Sbodio, M. and Tommasi, P., (2014). Smart traffic analytics in the semantic web with STAR-CITY: scenarios, system, and lessons learned in Dublin City. *Web Semantics: Science, Services, and Agents on the World Wide Web*, 27, pp.26-33.
- [30] Liang, G., (2015). Automatic Traffic Accident Detection Based on the Internet of Things and Support Vector Machine. *International Journal of Smart Home*, 9(4), pp.97-106.
- [31] Liu, S., Liu, Y., Ni, L.M., Fan, J. and Li, M., (2010). Towards mobility-based clustering. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 919-928). ACM.
- [32] Longadge, R. and Dongre, S., (2013). Class Imbalance Problem in Data Mining Review. *arXiv preprint arXiv:1305.1707*.
- [33] Madakam, S. and Ramaswamy, R., (2015), February. 100 New smart cities (India's smart vision). In *Information Technology: Towards New Smart World (NSITNSW), 2015 5th National Symposium on* (pp. 1-6). IEEE.
- [34] Mohamed, E.A., (2014). Predicting Causes of Traffic Road Accidents Using Multi-class Support Vector Machines. In *Proceedings of the International Conference on Data Mining (DMIN)* (p. 1).
- [35] Mulligan, C.E., and Olsson, M., (2013). Architectural implications of smart city business models: an evolutionary perspective. *Communications Magazine, IEEE*, 51(6), pp.80-85.
- [36] Obaidat, M.T., and Ramadan, T.M., (2012). Traffic accidents at hazardous locations of urban roads. *Jordan Journal of Civil Engineering*, 6(4), pp.436-447.
- [37] Olutayo, V.A., and Eludire, A.A., (2014). Traffic Accident Analysis Using Decision Trees and Neural Networks. *International Journal of Information Technology and Computer Science (IJITCS)*, 6(2), p.22.
- [38] Ong, R., Pinelli, F., Trasarti, R., Nanni, M., Renso, C., Rinzivillo, S. and Giannotti, F., (2011). Traffic jams detection using flock mining. In *Machine Learning and Knowledge Discovery in Databases* (pp. 650-653). Springer Berlin Heidelberg.
- [39] Pakgohar, A., Tabrizi, R.S., Khalili, M., and Esmaeili, A., (2011). The role of human factor in incidence and severity of road crashes based on the CART and LR regression: a data mining approach. *Procedia Computer Science*, 3, pp.764-769.
- [40] Perone, C.S., (2015). Injury risk prediction for traffic accidents in Porto Alegre/RS, Brazil. *arXiv preprint arXiv:1502.00245*.
- [41] Raut, S. and Karmore, S., (2015), March. Review on: Severity estimation unit of automotive accident. In *Computer Engineering and Applications (ICACEA), 2015 International Conference on Advances in* (pp. 523-526). IEEE.
- [42] Reby, D., Lek, S., Dimopoulos, I., Joachim, J., Lauga, J. and Aulagnier, S., (1997). Artificial neural networks as a classification method in the behavioral sciences. *Behavioral Processes*, 40(1), pp.35-43.
- [43] Shiau, Y.R., Tsai, C.H., Hung, Y.H. and Kuo, Y.T., (2015). The Application of Data Mining Technology to Build a Forecasting Model for Classification of Road Traffic Accidents. *Mathematical Problems in Engineering*, 2015.
- [44] Weiss, G.M., and Davison, B.D., (2010). Handbook of Technology Management.
- [45] Yin, C., Xiong, Z., Chen, H., Wang, J., Cooper, D. and David, B., (2015). A literature survey on smart cities. *Science China Information Sciences*, 58(10), pp.1-18.
- [46] Yousif, J.H., and AlRababaa, M.S., (2013). Neural Technique for Predicting Traffic Accidents in Jordan. *Journal of American Science*, 9(11).
- [47] Zhang, J., Wang, F.Y., Wang, K., Lin, W.H., Xu, X. and Chen, C., (2011). Data-driven intelligent transportation systems: A survey. *Intelligent Transportation Systems, IEEE Transactions on*, 12(4), pp.1624-1639.