

CRASH DATA REPORTING SYSTEMS IN FOURTEEN ARAB COUNTRIES: CHALLENGES AND IMPROVEMENT

Zahira ABOUNOAS¹, Wassim RAPHAEL², Yarob BADR³, Rafic FADDOUL⁴ Anne GUILLAUME⁵

^{1, 2, 3, 4} University-Saint Joseph (USJ), Faculty of engineering (ESIB), Beirut, Lebanon

³ United Nations Economic & Social Commission for West Asia (ESCWA), Beirut, Lebanon

⁵ Renault, LAB Laboratory for Accident Analysis, Paris, France

Abstract:

Traffic crash fatalities and serious injuries still represent a big burden for most Arab countries because the actual policies, strategies, and interventions are based on poorly collected data. Through this paper, we assessed the crash data reporting systems in Fourteen Arab countries via a survey conducted to identify the fundamental dysfunctions at the management and data collection levels. Then, to address some of the dataset problems, we had applied data mining technics to select a minimum of variables (crash, vehicle, and road user) that should be collected for a better understanding of crash circumstances. For this reason, three methods of selection (correlation, information gain, and gain ratio) and seven classifiers (naïve Bayes, nearest neighbour, random forest, random tree, J48, reduced error pruning tree, and bagging) were tested and compared to identify the variables that affect significantly the crashes severity. Decision trees family of classifiers showed the best performance based on the analysis of the area under the curve. The explanatory variables obtained from the data mining process were combined with other descriptive variables to maintain traceability. As a result, we produced hybrid lists of variables for the crash, vehicle, and road user, each contains 25 variables. Finally, in order to propose a cost-effective solution to switch from manual to electronic data collection, we got inspired by a tool used to track animals to create and customize a unified e-form for handheld devices, in order to ensure easy entering of the harmonized data for the entire region based on our selected lists of variables. The tool verified the countries requirements especially by enabling data collection and transfer with and without the internet, and by allowing data analysis through its built-in Geographic Information System (GIS) capabilities.

Keywords: road safety, information system, reporting system, variables selection, classification model.

To cite this article:

Abounoas, Z., Raphael, W., Badr, Y., Faddoul, R., Guillaume, A., 2020. Crash data reporting systems in fourteen Arab countries: challenges and improvement. *Archives of Transport*, 56(4), 73-88. DOI: <https://doi.org/10.5604/01.3001.0014.5628>



Contact:

1) zahira.abounoas@gmail.com, [<https://orcid.org/0000-0003-1920-0945>] - (corresponding author)

1. Introduction

In March 2010, the United Nations proclaimed during the general assembly the first decade of action for road safety, aiming to reduce 50% of road crash fatalities by 2020. However, in the Arab region instead of decreasing, road crash fatalities increased by + 4% (WHO estimated data) and +2% (officially reported data) between 2010 and 2016 (World Health Organization report, 2013; World Health Organization et al., 2018)¹.

The precondition to achieve our target is understanding the causes, circumstances and the location of crashes (Żukowska, 2015), which explain the importance of the collected crash data. However, based on the old principal of garbage in garbage out (GIGO) (Beasley, 2020), the quality of input data (collected crash data) affects directly the outputs (policies, strategies, interventions, and targets). Preliminary work was carried out affirmed that the quality of police crash data in the Arab region is questionable in terms of credibility, accuracy and completeness, especially because it is collected through non reliable manual reporting systems. Almatawah expressed the concerns in gulf cooperation council (GCC) countries (Oman, Kuwait, Bahrain, SA, UAE and Qatar) about the manual process of data collection which could lead to missing and incoherent data particularly for crash locations (Almatawah, 2014). In Morocco, road crash data is collected manually and computerized system appropriate to the systemic approach does not yet exist (Laaraj et al., 2018). In Egypt, due to the lack of an accurate reporting database, it is widely believed that the reported numbers are higher in reality (Khallaf & Yasseen, 2016). In Jordan, there is no one standard method for data collection (Dababneh et al., 2018). The Palestinian model relies entirely on text data, which leads to difficulties in data extraction and computerization. It also leads to the lengthy and error prone data entry (Sarraj, 2016). Besides completeness and accuracy problems, comparability issues appear between the different countries given the differences among the datasets in terms of collected variables, values and definitions. As illustrated, studies treating a single Arab country or a subset of countries do exist, but none focuses on the entire Arab region to extract

similarities and differences. Some of these studies are outdated.

To overcome similar data issues various crash datasets have been proposed all around the world, some are more exhaustive than the others. In Australia, each state has developed its own database and crash database system. In order to bring greater uniformity, state and territory road authorities agreed to work toward the implementation of common protocols for the collection of the fatality data, which enabled the establishment of a reliable national road fatality database (Montella et al., 2017). In the UK, the police reporting is done in an identical way across Great Britain (including England, Scotland, Wales and Northern Ireland) using the Stats19, which collects an accident record with (26 variables), vehicle records (22 variables), casualty records (14 variables) and contributory factors (6 factors) (UK department of transport, 2013). In the Netherlands, about 40 characteristics are recorded of those crashes that are registered by the police (SWOV, 2016). In Sweden, STRADA (Swedish TRaffic Accident Data Acquisition) has been used for the official accident statistics in Sweden (Trafikverket Swedish Traffic administration, 2019). Since 2003 the Swedish statistics are based on data reported by two sources the police and the emergency hospitals, about 97 variables are collected in total (Howard & Linder, 2014). For all of Europe Safety-net project suggested a minimum set of standardized data elements called CADaS (Common Crash Data Set). CADaS contains 88 Attributes (12 for crash, 37 for road, 17 for traffic unit, 22 for person) (Jha et al., 2020). In the USA, a more exhaustive list of variables is collected via the Fatality Analysis Reporting System (FARS) is using the Model Minimum Uniform Crash Criteria (MMUCC) which is more extended dataset (68 for person, 52 for crash, 94 for vehicle) (Bellis, 2015). Therefore, a question arise: is it more efficient to stick with the minimum set of variables, to use an extended one, or to produce a hybrid version? How to select the most relevant variables?

In recent years, data mining selection and classification models proved to be efficient in detecting variables with the greatest influence on car

¹ Arab countries taken into consideration for the % calculation are the ESCWA countries reported in both 2013 and 2018 WHO global reports (World Health Organization, 2013; World Health Organization et al., 2018)

crashes (Castro & Kim, 2016) and particularly important for data sets with large numbers of features because large data becomes useless without proper utilization (Hussain et al., 2018). These techniques were used to identify the significant predictors explaining fatal road accidents (Dadashova et al., 2016). A recent paper had also used data mining to classify crash severity of various traffic crashes (Ramya et al., 2019).

The selection and creation of an optimal list of variables is necessary but not sufficient to improve crash data reporting systems (CDRS) when data is manually collected. Manual data entry, always incurs mistakes and inaccuracy especially for crash location. Electronic recording methods make data collection faster and less susceptible to transcription errors. Additionally, digital forms incorporating Global Positioning System (GPS) with the combined use of GIS overcomes traditional problems associated with crash location, such as inaccuracies and collection mistakes (Montella et al., 2017). Even for road safety risk factors, digital data collection is convenient and reliable during roadside observational data collection, the productivity is higher compared with paper-based method (Mehmood et al., 2019).

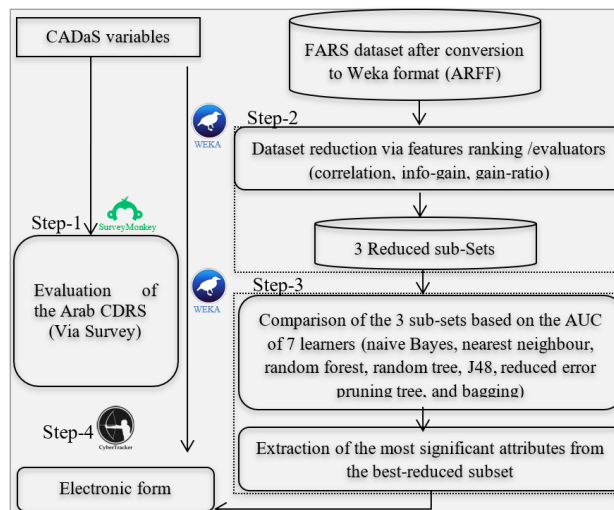
In brief, the main contributions of this paper are: a) To assess crash data reporting systems (CDRS) in 14 Arab countries. b) To propose remedies for some of

the identified problems, by applying data mining techniques to select a minimum set of variables (vehicle, crash, and user) that should be collected for an appropriate understanding of crash circumstances. c) To insert the selected variables in a unified customized e-form, in order to ensure easy data entry and harmonized data collection throughout the Arab region.

The geographic scope of the project encompasses all Arab countries members of UN-ESCWA. However, in this paper we will be focussing on the 14 countries which participated in the survey: Iraq (IQ), Jordan (JO), Kuwait (KW), Lebanon (LB), Mauritania (MR), Morocco (MA), Oman (OM), Palestine (PS), Qatar (QA), Saudi Arabia (SA), Sudan (SD), Syria (SY), Tunisia (TN) and Yamen (YE).

2. Methodologies

To achieve the objectives listed above, the CADaS list of variables is used in the survey to assess the variables collected by Arab CDRS (Step-1). Then, Dataset extracted from the Fatality Analysis Reporting System (FARS) for 2016 was tested to select explanatory variables for the severity level (dependent variable) (Steps 2 and 3). Finally in Step-4, variables from CADaS and FARS are combined in a customized e-form. Figure 1 summarises the entire process, noting that all used evaluators and learners will be explained later in this paper.



*ARFF: Attribute Relation File Format

Fig. 1. Holistic methodology summary

3. Material and methods

3.1. CDRS evaluation

In step-1, we assessed CDRS through a Survey-Monkey shared with the focal points of transport in 18 members' countries of UN-ESCWA. The survey contains 56 questions divided into six axes. In this paper the focus is given to road safety management, dataset (variables, values, and definitions) and data flow (from the collection to the storage). The dataset CADA's list of variables classified as High Important by Safety-net project.

3.2. The principle of selecting variables using Weka

Weka stands for Environment for Knowledge Learning. It is a data mining software written in Java. The tool was developed by the University of Waikato. Weka supports data mining tasks such as data pre-processing, clustering, classification, regression and feature selection. It was chosen because compared with R, Knime and RapidMiner, Weka needs less memory, work faster and it is provided with both GUI and CLI (Atnafu & Kaur, 2017). In this paper Weka is used for both variables selection and classification tests (in steps 2 & 3). First, we had used Weka to produce three reduced datasets, by removing irrelevant variables. The relevant variables were identified via three evaluators (Correlation, info Gain and Gain Ratio), then the attributes were ranked accordingly through a search option (Ranking). The main purpose of the selection step is to reduce the number of features to be collected by the crash investigator, which may reduce the effort and the storage requirement, but at the same time we want to ensure that the new selected subsets of variables are the most relevant ones.

Second, in order to compare the results and to get an idea of which reduced dataset had the best performance, we ran seven classifiers (NB, IBK, J48, RF, RT, RepT, and Bag) to find the best performance in the reduced subsets in terms of classifying severity. We compared the classifiers based on the AUC as a comparison and evaluation criteria.

3.3. Data selection algorithms

3.3.1. Correlation

It evaluates the worth of an attribute by measuring the Pearson's correlation coefficient between the

explanatory variable and the severity level. The main aim is to obtain a highly relevant features subset which contains features highly correlated with the severity class, but uncorrelated with each other (Kumar & Singh, 2016).

3.3.2. Gain-Ratio

It determines the value of attribute for selection by evaluating the gain with respect to the split information (Vinutha & Poornima, 2018). The attribute with the maximum gain ratio is selected as the splitting attribute.

$$\text{Gain Ratio (A)} = \frac{\text{Gain (A)}}{\text{SplitInfo (A)}} \quad (1)$$

Gain(A) tells us how much would be gained by branching on A. It is the expected reduction in the information requirement caused by knowing the value of A.

Where the split information value represents the potential information generated by splitting the training dataset D into v partitions corresponding to v outcomes on attribute A.

$$\text{SplitInfo}_A(D) = - \sum_{j=1}^v \frac{|D_j|}{|D|} \times \log_2 \left(\frac{|D_j|}{|D|} \right) \quad (2)$$

3.3.3. Info-Gain

This score reflects how much maximum information is obtained about the classes when a particular feature is used. In other words, the information gain is the amount by which the Shannon entropy of the class decreases; it reflects the additional information about the class provided by the attribute (Kumar & Singh, 2016).

If A is an attribute and C is the class, Eq.3 shows the IG equation:

$$\text{IG} = H(C) - H(C|A) \quad (3)$$

With the entropy of the class before (H(C)) and after observing the attribute (H(C|A))

$$H(C) = - \sum_{c \in C} p(c) \log_2 p(c) \quad (4)$$

$$H(C|A) = - \sum_{a \in A} p(a) \sum_{c \in C} p(c|a) \log_2 p(c|a) \quad (5)$$

3.4. Classifiers used to compare the reduced subsets

3.4.1. Naïve Bayes (NB)

NB is a simple probabilistic classifier based on applying Bayes theorem with strong (naive) independence assumptions. To classify a given instance $x = \{x_1, x_2, x_n\}$ (the n features are assumed independent), the model assigns to this instance the probabilities for each of k possible outcomes or classes C_k . (Theofilatos et al., 2019)

$$p(C_k|x) = \frac{p(C_k)p(x|C_k)}{p(x)} \quad (6)$$

3.4.2. K-nearest neighbours

k-nearest neighbor algorithm (kNN) is called Instance Based Learner (IBK) in Weka. It is a 'lazy learning' technique because little effort goes into building the classifier and most of the work is performed at the time of classification. An instance is classified by a plurality vote of its neighbors, with the instance being assigned to the class most common among its k nearest neighbors (Singh et al., 2017).

3.4.3. J48

J48 is the enhanced version of C4.5 decision tree (Kang & Michalak, 2018). The algorithm analyses the attribute list, divides the information in subset, identifies the attribute with most gain of information (which discriminates the various instances most clearly), and recognizes it as the decision parameter. Finally it classifies the information according to the decision parameter (Cuartas et al., 2015).

3.4.4. Random-Tree

It is a supervised Classifier. In standard tree each node is split using the best split among all variables. In a random forest, each node is split using the best among the subset of predictors randomly chosen at that node (Kalmegh, 2015).

3.4.5. Random-Forest

RF is an ensemble classifier that consists of many *RandomTrees*. It is created by taking a bunch of different samples of data and growing trees out of them (Belgiu & Drăguț, 2016). A random forest is like a black box that we can build and control. We can specify the number of trees we want in our forest; we can also specify the maximum number of features to be used in each tree. We cannot control

the randomness, but we can control which feature is part of which tree in the forest, or control which data point is part of which tree.

3.4.6. Rep-Tree

Basically Reduced Error Pruning Tree (REPT) is a fast decision tree learner which builds a decision tree using information gain as the splitting criterion, and prunes it using reduced error pruning (Kalmegh, 2015).

3.4.7. Bagging

Also called voting techniques which are used to combine the predictive power of multiple models in an attempt to surpass the performance of the best individual model (Zhang et al., 2019). Bagging plays an important role in the field of medical diagnosis.

3.5. The criteria of evaluation and comparison

The performance of classifiers is traditionally evaluated using the overall accuracy measure which is calculated using the following equations (Tharwat, 2018):

$$\text{Accuracy} = \frac{\text{Correct predictions}}{\text{Total number of examples}} \quad (7)$$

However with real data, in which data classes might have skewed distributions, imbalanced data sets, or unequal error, an accuracy comparison rule can perform poorly (Bhondave et al., 2014; S. Wang et al., 2015). The Area Under the Receiver Operator Characteristic (ROC) Curve denoted (AUC) has been proven to be a better performance metric in comparison with classification accuracy, it is now taken into account because it is more appropriate, that even the direct accuracy, for imbalanced data sets with different misclassification costs (Abellán & Castellano, 2017). AUC methodology was used for the first time in the context of electronic signal detection and problems with radar in the early 1950s. It is an evaluation metric that considers all possible classification thresholds to assess the capacity of the classifier to have a high sensitivity while not being excessively penalized by a decreasing specificity. Therefore, we will be using it to compare the results of the classifiers among the different datasets.

Suppose a data set consists of n samples, n_0 of them are from the positive class, and the remaining n_1 samples belong to the negative class. Each sample x_i

is associated with a score $s(x_i)$. To calculate AUC, these scores are first sorted in ascending order, then each of them is assigned with a rank starting from 1. After that, samples with the same score should be re-ranked by averaging the original ranks of them. Let r_1, r_2, \dots, r_{n_0} be the ranks of positive samples (R. Wang & Tang, 2009). AUC can be calculated using:

$$AUC = \frac{\sum_{i=1}^{n_0} (r_i - i)}{n_0 \times n_1} = \frac{\sum_{i=1}^{n_0} (r_i) - n_0 \frac{(n_0 + 1)}{2}}{n_0 \times n_1} \quad (8)$$

The following rules of thumb are used to evaluate the performance of classifiers using AUC (Wilson, 2018, p. 4):

$$\text{Classification Performance} = \begin{cases} \text{Not good, AUC} = 0.5 \\ \text{Poor, } 0.5 < \text{AUC} < 0.6 \\ \text{Fair, } 0.6 \leq \text{AUC} < 0.7 \\ \text{Acceptable, } 0.7 \leq \text{AUC} < 0.8 \\ \text{Excellent, } 0.8 \leq \text{AUC} < 0.9 \\ \text{Outstanding, AUC} \geq 0.9 \end{cases} \quad (9)$$

3.6. E-form for data collection (Cyber tracker)

In order to create a unified investigation e-form, we got inspired from tracking animals, and we had used *Cyber Tracker* instead of building a new application for handheld devices from scratch. The software was developed for the first time to track and monitor wildlife in South Africa (Liebenberg, 1999). Being an open source software, encouraged users in several fields to use it (Spanu & McCall, 2013), for example it was recently used in the Western European Shelf Pelagic Acoustic Survey to collect all positional, environmental and sightings data (O'Donnell et al., 2019). CyberTracker proved to be an efficient, cost-effective, user-friendly and versatile data collection and management tool (CyberTracker, 2020).

4. Discussion of the CDRS assessment results (Step 1)

To address the health burden of traffic crashes, 12 out of 16 Arab countries had set national strategies (three fully funded) (World Health Organization, 2018). However, even countries with fully funded national strategies did not showed a positive performance. Two principal factors might explain the absence of national strategies and the non-efficiency of the existing ones: a) dysfunctional national road-safety management system, and b)

dysfunctional data collection, which is the basis of national interventions.

As explained in the methods section, we had assessed these two types of dysfunctions through a survey shared with focal points of transport of 18 members' countries of ESCWA. Only 14 countries answered the questionnaire and the response rate differed from a question to another.

4.1. Safety management dysfunction

In order to understand how the system of traffic safety management works, we assessed four possible entities that might exist and contribute to road safety management in each country. Table 1 summarizes the assessment results for each entity by country.

First, the committee for traffic safety headed by 'the Prime Minister' and which might include the membership of the ministers addressing traffic safety. It meets periodically. Mainly it is focusing on formulating government policy and supervising the implementation of procedures according to the decisions taken. As a result seven countries do have committees, however most of them (5 out of the 7) do not involve ministry of communication which might explain why road safety issues are not promoted properly in the Arab region.


Second, the higher council composed of experts, representatives, companies, NGOs, and government departments. It meets several times per year, on specialized and specific topics. Its main mission is to formulate recommendations and proposals, support and carry out studies to improve knowledge about traffic safety, and evaluate the effectiveness of the implemented measures. As a result (8 out of 13) have a council.

Third, the national observatory which is acting as a secretary for the national traffic safety council is responsible for collecting, arranging, interpreting and disseminating statistical data related to traffic crashes, supervising and following up traffic crashes studies. As a result, only (2 out of 13) have a national observatory. The absence of such important entity may explain why the collected data is poor.

Fourth, a national lead agency dedicated to traffic safety management headed by the 'Minister of Transport or the Minister of Interior'. It is the entity which proposes and then supervises the implementation of the national traffic safety policy and ensures the seven institutional management

Table 1. Status of road safety Management

	National Committee	High Council	National Observatory	Lead agency (LA)	Score LA
Mauritania	Yes	Yes	No	No	NA
Yemen	No	No	No	Yes	1
Kuwait	No	Yes	No	No	NA
Qatar	Yes	Yes	Yes	Yes	6
Syria	Yes	Yes	No	No	NA
Jordan	No	No	No	No	NA
Iraq	No	Yes	No	No	NA
Lebanon	Yes	Yes	No	Yes	4
Morocco	Yes	Yes	No	Yes	3
Oman	No	No	No	No	NA
Palestine	Yes	Yes	No	No	4
Tunisia	No	No	No	Yes	3
Sudan	Yes	No	No	No	NA



Yes
No

functions defined by the world bank (Bliss & Breen, 2009): results focus, coordination, legislation, funding, promotion, monitoring and evaluation, knowledge transfer and producing the national strategy. A ranking score was given based on the number of management functions that the lead agency is doing. Notwithstanding the fact that some countries had a lead agency (4 out of 12), the ranking score was not promising. Qatar had the highest ranking (6 out of 8 functions are verified) and Yemen lowest one (1 out of 8 functions).

4.2. Data Dysfunctions

4.2.1. Possible existence of the electronic forms

As a result, only two countries do have an electronic form (Oman, Palestine), six countries have paper forms only (Sudan, Iraq, Tunisia, Syria, Morocco and Lebanon), and two countries do not have investigation forms (Kuwait, Sudan). Most of the Arab countries do not use digital forms to collect traffic crash data which affects data accuracy especially crash location.

4.2.2. Heterogeneous definitions for quantitative variables

Concepts definition are not homogenous, among countries. Hence, comparability between countries is undermined. The (Figure 2) pinpoints the definitions of severe injury and death used by each country.

4.2.3. Non-standardised values for qualitative variables

Using non-standardised values for some qualitative variables is another type of discrepancy that might make data merging and comparison more difficult, especially when each country uses its own categorization of values.

For the *collision type categorization*, collisions with vehicles hitting an obstacle are not taken into consideration in Palestine and Sudan. Sudan considers falling from a vehicle as collision type. Iraq and Saudi Arabia added a specific type of categorisation which is collision involving animals. Concerning *involved persons categorization*, 6 countries do not make difference between front and rear passengers (Syria, Iraq, Lebanon, Oman, and Palestine). Jordan doesn't take into consideration the cyclist as a possible value of road user.

For *crash location*, 67% use names of streets and indication of the Km, 27% use GPS, and 9% approximate description of the location.

4.2.4. Non standardised variables

By comparing the collected variables with CADaS list (H), we can note that many important variables are not collected. Table 2 details the variables collected by each country.

4.2.5. Data flow

The flowchart below is describing data flow from collection to the storage at country level (Figure 3).

Most countries do not fill the investigation form at the crash scene but at the office, which can affect data accuracy. Even, the person who fills the data at the office is not necessary the same person who

made the investigation at the crash scene, which might mislead the interpretation of the collected data. Only two countries perform real time transfer and grouping of data in the national database.

Table 2. Current collected and missing variables

		KW*	QA	SY	JO	IQ	LB	MA	OM	PS	TN	SD	SA**	MR	YE
Person	ID														
	Age														
	Gender														
	Nationality														
	Driving license age														
	Injury severity														
	Alcohol level														
	User type														
	Position on the vehicle														
	Distraction														
	Crash /Road	ID													
Crash date															
Crash time															
Weather															
Light															
Collision Type															
Functional road class															
AAD															
Speed limit															
Pavement															
Number of lanes															
Rural /Urban															
Obstacles															
Horizontal curve															
Vertical curve															
Presence of : Tunnel /Bridge/ Roundabout															
Vehicle	ID														
	Category														
	Make														
	Model														
	Engine power														
	Special use														
	Active safety														

* The two letters abbreviation represents the ISO code of each country's name (cited above in the introduction).

** SA variables are assessed based on data published at the open data website (Traffic Accident Statistics as of 1438 H - Traffic Accident Statistics as of 1438 H.XIs - Saudi Open Data, 2020)

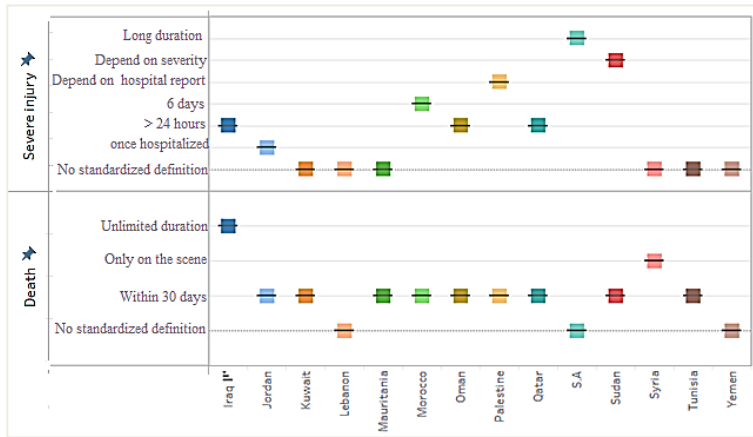


Fig. 2. Used definitions for deaths and severe injuries by country

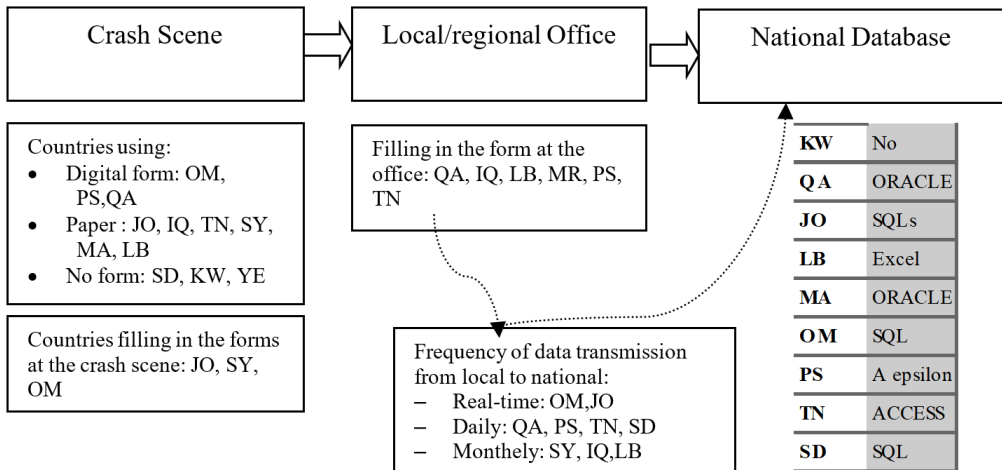


Fig. 3. Flowchart of data flow

Additional issues related to underreporting and data unreliability are expressed. 50 % of the countries reported that the police may go to the crash scene but not formally register it, because minor crashes are deemed not worthy of the administrative burden. Sometimes police does not go to the crash scene due to unavailability or proximity priorities. In addition, crash data may not be completely registered due to lack of training or skills, which may lead to input data errors. In some special cases, movement restrictions imposed by the occupation prevent the police from reaching some traffic collision sites (in palestine). Also, the reconciliation between the two

sides of the collision at the time of the collision (Iraq) may cause an underreporting.

5. Discussion of variables selection results (Steps 2 & 3)

As stated in the methodology and methods sections, test data were obtained from the FARS dataset (2016). Table 3 describes the dataset variables and classes with details. The idea was to select only “explanatory variables” that better explain the severity of the crash, which is the class. The selection results are discussed based on Figure 4 which summarizes the process of data selection.

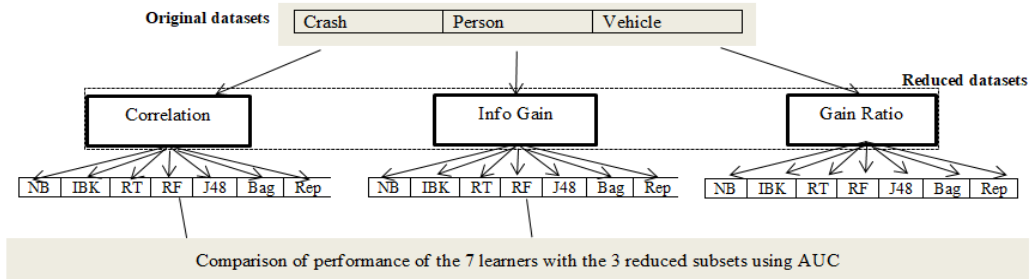


Fig. 4. Process of variables selection from FARS dataset

Table 3. Description of the original dataset used for our experiments

Datasets	Variables	Classes (number of instances)
Crash	49	a: single death (32280) b: more than one death (2468)
User	68	a: no apparent injury (2126) b: possible injury (716) c: minor injury (979) d: serious injury (887) e: fatal injury (3780)
Vehicle	86	a: zero fatalities (24408) b: with fatalities (28306)

5.1. Comparison of Crash reduced data sets

As we can notice, the dataset used for this binary classification is imbalanced in terms of instances number in each classes (classes a and b are described in table 3). For this reason, we reweighted the Cost-Sensitive Matrix, after several tests the matrix $\begin{pmatrix} 0 & 1 \\ 5 & 0 \end{pmatrix}$ showed the highest AUC and simultaneously ensured a good accuracy (>80%). Applying REPT on Info gain subset produced the largest AUC for both classes. According to Eq.9 when the AUC>0.8 (Table 4) the accuracy performance is *excellent*. The list of crash variables resulting from Info-gain selection is included in Figure 5.

5.2. Comparison of Person reduced data sets

The AUC values represent the weighted average for the five classes of person severity listed in table 3, to simplify the comparison results of this multi-classification. Again, another decision three algorithm proved its performance, which is Random Forest. RF applied to the info-gain subset shows an AUC>0.9 (Table 5), the classification performance is *outstanding* (Eq.9). The list of person variables resulting from Info gain selection is included in Figure 5.

Table 4. Metrics results of applying the classifiers* to the Crash reduced dataset

Learner	Correlation		Info-gain		Gain-Ratio		
	Accuracy	.AUC	Accuracy	.AUC	Accuracy	.AUC	
NB	85.80	0.760	85.95	0.762	85.91	0.762	a
		0.760		0.762		0.762	b
IBK	84.15	0.784	83.95	0.802	84.17	0.793	a
		0.784		0.802		0.793	b
J48	84.1	0.774	83.36	0.807	83.74	0.801	a
		0.774		0.807		0.801	b
RF	84.22	0.816	83.53	0.825	84.07	0.822	a
		0.816		0.825		0.822	b
RT	83.93	0.745	83.64	0.778	84.12	0.768	a
		0.744		0.779		0.768	b
Bag	83.55	0.800	91.82	0.791	92.11	0.811	a
		0.800		0.791		0.811	b
REPT	83.72	0.824	83.43	0.831	83.58	0.830	a
		0.824		0.831		0.830	b

*With naive Bayes (NB), k-nearest neighbour algorithm called also Instance Based Learner (IBK), random forest (RF), RT (random tree), reduced error pruning tree (RepT), and bagging (Bag).

5.3. Comparison of Vehicle reduced data sets

Random Tree applied to the info-gain subset shows an AUC>0.9 (Table 6), according to Eq.9, the classification performance is *outstanding*. Similar to the tables, the list of person variables resulting from Gain Ratio selection is included in Figure 5.

Table 5. Metrics results of applying the classifiers to the Person reduced dataset

Dataset from Learner:	Correlation		Info-gain		Gain-Ratio	
	Accuracy	Weighted Avg.AUC	Accuracy	Weighted Avg.AUC	Accuracy	Weighted Avg.AUC
NB	80.66	0.962	80.81	0.964	81.21	0.959
IBK	79.27	0.919	78.7	0.910	81.05	0.947
J48	81.25	0.949	80.31	0.940	81.46	0.955
RF	80.77	0.957	81.20	0.967	81.42	0.957
RT	78.80	0.911	75.69	0.897	80.82	0.945
Bag	80.83	0.963	80.35	0.955	81.32	0.960
REPT	80.83	0.957	79.73	0.946	81.27	0.956

Table 6. Metrics results of applying the classifiers to the Vehicle reduced dataset

Dataset from: Learner	Correlation		Info gain		Gain Ratio		
	Accuracy	.AUC	Accuracy	.AUC	Accuracy	.AUC	a
NB	79.53	0.908	83.6	0.913	80.34	0.916	b
		0.908		0.913		0.916	a
IBK	79.53	0.906	83.77	0.896	82.73	0.916	b
		0.92		0.895		0.616	a
J48	81.51	0.909	83.38	0.915	83.42	0.918	b
		0.909		0.931		0.918	a
RF	81.19	0.907	N/A	N/A	82.23	0.899	b
		0.907		N/A		0.899	a
RT	80.87	0.903	76	0.802	83.2	0.923	b
		0.903		0.802		0.923	a
Bag	77.70	0.909	81.33	0.918	79.21	0.617	b
		0.909		0.918		0.617	a
REPT	81.22	0.912	53.14	0.530	82.71	0.914	b
		0.912		0.530		0.914	a

5.4. Final List of variables

The tests which are applied on our three reduced subsets revealed that decision tree algorithms are slightly more performant comparing with the other classifiers. As a result, we choose variables subsets in which the best classifier showed the maximal AUC. The selected variables (from the FARS dataset) are the most explanatory ones for the crash severity; to differentiate it we had put a (F) in front of each.

However, the selected variables do not include descriptive variables to keep traceability and to share responsibility (E.g. IDs, names, makes...). Those descriptive variables, which are needed for the record, are extracted from CADaS list of

variables; they are indicated in the list with a (C). As a result, Figure 5 summarises a hybrid combination of variables inspired from both CADaS (C) and Fars (F). Each table contains 25 variables.

Knowing that the attribute “First Harmful event” belongs to each of the three tables of raw data (Vehicle, Crash, and Road User). It is not surprising that this variable appeared at the end of the selection process three times as important explanatory variable. However, in the e-form we are going to collect it only once (in the crash table).

Another striking finding is that almost 50% of the crash variables are related to road environment at the crash point. For this reason, we are going create for it a new table named “Road” in the e-form.

Accident	Vehicle	User
<ul style="list-style-type: none"> • Accident ID : (C) • Accident Date: (C) • Accident Time: (C) • Weather Conditions: (C) • Light Conditions: (C) • Accident Type Variables: (C) • Number of occupant in Motor Vehicles (F) • Pedestrians: (F) • Drunk Drivers: (F) • First Harmful Event(F) • Manner of Collision (F) • Number of Vehicle in crash (F) • Rural/urban: (C)&(F) • Road Functional Class - Road 1: (C)&(F) • Road Functional Class road 2: (C)&(F) • Speed Limit - road 1: (C)&(F) • Speed Limit road 2 : (C)&(F) • Junction: (C) • Surface Conditions : (C) • Carriageway Type : (C) • Number of Lanes: (C) • Work Zone Related: (C) • Location Coordinate X, Y (C) • Vertical signalisation (C) • Horizontal signalization (C) 	<ul style="list-style-type: none"> • Accident ID (C) • Vehicle ID (C) • Vehicle Type (C) • Vehicle Special function (C) • Engine power (C) • Active safety equipment (C) • Make (C) • Model (C) • Registration year (C) • Registration country (C) • Vehicle manoeuvre (C)&(F) • First point of impact (C) • First Object Hit in Carriageway(C) • First Object Hit off Carriageway (C) • Vehicle Insurance (C) • First Harmful Event (F) • Hit and run (C)&(F) • Vehicle trailing(C)&(F) • Jackknife(F) • Gross Vehicle Weight Rating(F) • Rollover(F) • Extent of Damage(F) • Most Harmful Event(F) • Critical Event- Pre-crash (F) • Vehicle Configuration (F) 	<ul style="list-style-type: none"> • Accident ID (C) • Vehicle ID (C) • Person ID (C) • Age (C)&(F) • Nationality (C)&(F) • Gender (C) • Road user type (C)&(F) • Injury severity/ Mais (C)&(F) • Alcohol test status(C)&(F) • Alcohol test type (C)&(F) • Alcohol result (C)&(F) • Drug test(C)&(F) • Drug test result (F) • License issue date (C) • Driving license validity (C) • Safety equipment (Rest_use) (C)&(F) • Position in the vehicle (C)&(F) • Distracted by device (C) • Psychophysical/ physical condition (C) • Death time (if applicable) (F) • Transported to First Treatment Facility (F) • Died at scene (if applicable) (F) • Ejection (F) • Air Bag(F) • First Harmful Event(F)

Fig. 5. Final list of variables

6. Unified electronic investigation form (Step 4)

A considerable amount of mobile applications exist and ensure mobile data collection based on the use of internet at crash scene either for geo-localising the crash or for making data transfer from the handheld device (mobile/PDA) to the database. However, during the survey some countries expressed their inability to use internet at the crash scene location because of its non-availability, and in order to avoid security issues since the collected data contained some personal information like the names and the IDs.

Moreover, we were looking for a tool that ensures the permanence, at low cost, easy and flexible to use. The unique tool inspired from tracking animals verified our needs and more. It allowed acquiring, geo-referencing, storing and transferring local spatial knowledge from the handheld device equipped with a GPS (CyberTracker, 2020), and ensure some additional benefits like recording voices notes and taking photos illustrating the conditions and descriptions of the infrastructure, and the vehicles (Fig. 6).

The new data flow was described as following:

- a) Transfer: Synchronized data transfer can be done once the investigator arrived to the police office (via cable or secured intranet) allowed data to be extracted from the handheld device.
- b) Analysis: Unlike other applications, through desktop application data can be queried and visualized in table or map form (using built-in GIS capability). Queries can be saved in reports, which are updated when new data is collected.
- c) Export: Data can also be exported in a wide variety of formats for use in other programs or

- d) Storage: depending on the country wishes, data can be transferred remotely (synchronized) either to a server database (MySQL, Microsoft SQL Server and PostgreSQL) as illustrated in Fig. 7, or to online databases (Esri ArcGIS Online, SMART, and Earth-Ranger).

The selected list of variables and their possible values were inserted in a customized electronic form using sequence of interlinked screens in the Cyber-tracker.

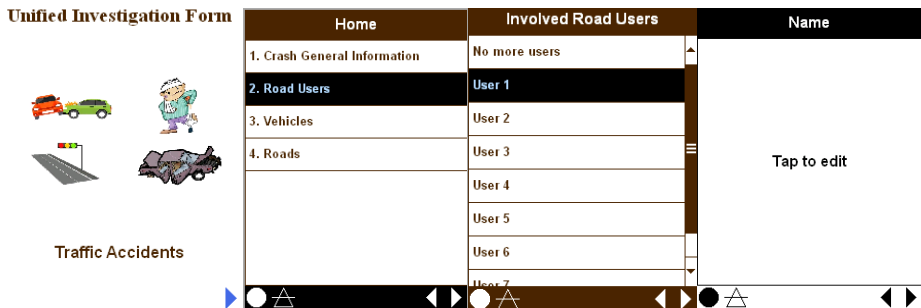


Fig. 6. Icon based screens from the e-form prototype (Mobile view)

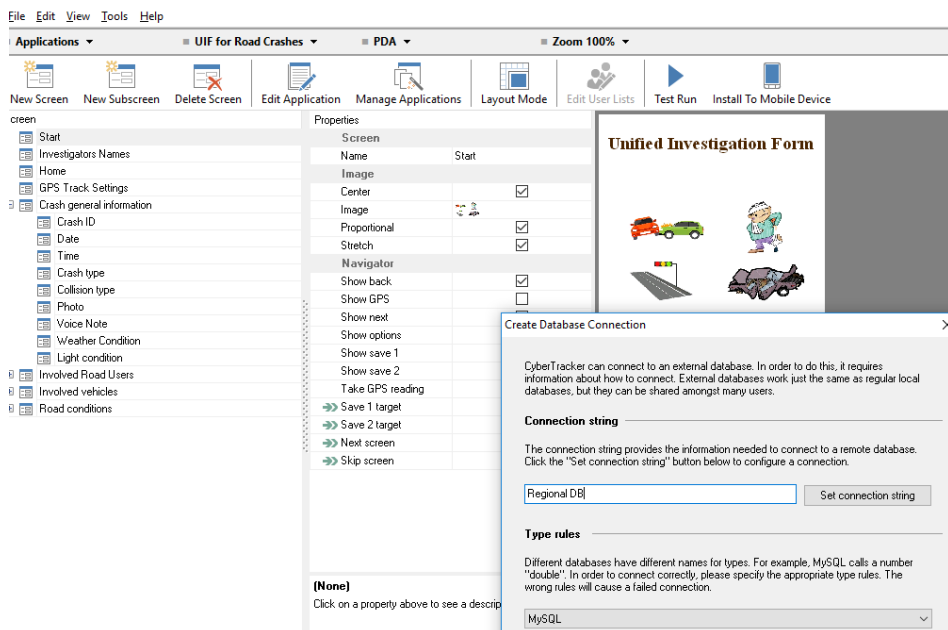


Fig. 7. Desktop view, database connection

7. Conclusions

This study is the first step towards achieving the regional vision of UN-ESCWA of establishing an Arab Road Safety Observatory, aiming to present regional guidance and to facilitate regional comparisons. In brief the paper investigated some CDRS issues and their remedy:

- a) To identify CDRS issues, an assessment was conducted and proved that these systems suffer from several management and data dysfunctions, specifically the heterogeneous, insufficient, and non-standardized variables which are collected manually.
- b) To improve variables quality, new collected datasets were proposed based on reducing FARS datasets. The reduced datasets were selected through three selection methods of identify the relevant variables. In order, to compare the results and to get an idea of which reduced dataset had the best performance, we had tested seven classifiers in terms of classifying *severity*. Based on AUC comparison, decision trees algorithms showed better performance on selecting best reduced datasets.
- c) To improve the collected datasets, not only explanatory variables are needed, but also descriptive ones to keep tracking and to share responsibility. The latter were extracted from CADaS. As a result we got a hybrid list of 25 variables for each of the three components (crash, person and vehicle).
- d) To standardize and harmonize the collected data at the bottom (because it is more practical) a unified e-form for data collection was suggested based on the hybrid list of variables. To build the new e-form, no big investments are needed to start from scratch. Existing tools like CyberTracker is tested and it has proved its efficiency. Getting inspired from other fields is a key.

Acknowledgements

The entire project of research was made possible by a grant from Renault Foundation; we are indebted to its generous support. We gratefully acknowledge the help provided by UN-ESCWA focal points of transport who participated to the conducted survey for assessing crash data reporting systems in the Arab region. We want to extend our special thanks to the Science Excellence programme of the Ministry of Science and Higher Education in Poland,

which co-founded the publication of this paper. We gratefully acknowledge the constructive comments of the anonymous reviewers.

References

- [1] Abellán, J., & Castellano, J. G. (2017). A comparative study on base classifiers in ensemble methods for credit scoring. *Expert Systems with Applications*, 73, 1–10. <https://doi.org/10.1016/j.eswa.2016.12.020>
- [2] Almatawah, D. J. (2014). *Towards Improving Crash Data Management System in Gulf Countries*. 4(9), 6.
- [3] Atnafu, B., & Kaur, G. (2017). *Analysis and Predict the Nature of Road Traffic Accident Using Data Mining Techniques in Maharashtra, India*. 4(10), 10.
- [4] Beasley, R. E. (2020). *Essential ASP.NET Web Forms Development: Full Stack Programming with C#, SQL, Ajax, and JavaScript*. Apress. <https://doi.org/10.1007/978-1-4842-5784-5>
- [5] Belgiu, M., & Drăguț, L. (2016). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, 114, 24–31. <https://doi.org/10.1016/j.isprsjprs.2016.01.011>
- [6] Bellis, E. (2015). *FARS Analytical User's Manual*. 592.
- [7] Bhondave, R., Kalbhor, M., Shinde, S., & Rajeswari, K. (2014). *Improvement of Expectation Maximization Clustering using Select Attribute*. 1.3, 503–503.
- [8] Bliss, T., & Breen, J. (2009). *Country Guidelines for the Conduct of Road Safety Management Capacity Reviews and the Specification of Lead Agency Reforms, Investment Strategies and Safe System Projects*. 329.
- [9] Castro, Y., & Kim, Y. J. (2016). Data mining on road safety: Factor assessment on vehicle accidents using classification models. *International Journal of Crashworthiness*, 21(2), 104–111. <https://doi.org/10.1080/13588265.2015.1122278>
- [10] Cuartas, K., Anzola, J., & Tarazona, G. (2015). *Classification Methodology Of Research Topics Based In Decision Trees: J48 And Randomtree*. 8, 19413–19424.

- [11] CyberTracker. (2020). *CyberTracker GPS Field Data Collection System—Home*. <https://www.cybertracker.org/>
- [12] Dababneh, A., Fouad, R. H., & Majeed, A. J. H. (2018). *Assessment of Occupational Safety and Health Performance Indicators for Jordan*. 8.
- [13] Dadashova, B., Arenas-Ramírez, B., Mira-McWilliams, J., & Aparicio-Izquierdo, F. (2016). Methodological development for selection of significant predictors explaining fatal road accidents. *Accident Analysis & Prevention*, 90, 82–94. <https://doi.org/10.1016/j.aap.2016.02.003>
- [14] Howard, C., & Linder, A. (2014). *Review of Swedish experiences concerning analysis of people injured in traffic accidents*. www.vti.se/publications
- [15] Hussain, S., Abdulaziz Dahan, N., Ba-Alwi, F. M., & Ribata, N. (2018). Educational Data Mining and Analysis of Students' Academic Performance Using WEKA. *Indonesian Journal of Electrical Engineering and Computer Science*, 9(2), 447. <https://doi.org/10.11591/ijeecs.v9.i2.pp447-459>
- [16] Jha, A. N., Tiwari, G., & Chatterjee, N. (2020). Road Accidents in EU, USA and India: A critical analysis of Data Collection Framework. In P. K. Kapur, O. Singh, S. K. Khatri, & A. K. Verma (Eds.), *Strategic System Assurance and Business Analytics* (pp. 419–443). Springer Singapore. https://doi.org/10.1007/978-981-15-3647-2_31
- [17] Kalmegh, S. (2015). *Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News*. 2(2), 9.
- [18] Kang, K., & Michalak, J. (2018). *Enhanced version of AdaBoostM1 with J48 Tree learning method*. 4.
- [19] Khallaf, R. I., & Yasseen, A. Y. (2016). Improvement of Road Safety Within the Oil and Gas Industry and its Effect on the Community—Case Study. *SPE International Conference and Exhibition on Health, Safety, Security, Environment, and Social Responsibility*. SPE International Conference and Exhibition on Health, Safety, Security, Environment, and Social Responsibility, Stavanger, Norway. <https://doi.org/10.2118/179422-MS>
- [20] Kumar, K., & Singh, J. (2016). Network Intrusion Detection with Feature Selection Techniques using Machine-Learning Algorithms. *International Journal of Computer Applications*, 150(12), 1–13. <https://doi.org/10.5120/ijca2016910764>
- [21] Laaraj, N., Boutahari, S., & Jawab, F. (2018). The Road Safety Information Systems Appropriate to the Systemic Approach: The Case of Morocco. *2018 IEEE 5th International Congress on Information Science and Technology (CiSt)*, 74–79. <https://doi.org/10.1109/CIST.2018.8596590>
- [22] Liebenberg, L. (1999). *Rhino Tracking with the CyberTracker Field Computer*. 27, 59–61.
- [23] Mehmood, A., Taber, N., Bachani, A. M., Gupta, S., Paichadze, N., & Hyder, A. A. (2019). Paper Versus Digital Data Collection for Road Safety Risk Factors: Reliability Comparative Analysis From Three Cities in Low- and Middle-Income Countries. *Journal of Medical Internet Research*, 21(5), e13222. <https://doi.org/10.2196/13222>
- [24] Montella, A., Chiaradonna, S., Criscuolo, G., & De Martino, S. (2017). Perspectives of a web-based software to improve crash data quality and reliability in Italy. *2017 5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, 451–456. <https://doi.org/10.1109/MTITS.2017.8005714>
- [25] O'Donnell, C., O'Malley, M., Lynch, D., & Mullins, E. (2019). *WESPAS Cruise Report*.
- [26] Patil, M. D., & Sane, D. S. S. (2014). *Effective Classification after Dimension Reduction: A Comparative Study*. 4(7), 4.
- [27] Ramya, S., Reshma, Sk., Manogna, V. D., Saroja, Y. S., & Gandhi, G. S. (2019). Accident Severity Prediction Using Data Mining Methods. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 528–536. <https://doi.org/10.32628/CSEIT195293>
- [28] Sarraj, Y. (2016). *Developing Road Accidents Recording System in Palestine*. 30, 188–204.
- [29] Singh, A., N., M., & Lakshmiganthan, R. (2017). Impact of Different Data Types on Classifier Performance of Random Forest,

- Naïve Bayes, and K-Nearest Neighbors Algorithms. *International Journal of Advanced Computer Science and Applications*, 8(12). <https://doi.org/10.14569/IJACSA.2017.081201>
- [30] Spanu, V., & McCall, M. K. (2013). Eliciting Local Spatial Knowledge for Community-Based Disaster Risk Management: Working with Cybertracker in Georgian Caucasus. *International Journal of E-Planning Research (IJEPR)*, 2(2), 45–59. <https://doi.org/10.4018/ijepr.2013040104>
- [31] SWOV. (2016). *Data_sources.pdf*. https://www.swov.nl/sites/default/files/bestanden/wegwijzer/data_sources.pdf
- [32] Tharwat, A. (2018). Classification assessment methods. *Applied Computing and Informatics*. <https://doi.org/10.1016/j.aci.2018.08.003>
- [33] Theofilatos, A., Chen, C., & Antoniou, C. (2019). Comparing Machine Learning and Deep Learning Methods for Real-Time Crash Prediction. *Transportation Research Record: Journal of the Transportation Research Board*, 2673(8), 169–178. <https://doi.org/10.1177/0361198119841571>
- [34] *Traffic Accident Statistics as of 1438 H - Traffic Accident Statistics as of 1438 H.xls—Saudi Open Data*. (2020). https://data.gov.sa/Data/en/dataset/traffic-accident-statistics-as-of-1438-h/resource/0e1a41f8-abee-4b07-9c28-dca768b30af6?view_id=06d50d94-458c-4a36-9edd-7c991e624ec0
- [35] Trafikverket Swedish Traffic administration. (2019). *Data Collection* [Text]. Trafikverket. <https://www.trafikverket.se/en/startpage/operations/Operations-road/vision-zero-academy/Vision-Zero-and-ways-to-work/data-collection/>
- [36] UK department of transport. (2013). *Reported Road Casualties in Great Britain: Guide to the statistics and data sources*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/259012/rrcgb-quality-statement.pdf
- [37] Vinutha, H. P., & Poornima, B. (2018). An Ensemble Classifier Approach on Different Feature Selection Methods for Intrusion Detection. In V. Bhateja, B. L. Nguyen, N. G. Nguyen, S. C. Satapathy, & D.-N. Le (Eds.), *Information Systems Design and Intelligent Applications* (Vol. 672, pp. 442–451). Springer Singapore. https://doi.org/10.1007/978-981-10-7512-4_44
- [38] Wang, R., & Tang, K. (2009). Feature Selection for Maximizing the Area Under the ROC Curve. *2009 IEEE International Conference on Data Mining Workshops*, 400–405. <https://doi.org/10.1109/ICDMW.2009.25>
- [39] Wang, S., Li, D., Petrick, N., Sahiner, B., Linguraru, M. G., & Summers, R. M. (2015). Optimizing area under the ROC curve using semi-supervised learning. *Pattern Recognition*, 48(1), 276–287. <https://doi.org/10.1016/j.patcog.2014.07.025>
- [40] Wilson, J. (2018). Universal screening with automated essay scoring: Evaluating classification accuracy in grades 3 and 4. *Journal of School Psychology*, 68, 19–37. <https://doi.org/10.1016/j.jsp.2017.12.005>
- [41] World Health Organization. (2013). *Global status report on road safety 2013 supporting a decade of action*. World Health Organization. <http://site.ebrary.com/id/10931312>
- [42] World Health Organization, issuing body, & ProQuest (Firm). (2018). *Global status report on road safety 2018*. <https://ebookcentral.proquest.com/lib/qut/detail.action?docID=5910092>
- [43] Zhang, H., Rangrej, J., Rais, S., Hillmer, M., Rudzicz, F., & Malikov, K. (2019). Categorizing Emails Using Machine Learning with Textual Features. In M.-J. Meurs & F. Rudzicz (Eds.), *Advances in Artificial Intelligence* (Vol. 11489, pp. 3–15). Springer International Publishing. https://doi.org/10.1007/978-3-030-18305-9_1
- [44] Żukowska, J. (2015). Regional implementation of a road safety observatory in Poland. *Archives of Transport*, 36(4), 77–85. <https://doi.org/10.5604/08669546.1185212>