A Deep-Learning Model for Predicting and Visualizing the Risk of Road Traffic Accidents in Saudi Arabia: A Tutorial Approach

Maram Alrajhi[1], Mahmoud Kamel[2]
Information Systems Department, Faculty of Computing and Information Technology
King Abdulaziz University, Jeddah, Saudi Arabia

***Abstract*—Around the world, road traffic accidents (RTAs) cause significant concerns for decision makers and researchers on traffic safety. The diversity, rarity, and interconnectivity of historical data on factors causing car accidents point to the need for more focused studies for analyzing, predicting, and visualizing the risk of accidents over the short and long term for preventive purposes. There are many techniques and tools applied to analyze, forecast, and visualize risk. Most RTA studies have applied linear time-series methods to forecasting the risk with limited studies applying machine-learning and deep-learning techniques, especially in Saudi Arabia. Recently, many global studies have applied long short-term memory (LSTM) networks, which can be used to automatically learn the temporal dependence structures for challenging time-series forecasting problems. This paper displays a tutorial for designing a prototype of an interactive analytical tool based on a multivariate LSTM model for time-series data to predict future car accidents, fatalities, and injuries in the Kingdom of Saudi Arabia (KSA). This interactive tool visualizes the real data with the predicted values regionally in a web browser with Python. The tutorial represents the annual data of the period between 1417 (1996) and 1433 (2013), then uses the data with some contributing factors, such as population, gender, nationality, number of vehicles, and length of road, to generate the input data and predict the future values of accidents, fatalities, and injuries up to the year 1452 (2030). After that the real and predicted values are visualized regionally on an interactive map that represents the degree of risk. Finally, the paper discusses the evaluation and utilization of the proposed prototype in the future in the field of road safety.**

Keywords*—*LSTM for time-series forecasting; deep learning; RTA; data visualization; interactive map; Saudi Arabia

I. Introduction

Predicting road traffic accidents (RTAs) can improve road safety for both travelers and road-safety administrators. Nowadays, the accelerated development in data-collection techniques facilitates the availability of big datasets. Several studies have discussed the prediction of RTA risk regarding multiple influencing factors. Some have considered human factors [1], and others, vehicle factors [2]; still others have concentrated on road and environmental factors [3]. The rest have combined all these factors to predict risk [4].

Two approaches exist for estimating the time-series forecasting of RTAs. The first approach is a regression problem that forecasts the number of accidents based on the attributes of the accident dataset [5]. The second is a classification problem that predicts the severity of crashes based on the crash dataset. The autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) are two basic prediction methods that are notable for time-series prediction models based on regression [6] Most statistical models have difficulty dealing with complexity, randomness, irregularity, and the nonlinearity of real data, limiting the precision of their predictions; machine-learning models as in [7] can obtain more accurate predictions when compared with the traditional models because of repeated training iterations and learning approximation mechanisms [8]. However, in the case of time-series prediction, a lack of efficient processing of sequence dependencies between input variables is a problem in some machine-learning methods [9].

Recently, specific patterns of deep learning, especially recurrent neural networks (RNNs) [10] and long short-term memory (LSTM) algorithms [11], have been successfully applied to time-series forecasting tasks. RNNs are the most efficient method [12] as they are a kind of artificial neural network (ANN) containing connected nodes in a loop form. Looping allows the internal state of the neural network to manifest dynamic timing behavior. However, some problems, such as gradient disappearance, often occur during the training of RNNs with the increase in the length of the processing time series, especially in networks using conventional activation functions, such as sigmoid or tanh functions; this limits their prediction accuracy.

The LSTM method [13] is generally one of the most efficient approaches for dealing with time-series forecasting problems; here, a simple RNN is employed with some multithreshold gates for resolving and forgetting memory problems with long period of time. The structure of LSTM is inspired by cognitive neuroscience; some researchers have even introduced attention mechanisms into the coding–decoding framework and encoded semantics in long-term memory to improve the selection of input sequences and the information-processing capabilities of the neural multiorder network [14]. Recently, LSTM with different mechanisms has been broadly applied and performed well in diverse types of deep-learning tasks, such as image captioning [15], visual question answering [16], traffic flow [17], road accidents [18], and speech recognition [19].

For visualizing the volume of accident risk, many data visualization techniques, such as interactive maps, bar charts,

bubble charts, and combinations of more than one tool [20], aim to facilitate user understanding of information interactively in a web browser [21]. Several web-programming languages, such as Python and JavaScript, contribute to designing these interactive tools [22]. Some studies have analyzed the time-series problem of RTAs regarding different factors, even focusing on classification or regression. Previous studies have noted some machine-learning techniques that can be used for forecasting RTA data. The present study aims to predict and visualize RTA data by applying LSTM with Python.

This paper is presented as follows: Related work is described in Section II. Then, the whole process of the proposed experiment is considered in Section III. The evaluation of the experiment's results is given in Section IV. Section V discusses the utilization and improvement of the proposed prototype. Finally, the conclusion is provided in Section VI.

II. Related Work

A. Time-Series Forecasting with Long Short-Term Memory

At one time, statistical methods—especially error, trend, and seasonal ETS models and ARIMA models—performed better than other methods in terms of time-series forecasting [23]. However, traditional time-series forecasting has been affected by the advancements in the field of deep learning for time-series prediction. In recent years, there has been attention paid to RNNs and LSTM, with their deep-learning applications for many disciplines, including natural language processing, finance, and computer vision. Deep-learning methods have the capability of identifying patterns and structures of data, such as complexity and nonlinearity, in time-series predicting [24]. LSTM is used in such predictive research for time-series data in different fields. The experiment in [25] solves the location prediction problem using time-series analysis. The data are transformed into a multivariate time series, and this is predicted using RNN and its variants (bidirectional RNN [BRNN] and LSTM). These developed models learn the movement behavior of a specific vehicle. Analyzing the learning of many different vehicles, it is observed that vehicles that travel across larger geographic areas, and thus have a more complex movement pattern, also have a higher prediction error. The problem of high variance is approached by using regularization and dropouts. The study shows from the experiments that the LSTMs have better predictions when compared with simple RNNs and BRNNs in analysis and implementation of the near-time future prediction of truck locations.

Another study [26] in the finance field compared LSTM and support vector regression (SVR) for implementing a robust forecasting model to facilitate predicting phone prices in European stores. The study applied LSTM for its architecture to consider the problems that are not addressed by classic RNNs. In addition, the study utilized the support vector machine (SVM) method for both classification and regression because of its ability as a potent and accurate machine-learning method for univariate models, as shown in variant studies. In terms of introducing more variables for the multivariate approach, the models exhibited better prediction performance. By comparing the results, it was found that the SVR model can forecast the next-day price with a root mean squared error (RMSE) value of 33.43 euros for the univariate model. Nevertheless, applying multivariate models, the LSTM RNN gives the most accurate forecasting result for the next day's price, with an RMSE of 23.640 euros.

In the case of model flexibility to deal with missing values, [17] discussed developing the LSTM model, the model masking with vector M to perform two functions in predicting the traffic flow, called LSTM-M. This advanced approach can assume traffic flow data despite the missing values. The traffic flow dataset used in the study covered the time interval between April 4th, 2015 and January 3rd, 2016, including spring and winter seasons, by considering all-weather fluctuations for rainy, snowy, and sunny days. The study area relied on data collected from different locations of Hangzhou elevated road. In the data gathering step, the study faced missing data related to describing the working condition of the emergency detection and road equipment. The study applied multiscale temporal smoothing to infer the missing data and determine the prediction for the residual values. However, the time interval affected the rate of missing data. In a particular period, there was a small occurrence probability of overall data loss appearing with the larger time interval. It is unreasonable to have a large time interval when the traffic situation should be reflected accurately in a short period. As a result, the LSTM-M's superiority exceeds that of LSTM, producing a gain of over 1.5% in the average accuracy. When the traffic flow is massive, this implies a large mean absolute error (MAE) value, meaning that a greater presence of vehicles on the road may give rise to traffic accidents or bursts, and the traffic flow series may lead to fluctuations.

Author in [26] introduced two forecasting tasks based on the LSTM model. The first task was recognizing the unique features of the traffic data by applying deep LSTM at the peak hour for forecasting traffic. The other task was dealing with traffic datasets collected from 2,018 loop detectors, located along the Los Angeles arterial streets and highways. These loop detectors showed the datasets as large-scale and real-world datasets because they covered 5,400 miles cumulatively for intervals between May 19th and June 30th, 2012. The second task utilized a mixture of deep LSTM architecture that consolidated deep LSTM with a stacked autoencoder. The loop detectors picked three essential types of accidents, which were as follows: minor injuries, rolling, and major injuries. Approximately 1,650 sensors detected 6,811 accidents. There were different attributes associated with each accident, like the affected traffic direction and downstream post mile. The framework for the two situations proved the significant performance improvement in large-scale real-world traffic data. The suggested approach showed 30–50% superiority over the baseline, which was achieved by training end to end with proper regularization. In addition, the study described a novel technique to illustrate the model with signal stimulation and engaging summaries noted from the trained neural network (NN).

Regarding prediction of the frequency of traffic accidents using spatial data, [18]represented the risk of traffic accidents based on frequency, constructing a deep-learning model relying on LSTM for capturing the regional spatial and temporal correlation patterns. The studied dataset considered

traffic accidents based on different features, such as time, and the global positioning system coordinates of Beijing between 2016 and 2017. The model performance relied on comparing RMSE values, with the predicted risk map illustrating the effectiveness and accuracy of the proposed model. The study showed that the influences of temporal–spatial features, big traffic accident data, and the deep RNN on gaining accurate traffic accident risk prediction. The traffic accident warning system utilized in the study approach can help people avoid traffic accidents by choosing safer regions. However, the study had some limitations in that it utilized the traffic accident data for prediction, without other related data, such as human mobility, traffic flow, special events, and road characteristics, which may also be vital for traffic accident risk forecasting. The prediction results were coarse grained, and therefore, could not afford the level risk forecast of the road accident. Nevertheless, the approach can be efficiently utilized for road network–based forecasting. Therefore, future research should be conducted to combine the urban road network structure and comprehensive features related to traffic accidents to generate more reliable prediction results. Indeed, the LSTM approach's ability to generate predictions using the temporal features of traffic has been illustrated. However, LSTM algorithms fail to capture the spatial features and represent them on the map [27]. For this reason, the present study isolated the spatial features from prediction and took only their benefit to represent real and predicted accidents, fatalities, and injuries on an interactive map.

B. Visualizing Risk

Data visualization is a successful way to encode information so that our eyes can observe it and our brains can comprehend it. This is much more a science than an art, and we can only achieve successful results by studying human thought. The purpose is to translate abstract information into visual representations that can accurately, efficiently, easily, and meaningfully decode information. Data visualization is used to preset awareness of multiple risks in the world or make a decision concerning a specific risk. There are different purposes for using choropleth maps of a specific country or area. For example, maps are used to visualize diseases[28], flooding [29], traffic accident [11], and other risks and social issues. There are various tools for visualizing data; as discussed in [30] these include technical and nontechnical tools. Author in [20] applied bar and row charts with colored maps using a library called Dimensional Charting JavaScript (dc.js). This library allows highly efficient exploration of large data on accidents extracted from newspapers by the text-mining process. The dc.js library facilitates visualizing data and analyzing it in an interactive way on mobile devices or in a web browser. The map also has bar charts with color variations to visualize the statistics of accidents, with the degree of risk according to accident locations. To create online interactive maps with charts, plotly and folium are powerful and famous libraries for visualizing data by python mentioned in[30].

Folium is a python wrapper for Leaflet.js; this library facilitates the binding of the data to the map. The tool was developed to take in the trajectory of paths and plot them on a map. In one work, the folium library, with the help of Java code, contributed to plotting big datasets into an interactive geotemporal map of Morocco for assessing the air quality affected by significant correlations between emissions spread in air; this was done by incorporating the set of industrial activities associated with thermal power factories, plants, ports, and transportation[31]. Another study [32]applied folium to creating dynamic and interactive energy maps along with different informative charts allowing different stakeholders to view the characterization of the energy performance of buildings located in different areas in Italy, supported by different data and knowledge-visualization techniques. In addition [25], used folium with Matplotlib and plotly for visualizing the data in the near-time future location prediction of trucks' locations under location-based services to provide valuable services in the transportation sector to SCANIAs company customers in Taiwan.

C. Saudi Arabia as a Case Study

The RTA problem has a limited study history in traffic safety and traffic awareness literature in the Kingdom of Saudi Arabia (KSA). The total publication number for the traffic safety field and traffic awareness field is around 50 studies as identified in [33]. Author in [33] examined the causality relationship between the RTA, GDP, population, road miles, road vehicles, and number of driver's licenses in the KSA for 1971–2012, using a multivariate framework analysis. The study applied the autoregressive distributed lag (ARDL) model for cointegration in the KSA, employing the cointegration test. The results showed that the variables are cointegrated, meaning that there is a stable long-run relationship between RTA and its determinants, although there may be deviation in the short-run; thus, RTAs and the independent variables have bidirectional causality in the KSA.

After 2016, some studies discussed the RTA issue in relation to different fields, such as data mining for analyzing RTAs[34] using the internet of things (IoT) for classifying reasons for RTAs [35] employing the IoT to create e-awareness application against RTAs[35], and detecting RTAs via the IoT with ANNs[36]. In terms of prediction[37], performed the first study was conducted in 1995 and used three ARIMA models with a Box–Jenkins methodology for time-series data to predict the total traffic accidents, injuries, and fatalities in the KSA. The three models developed in this study showed the need for a quick revision of the current traffic safety programs in Saudi Arabia since no decrease was evident in the future predictions of the numbers of accidents, injuries, and fatalities. An improvement in traffic safety plans is urgently needed to reduce these numbers in the future. Author in [6] showed that the ARIMA model can forecast RTAs using time-series data. It introduced the applicant model for making RTA forecasts for up to 7 years. The prediction values of traffic accidents showed that there will be increasing deaths and injuries in the coming years.

Predicting RTAs is still a highly challenging problem because of the diversity of factors causing accidents from one place to another. In addition, data resources are restricted in some countries, limiting the quality of prediction results and visualization tools. This paper aims to elaborate on the concept of RTA prediction using deep-learning techniques, with the Kingdom of Saudi Arabia (KSA)—where such research is still lacking—receiving focus as a case study. This tutorial can be a

starting point for considering the issue more deeply in future research. The paper introduces the design of an online interactive map to visualize both real and predicted RTAs, fatalities, and injuries for all KSA regions in the years 1417 to 1452, based on the Hijri Calendar (AH), according to the annual historical data.

III. Designing the Process

The entire process of the experiment is an interactively advanced solution to visualize the actual and predicted numbers of accidents (As), fatalities (Fs), and injuries (Is). This process involves Python programs with five leading platforms—three for future prediction models and two for visualization. Therefore, the complete process can help address the global need for RTA prediction. First, it is necessary to prepare the historical data for visualizing the training dataset. The preparation step requires some preprocessing to prepare the data for prediction.

After predicting, the Python program interactively visualizes the number of road accidents, fatalities, and injuries in a web browser. There are different tools that can be utilized for visualization purposes, such as geographical maps and histograms. Visualizing data contributes to facilitating the understanding of complicated statistics in an easy, accurate way in a short time; due to these features, the concept of visualizing data has become popular on the internet. Fig. 1 illustrates the sequence of the experimental process. The explanation below identifies the steps in the figure, considering both directions.

A. Preparing Raw Data

The study dataset comprises numerical data obtained from reports of the KSA's General Department of Traffic, Ministry of Interior, Saudi Open Data, and General Authority for Statistics. The data contain the statistics on accident, injury, and fatality records classified in regional form between 1417 (1996) and 1433 (2013). In addition to the necessary information—that is, totals of "A," "I," and "F" per KSA region—the dataset also contains many valuable features related to accidents, such as vehicle numbers (N), road lengths (L), and population statistics (P) for males (Pm) and females (Pf) in each region. The dataset is arranged in the form of CSV/Excel files to be ready for the next step. Ultimately, we attain four files (accidents, injuries, fatalities, and factor attributes) based on the regional classification. The dataset requires some preprocessing operations to be ready for modeling. There is a need to perform a set of preprocessing tasks with the help of Python libraries to import, read, operate, clean, interpolate, extrapolate, and visualize data. We import Pandas to read, manipulate, and perform mathematical operations on the data in the form of CSV/Excel files. In addition, we use this library to handle the missing values via cleansing, extrapolating, and interpolating data. Other libraries used are Numpy to perform complex mathematical array operations and Matplotlib for data visualization in the form of line charts.

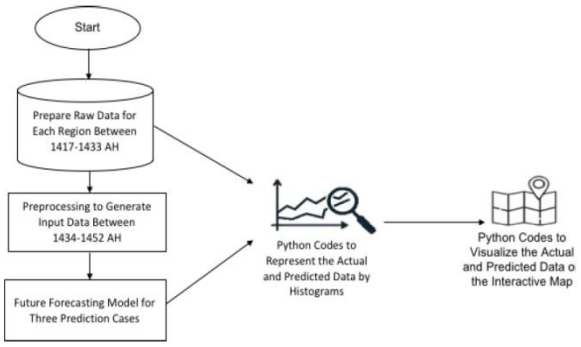


Fig. 1. The Steps Performed in the Tutorial.

B. Preprocessing to Generate Input Data

There is a need to populate all the data for 1434–1452 AH because there are no visible data for this period. This is done by calculating the average annual growth percentage for each input feature column per region after preprocessing the missing values. The four following features are used to make the required predictions:

- Male population (Pm).
- Female population (Pf).
- Number of registered vehicles (N).
- Length of roads in kilometers (L).

All the mentioned features with the original statistics on accidents, fatalities, and injuries can be accessed for each region in the project folder; Fig. 2 and 3 illustrate these data before and after preprocessing. Based on the provided dataset for 1417–1433, we build the generating input data for each region for the years 1434–1452, as illustrated in Fig. 3, to predict the following three quantities for every given region in the dataset: As, Fs, and Is. After generating the input data, the data must be normalized before regression by standardizing the column-wise data along the mean and dividing the column-wise data by the standard deviation for each column.

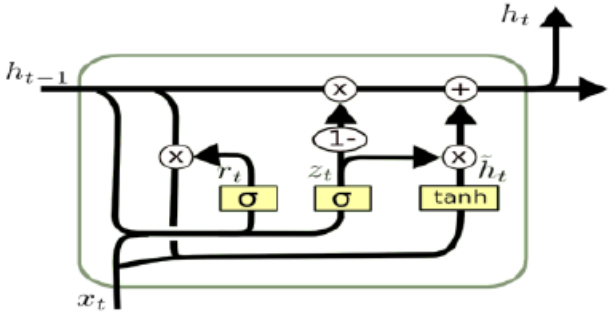


Fig. 2. Structure of the LSTM Cell that Describe (1), (2), (3), and (4).

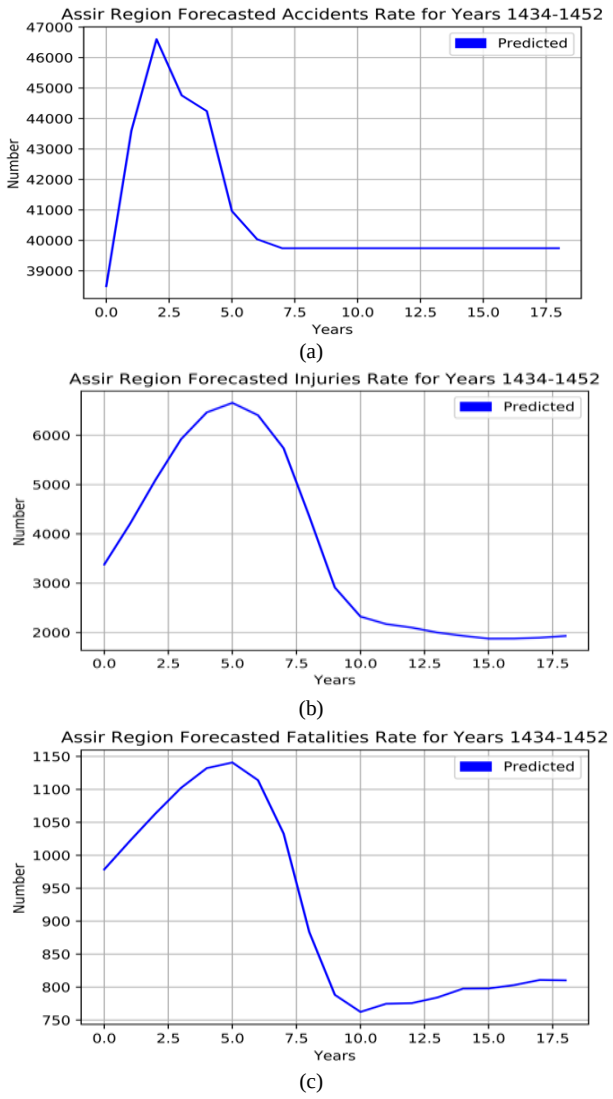


Fig. 3. (a). Future Accidents Prediction for the Assir Region. (b). Future Injuries Prediction for Assir Region. (c). Future Fatalities Prediction for Assir Region.

C. Future Forecasting Model

In this phase, the data are split into two groups, with the years 1417–1433 for training and 1434–1452 for testing. Then, the data are transformed into a sequence-to-sequence learning model. The data must pass through a set of tasks with the support of LSTM. LSTM is the most popular variant of RNN, representing a classic deep-learning method for handling sequence learning tasks. The Python libraries and Keras, which is based on Tensorflow, are used to import, read, operate, clean, interpolate, extrapolate, and build the models. The preprocessing returns a reframed data frame, which is first split into the training and test data based on the number of years, then split based on the inputs and outputs of both separated data frames. We use LSTM for the time-series data forecasting, which requires a specific three-dimensional input format of our dataset, where we need the size of input rows, timestep value, and feature length value. The model is multivariate, comprising (N, 1, M), with N representing the sample number, 1 the timestep, and M the number of multiple features depending on factors selected for the study. For each region; the multivariate model is constructed by 16 samples and one timestep with five features regarding each predicted quantity for 19 years. The model is executed based on one layer of four LSTM neurons, converging to one layer of a single dense neuron. The current models are run for 10 epochs each, with a batch size of 2, where the data are kept unshuffled to maintain the sequence. The model's loss function is the mean absolute error (MAE), with the Adam algorithm being the optimizer. The training data comprise a sequence of features that are passed into a layer of a number of LSTM neurons. Each LSTM neuron has the following inner workings [38]:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]), \quad (1)$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]), \quad (2)$$

$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t]), \quad (3)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t, \quad (4)$$

where h_{t-1} is the output from the previous iteration of the same neuron, h_t denotes the output of the neuron for the same iteration, and x_t denotes the input for the current iteration. The adam optimizer is used for the weight optimization. One dense layer specifies the need to estimate a single value, that is, the number of accidents, injuries, or fatalities for a given instance of time. The loss function represents the difference between the real and predicted value, which helps in updating the weights, as shown in Fig. 2.

Finally, the model represents three graphs for each region, indicating the predicted values for the future As, Is, and Fs. The following figures illustrate the model in three cases for the Assir region sample—Fig. 3(a) for accidents, Fig. 3(b) for injuries, and 3c for fatalities. Table I shows the model values for each year based on the five features for the Assir region to predict A, I, and F.

TABLE. I. MODEL CASE RESULTS FOR THE ASSIR REGION SAMPLE

o. of Year	*Year*	*Predicted Accident Values*	*Predicted Injury Values*	*Predicted Fatality Values*
1	1434	38497	3379	978
2	1435	43603	4199	1021
3	1436	46601	5096	1063
4	1437	44758	5921	1102
5	1438	44240	6464	1132
6	1439	40959	6657	1140
7	1440	40038	6407	1113
8	1441	39742	5738	1032
9	1442	39743	4359	883
10	1443	39743	2913	788
11	1444	39743	2324	762
12	1445	39743	2171	774
13	1446	39743	2100	775
14	1447	39743	2000	784
15	1448	39743	1933	797
16	1449	39743	1876	797
17	1450	39743	1877	803
18	1451	39743	1895	810
19	1452	39743	1930	810

D. Representing Data by Histograms

The step of representing data using graphs has a Python code to create colorful histograms with legends to specify the degree of risk using Matplotlib and Plotly, which help stream the data to be analyzed and visualized. The histograms represent the real and predicted numbers of accidents, injuries, and fatalities, with colored legends to indicate the degree of risk in each region. Fig. 4(a) illustrates the histogram of real accidents for the Assir region in 1417–1433, while Fig. 4(b) shows the predicted values of accidents in 1434–1452 for the same region.

E. Representing Data on the Interactive Map

The phase of applying data visualization as an interactive map combines science and art. Using the map, the user can obtain information clearly and efficiently from information graphics constructed using meaningful information from data attributes or variables. The perfect template of data visualization will give effective results, aiding in data unification and analysis for the end user. The purpose of using data visualization is for promoting the eye's ability to differentiate extensive, diverse data and vast coherent information, serving a clear purpose in a small space. Web browser media support the designed information being displayed colorfully and interactively. The related information on road accidents, injuries, and fatalities is extracted to the interactive map, then visualized regionally according to the following entities:

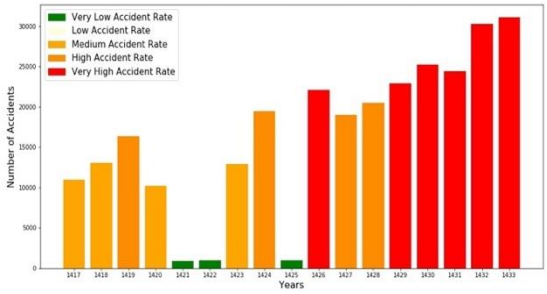
- The number of accidents, injuries, and fatalities for RTAs in 1417–1433.
- The number of accidents, injuries, and fatalities for RTAs in 1434–1452 (2030).

For data visualization, there are many different approaches and tools that can be employed. In this process, the Folium library (i.e., one of the Python libraries for interactive visualization) is used. The creation of an interactive map by Python, including a set of technical details, is illustrated in Fig. 5.

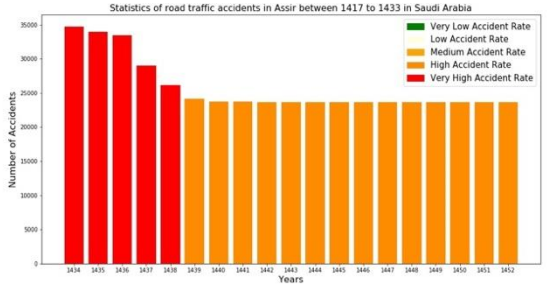
The visualizing phase represents the real and generated A, I, and F data and the results of predictions in the form of histograms on an interactive map using Python libraries like Folium. The interactive map is created using Python via the steps illustrated in Fig. 5. Folium is a Python library that allows visualizing the preprocessed data on Leaflet.js maps. This document covers and explains the process of making interactive choropleth maps from a given dataset on a map using the Folium library. Before visualizing the data on the KSA map, it is necessary to define the boundaries of the country and its regions to enable identifying the area on the world map being used. To obtain the GeoJSON file of a country, there are also a few steps that need to be taken:

- First, it is necessary to download the shape file of the geographical location that we are interested in. A shape file is a digital vector storage format for storing geometric location and associated attribute information. Country shape files can be downloaded from gadm.org.
- After downloading the shape file, it is necessary to convert it into a GeoJSON file for use with Folium to display all the region boundaries on the map. Converting the shape file to a GeoJSON file can be done through the MapShaper website. By uploading the shape zip file to the website, it is possible to export a GeoJSON file from the website of the same area/country.

The following screenshots of the KSA online map illustrate the related data visualizations for each region according to any of the following selected features: accidents, injuries, fatalities, future accidents, future injuries, and future fatalities; the related histogram appears after pressing the location icon of the region. In Fig. 6(a), the map represents the real numbers of accidents in the Assir region in 1417–1433. The same thing occurs when selecting the future accident case to represent accidents in Assir in 1434–1452, as shown in Fig. 6(b). As the samples show in Fig. 6(a) and 6(b), there is an identical histogram indicating the numbers of As, Is, and Fs separately for each region, covering the real and predicted cases. All these histograms are connected to the map internally to calculate the degree of risk, represented by the colored regions. The risk level for the map regions is based on the total numbers of As, Is, and Fs during the study years.



(a)



(b)

Fig. 4. (a). Risk Histogram of Real Accidents Statistics for Assir region in 1417-1434. (b). Risk Histogram of Predicted Accidents Statistics for Assir Region in 1417-1434.

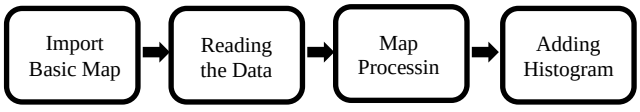
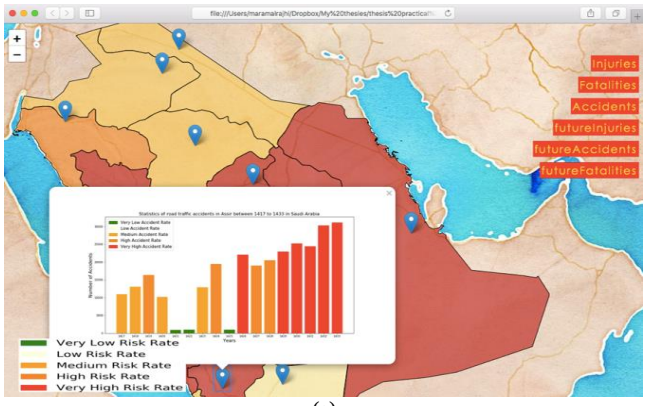
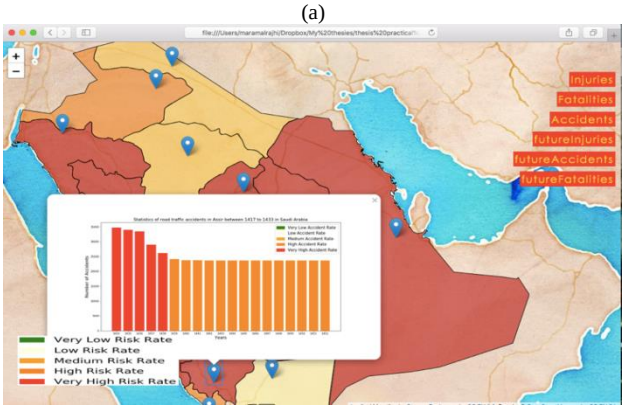


Fig. 5. Interactive Map Design Process.



(a)



(b)

Fig. 6. (a). Assir Case Illustrating the Data Visualization of Accidents Statistics in 1417–1433. (b). Assir Case Illustrates the Data Visualization of Predicting Accidents Statistics in 1434-1452.

IV. Evaluation

For representing information, it is important to maintain a high accuracy rate. This is crucial for evaluating the performance of the interactive map process. There are two approaches that are essential in evaluating the prototype process. The first approach involves measuring the accuracy of the predicted results. It is necessary to identify and estimate the efficiency and effectiveness of the LSTM with the dataset by calculating the mean squared error (MSE) or use other estimator tools for estimation purposes. Most of the evaluation metrics can be computed using Scikit-Learn in Python. The second approach is using feedback from decision makers or researchers as the end users of the whole process, focusing on how the interactive map can help for different purposes (improving decisions, use as an analytical reference, or for spreading awareness) by measuring criteria like perceived enjoyment, usefulness, and ease of use.

This paper checked the first approach by calculating the MAE, MSE, and root MSE (RMSE). Equations (5), (6), and (7) illustrate the syntax of the mentioned metrics:

$$MAE = \frac{\sum_{t=1}^{n}|A_t - F_t|}{n}, \quad (5)$$

$$MSE = \frac{\sum_{t=1}^{n}(A_t - F_t)^2}{n}, \quad (6)$$

$$RMSE = \sqrt{\frac{\sum_{t=1}^{n}(A_t - F_t)^2}{n}}, \quad (7)$$

TABLE. II. Error Rates for All Cases of Future Model

Error Rate	*Accident Case*	*Fatalities Case*	*Injuries Case*
MAE	3736216722981	848	1757
MSE	1340508538934528620363776	1233102	4180412
RMSE	318260542776	1110	2044

where A_t indicates the real accident, injury, or fatality values in a certain year t, F_t indicates the corresponding forecasted value for each A_t, and n represents the total dataset years. For the study model's evaluation, Python calculates the mentioned metrics for each case as in Table II.

The study faced data limitations in that it was forced to deal with annual data; just 14 datapoints were collected for each region, representing the number of years differentiating the study data. This resulted in the establishment of a small dataset to construct the time-series problem to be solved by deep learning. For future works, it would be better to consider cities or parts of cities instead of regions. In addition, for obtaining a better result with LSTM, it is important to have a long period of time to include hundreds or thousands of datapoints. For 14 datapoints, it would be better to have extra datasets on a monthly, daily, or hourly basis than the annual standard to apply LSTM perfectly with time-series problems. Perhaps, for the study dataset, it would be a good idea to combine another tool like a Convolutional Networks (CNN) with LSTM, as accomplished in [39], to improve the performance of short-term forecasting via a proposed new method called CNN & LSTM traffic flow prediction method (CLTFP) for forecasting future traffic flow data.

V. Utilization and Improvement

The proposed LSTM process can be applied online, particularly for statistical presentations, providing in-depth details and interactive representation. This is because data visualization offers an easy template that facilitates a comparison of different parts of the data, displaying a huge amount of data in a small area and a lot of coherent information to serve a clear goal. The internet is a public medium, and displaying information via web browsers is highly useful. We can see some popular newspapers and statistical and weather organizations using visual data representation to present informative reports and predictions.

The study process can be further enhanced. The procedure is practiced as a prototype for historical data to analyze, predict, and visualize the numerical data on accidents, injuries, and fatalities related to RTAs. In this tutorial approach, limited data have been used in regional-based research. The results can be made more dynamic by automatically processing these statistical data from different official web resources. Several factors can be added to improve deep learning, such as the weather, time of occurrence, and spatial location of incidents in relation to Saudi cities. One important factor is that only information about road accidents, injuries, and fatalities is extracted and visualized because of limitations on resources in the KSA case study; yet, there are still many critical details that should be included to evaluate the risk more accurately.

Concerning KSA Vision 2030, it is essential—as the KSA is an economically influential country—to be aware of death statistics every day. This whole process can be adopted by decision makers to reduce the considerable losses in health, social, and economic life, and it can be configured from our social responsibility. Limitations in the available data cause problems in evaluating the best prediction method to analyze them and indicate that the results will be inaccurate. This issue can be reconstructed for further enhancement of this process by trying other methods and expanding the dataset in more depth. Adding many related factors can be extremely helpful for enhancing the interactive map. In addition, working with live and spatial data can provide more useful and creative elements.

VI. Conclusion

The design of an interactive map prototype was described in this paper. The work presented a tutorial approach to illustrate the analytical and practical process of generating a prototype. Predicting the future of accidents, fatalities, and injuries was done based on historical data collected from some KSA websites. The collected data were saved in CSV files and converted to the sequence-to-sequence problem solved by LSTM in Python by Keras. Then, the real and predicted data were represented in ranked, colored histograms to act on the map in the form of combining the art and science to allow easy observation of risk. In the future, this process can be extended and evaluated more deeply after the disposal of the data limitation problem. New trends in data science can be added to this prototype to attract the attention of decision makers and the public to the concern of road accidents in the KSA.

References

[1] J. J. Rolison, S. Regev, S. Moutari, and A. Feeney, "What are the factors that contribute to road accidents? An assessment of law enforcement views, ordinary drivers' opinions, and road accident records," Accid. Anal. Prev., vol. 115, pp. 11–24, February 2018.

[2] A. A. Mohammed, K. Bin Ambak, A. M. Mosa, and D. Syamsunur, "Classification of traffic accident prediction models: a review paper," Int. J. Adv. Sci. Eng. Technol., vol. ISSN, no. 6, pp. 2321–9009, 2018.

[3] M. M. Islam, M. Alharthi, and M. M. Alam, "The impacts of climate change on road traffic accidents in Saudi Arabia," Climate, vol. 7, no. 9, pp. 1–13, 2019.

[4] N. T. Ratrout, S. Chowdhury, U. Gazder, and S. M. Rahman, "Characterization of crash-prone drivers in Saudi Arabia—a multivariate analysis," Case Stud. Transp. Policy, vol. 5, no. 1, pp. 134–142, 2017.

[5] C. Liu and A. Sharma, "Using the multivariate spatio-temporal Bayesian model to analyze traffic crashes by severity," Anal. Methods Accid. Res., vol. 17, pp. 14–31, February 2018.

[6] M. Al-zyood, "Forecast car accident in Saudi Arabia with ARIMA models," Int. J. Soft Comput. Eng., vol. 7, no. 3, pp. 30–33, 2017.

[7] C. Voyant et al., "Machine learning methods for solar radiation forecasting: a review," Renew. Energy, vol. 105, pp. 569–582, May 2017.

[8] J. Chen et al., "A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide," Environ. Int., vol. 130, September 2019.

[9] Y. Qin, D. Song, H. Cheng, W. Cheng, G. Jiang, and G. W. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," IJCAI Int. Jt. Conf. Artif. Intell., vol. 4, pp. 2627–2633, August 2017.

[10] M. Sameen and B. Pradhan, "Severity prediction of traffic accidents with recurrent neural networks," Appl. Sci., vol. 7, no. 6, p. 476, 2017.

[11] Z. Yuan, X. Zhou, and T. Yang, "Hetero-ConvLSTM: A Deep Learning Approach to Traffic Accident Prediction on Heterogeneous Spatio-Temporal Data," in Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2018, pp. 984–992.

[12] DE Rumelhart, GE Hinton, and RJ Williams, "Learning representations by back-propagating errors," Nature, vol. 323, pp. 533–536, 1968.

[13] G. P. Zhang, "Neural networks for time-series forecasting," Handb. Nat. Comput., vol. 4, pp. 461–477, 2012.

[14] J. Verdegaal, "External memory enhanced sequence-to-sequence dialogue systems," University of Amsterdam, 2018.

[15] K. Xu, H. Wang, and P. Tang, "Image captioning with deep LSTM based on sequential residual," in Proceedings—IEEE International Conference on Multimedia and Expo, no. 2017, pp. 361–366.

[16] C. Ma et al., "Visual question answering with memory-augmented networks," in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2018, pp. 6975–6984.

[17] Y. Tian, K. Zhang, J. Li, X. Lin, and B. Yang, "LSTM-based traffic flow prediction with missing data," Neurocomputing, vol. 318, pp. 297–305, November 2018.

[18] H. Ren, Y. Song, J. Wang, Y. Hu, and J. Lei, "A deep learning approach to the citywide traffic accident risk prediction," in IEEE Conference on Intelligent Transportation Systems, Proceedings, ITSC, November 2018, pp. 3346–3351.

[19] R. Trianto, T. C. Tai, and J. C. Wang, "Fast-LSTM acoustic model for distant speech recognition," in 2018 IEEE International Conference on Consumer Electronics, ICCE 2018, January 2018, no. 2, pp. 1–4.

[20] H. Akhter, "Information extraction and interactive visualization of road accident related news," Int. J. Comput. Appl., vol. 128, no. 5, pp. 37–40, 2015.

[21] S. Murray, Interactive Data Visualization for the Web, 2nd ed., Sebastopol CA, O'Reilly, 2017.

[22] S. H.-S. Brian Suda, "The 38 best tools for data visualization | Creative Bloq," Creative Blog Art and Design Inspiration, 2017. [Online]. Available: https://www.creativebloq.com/design-tools/data-visualization-712402. [Accessed: 22 February 2018].

[23] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, "Statistical and machine learning forecasting methods: concerns and ways forward," PLoS One, vol. 13, no. 3, pp. 1–26, 2018.

[24] H. P. Sascha Krstanovic, "Ensembles of recurrent neural networks for robust time series forecasting," in International Conference on Innovative Techniques and Applications of Artificial Intelligence, 2017, pp. 34–46.

[25] A. Srinivasan, "Near-time predictions of future truck locations," Chalmers University of Technology, 2017.

[26] R. Yu, Y. Li, C. Shahabi, U. Demiryurek, and Y. Liu, "Deep learning: a generic approach for extreme condition traffic forecasting," in Proceedings of the 2017 SIAM International Conference on Data Mining, 2017, pp. 777–785.

[27] H. Yu, Z. Wu, S. Wang, Y. Wang, and X. Ma, "Spatiotemporal recurrent convolutional networks for traffic prediction in transportation networks," Sensors (Switzerland), vol. 17, no. 7, 2017.

[28] N. A. Samat and L. W. Mey, "Malaria disease mapping in Malaysia based on Besag-York-Mollie (BYM) Model," J. Phys. Conf. Ser., vol. 890, no. 1, 2017.

[29] B. Elboshy, S. Kanae, M. Gamaleldin, H. Ayad, T. Osaragi, and W. Elbarki, "A framework for pluvial flood risk assessment in Alexandria considering the coping capacity," Environ. Syst. Decis., vol. 39, no. 1, pp. 77–94, 2019.

[30] E. G. Caldarola and A. M. Rinaldi, "Big data visualization tools: a survey: the new paradigms, methodologies and tools for large data sets visualization," in DATA 2017—Proceedings of the 6th International Conference on Data Science, Technology and Applications, 2017, pp. 296–305.

[31] B. B. Semlali and E. A. Chaker, "Towards remote sensing datasets collection and processing," Int. J. Embed. Real-Time Commun. Syst., vol. 10, no. 3, pp. 49–67, 2019.

[32] T. Cerquitelli et al., "Exploring energy performance certificates through visualization," in CEUR Workshop Proceedings, August 2019, vol. 2322.

[33] M. M. Ageli and A. M. Zaidan, "Road traffic accidents in Saudi Arabia: An ADRL approach and multivariate granger causality," Int. J. Econ. Financ., vol. 5, no. 7, pp. 26–31, 2013.

[34] I. Al-Turaiki, M. Aloumi, N. Aloumi, and K. Alghamdi, "Modeling traffic accidents in Saudi Arabia using classification techniques," in 2016 4th Saudi International Conference on Information Technology (Big Data Analysis), KACSTIT 2016, 2016, pp. 1–5.

[35] M. A. Alharbe, "Awareness ability and influences on raising of traffic accidents through the content of social media in the internet of things: a practical empirical study by the internet of things and multimedia on university students in western Saudi Arabia," in Proceedings of the International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud), I-SMAC 2018, 2019, pp. 48–51.

[36] N. Dogru and A. Subasi, "Traffic accident detection using random forest classifier," in 2018 15th Learning and Technology Conference, L and T 2018, 2018, pp. 40–45.

[37] A. S. Al-Ghamdi, "Time series forecasts for traffic accidents, injuries, and fatalities in Saudi Arabia," J. King Saud Univ.—Eng. Sci., vol. 7, no. 2, pp. 199–217, 2018.

[38] C. Olah, "Understanding LSTM networks," Blog, 2015. [Online]. Available: http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

[39] Y. Wu and H. Tan, "Short-term traffic flow forecasting with spatial-temporal correlation in a hybrid deep learning framework," vol. v1, pp. 1–14, December 2016.