University of Bahrain

College of Science

Department of
Postgraduate Programs

A Machine Learning Approach to Mitigate Transportation Traffic Congestions

Thesis submitted in partial fulfilment of the requirements for the degree of M.Sc. In Big Data Science and Analytics

Prepared by:
Aysha Khaled Mubarak
ID: 20092905

Supervised by:

Dr. Wael Elmedany
Associate Professor
University of Bahrain

Prof. Nabil M. Hewahi
Professor
University of Bahrain

Kingdom of Bahrain
September 2022

University of Bahrain

College of Science

Department of
Postgraduate Programs

A Machine Learning Approach to Mitigate Transportation Traffic Congestions

Thesis submitted in partial fulfilment of the requirements for the degree of M.Sc. In Big Data Science and Analytics

Prepared by:
Aysha Khaled Mubarak
ID: 20092905

Supervised by:

Dr. Wael Elmedany
Associate Professor
University of Bahrain

Prof. Nabil M. Hewahi
Professor
University of Bahrain

Kingdom of Bahrain
September 2022

Defence Committee's Approval

University of Bahrain
Deanship of Graduate Studies
And scientific Research

جامعة البحرين
عمادة الدراسات العليا والبحث العلمي

Thesis Discussion Committee Decision

Student Name: **AYSHA KHALED MUBARAK**	Student ID: **20092905**
Department: Choose an item.	College: **Science**
Program: **M.Sc. in Big Data Science and Analytics**	
Thesis Title: **A Machine Learning Approach to Mitigate Transportation Traffic Congestions**	
Supervisor Name: **Dr. Wael Elmedany + Prof. Nabil Hewahi**	Student GPA: **3.78**

The Discussion Committee met on 19-Oct-22 In S41-1002. Upon discussing the student whose details are stated above in public, the discussion committee recommends the following:

The student is awarded a grade Choose an item. in the thesis subject.

☐ Without making any changes to the thesis.

☒ After making the modifications attached with this decision.

Discussion Committee Members Approval:

Name		Signature
1. Dr. Wael Elmedany + Prof. Nabil Hewahi	(Advisor)	
2. Dr. Riadh Ksantini	(Internal Examiner)	
3. Dr. Adel Khelifi	(External Examiner)	Adel Khelifi

Approved at the University Council 6th meeting, 2013
In accordance with Decision No. 736/2013 dated 14/3/2013
Modified by University Council Decision No. 2043/2013 dated 20/11/2013.

Abstract

Despite the benefits of Intelligent Transportation Systems (ITS) in enhancing the performance of the transportation mechanize and facilitate the fast rhythm of the citizens' lifestyle, still there is a suffer from the traffic congestion. As with the increase in population and urban growth, the traffic congestion is increased as well. Therefore, this study is conducted to mitigate the traffic congestion using Machine Learning (ML) approaches. It is aimed to mitigate the traffic congestion in Kingdom of Bahrain by predicting the solution needed from construction and traffic engineering perspective. This research used a primary data, which is collected by the author from the reports and documents provided by the Ministry of Works who is responsible of the roads' construction in Bahrain. The extracted dataset has five class labels and were imbalanced classed, therefore, three oversampling methods were applied on the dataset. After that, three datasets result from each oversampling methods; Synthetic Minority Oversampling Technique (SMOTE), Adaptive Synthetic (ADASYN), Random Oversampling (ROS) and the fourth one is the hybrid and the combination of the three datasets. The models are built using five ML approaches; Decision Tree, Random Forest, k-Nearest Neighbours, Support Vector Machine and Multilayer Perceptron. All the models applied five times; one before applying oversampling and one for each oversampling method including the hybrid. The hybrid dataset produced the high accuracy among all the ML models. K-Nearest Neighbours and Random Forest resulted the best accuracy 100% and 99.9%. This study proved that the ML approaches are empowering the decision making in taking the suitable solution for road construction to mitigate the traffic congestion.

Dedication

I dedicate my thesis heartily and proudly to my amazing parents, my father Khalid who made me who I am today, my mother Muna for supporting me in every and single step, my beloved husband Abdulla for being my supportive backbone in this journey, my gorgeous twin daughters Eman and Muna, and my wonderful brother and sisters Mariam, Fatima and Amina for all the encouragement and endless support I got during doing my thesis.

Acknowledgements

From the conception of the topic until the final submission of the thesis, I have gained a lot of experience in many aspects. It has been a real challenge that has been both motivating and encouraging.

Many people have helped with this project, and it would not have been feasible to finish it without them. First and foremost, I want to thank the program coordinator, Dr Sawsan Hilal and my both supervisors, Dr Wael Elmedany and Dr Nabil M. Hewahi, for supporting my thesis and for their patience, excitement, encouragement, and scientific contributions that helped me finish this work.

Moreover, I would like to thank all the engineers in the Ministry of Works from the roads sector; Eng. Hosam Mohammad, Eng. Ishaq Akbar and Eng. Aysha Almansoori for their willingness to share their knowledge and experiences, as well as dedicating frequent meetings and opening their doors always for any help needed to complete this thesis.

Finally, I would like to express my gratitude to all my supporters; to name a few; Ahmad, Hajar, Amal, Isa, Latifa, Rana, Buthaina, Noor and Narjes for always boosting my confidence, and their support throughout this journey.

Table of Contents

Defence Committee's Approval ..... A

Abstract ..... B

Dedication C

Acknowledgements ..... D

Table of Contents ..... E

List of Tables ..... G

List of Figures ..... H

List of Abbreviations I

Chapter One Introduction to the Study 2

1.1. Introduction and Background 2

1.1.1. Transportation Traffic Overview 3

1.1.2. Intelligent Transportation Systems 4

1.1.3. Research's Machine Learning Approaches 5

1.2. Research Problem 7

1.3. Research Objectives 9

1.4. Research Significance 9

1.5. Research Structure 10

Chapter Two Literature Review 12

2.1. ML for Traffic Congestion Prediction 13

2.1.1. The Dataset Used for Traffic Congestion Prediction 13

2.1.2. ML Algorithms Used for Traffic Congestion 15

Chapter Three Methods and Procedures 28

3.1. Methodology 28
3.2. Research Study Population 28
3.3. Research Study Tools 29
3.4. Data Collection and Procedures 30
3.4.1. Data and Variable Extraction 31
3.4.2. Data Preparation 42
3.5. ML Approaches 45
3.5.1. ML Models 45
3.5.2. Evaluation Criteria 46
Chapter Four Results and Discussion 48
4.1. Results 48
4.1.1. Scenario 1: Results Before Oversampling 48
4.1.2. Scenario 2: Results After Oversampling 49
4.1.3. Scenario 3: Using the Hybrid Dataset 53
4.1.4. Summary of the Findings 54
4.2. Discussion 55
Chapter Five Conclusion 59
5.1. Conclusion 59
5.2. Limitations 61
5.3. Recommendations and Future Work 61
References 63
Appendices .....
الملخص .....

List of Tables

Table (1) Reasons for Traffic Congestions.....3
Table (2) Summary of Literature Review.....21
Table (3) Review for ML models.....23
Table (4): Variables Description.....32
Table (5) Lane Configuration Arrows Equivalence.....38
Table (6) Decisions Description.....42
Table (7) Variables Datatype After Data Cleaning.....43
Table (8) Models' Parameters.....46
Table (9) Results of Fitted Models in Scenario 1.....49
Table (10) Results of Fitted Models in Scenario 2 - SMOTE.....50
Table (11) Results of Fitted Models in Scenario 2 - ADASYN.....51
Table (12) Results of Fitted Models in Scenario 2 - ROS.....52
Table (13) Results of Fitted Models in Scenario 3.....54

List of Figures

Fig. 1: Number Of Publications On ITS, Including ML-Based Techniques (2010 To 2020) (Yuan et al., 2021).....12

Fig. 2: Schematic Diagram of implemented methodology.....30

Fig. 3: Map Sample of Four Junction.....34

Fig. 4: Map Sample of One Junction.....34

Fig. 5: Field Movement Count.....35

Fig. 6: Full Report of J1-AM in Synchro.....36

Fig. 7: Part of Synchro Report for Muharraq Ave/SZH Junction in the Three Peaks.....36

Fig. 8: Part of Sidra reports.....37

Fig. 9: Left Map – SHZ with Hwy71 Junction latest Satellite Map, Right Map - SHZ with Hwy71 Junction Satellite Map on 2015.....39

Fig. 10: Part of Synchro Report - Intersection Summary.....40

Fig. 11: A sample of an overall ranking for the solution options to enhance the junction.....41

Fig. 12: Visualising the Imbalanced Classes in the Dataset.....44

Fig. 13: The Dataset's Instances and Attributes After SMOTE.....49

Fig. 14: Visualising Dataset's Classes After SMOTE.....50

Fig. 15: The Dataset's Instances and Attributes After ADASYN....51

Fig. 16: Visualising Dataset's Classes After ADASYN.....51

Fig. 17: The Dataset's Instances and Attributes After ROS.....52

Fig. 18: Visualising Dataset's Classes After ROS.....52

Fig. 19: The Hybrid Dataset's Instances and Attributes.....53

Fig. 20: The Hybrid Dataset's Classes Visualizing.....53

Fig. 21: The Accuracy of the Models Using all the Datasets.....55

List of Abbreviations

Term	Full Form
ITS	Intelligent Transportations Systems
ML	Machine Learning
SVM	Support Vector Machine
DT	Decision Tree
RF	Random Forest
MLP	Multi-Layer-Perceptron
KNN	K-Nearest Neighbours
SMOTE	Synthetic Minority Oversampling Technique
ROS	Random Oversampling
ADASYN	Adaptive Synthetic
RMSE	Root Mean Square Error
MOW	Ministry of Works, in Kingdom of Bahrain
LGV	Light Goods Vehicles
HGV	Heavy Goods Vehicles
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Units
ARIMA	Auto-Regressive Integrated Moving Average
CNN	Convolutional Neural Network

Chapter One

Introduction to the Study

1.1. Introduction and Background

1.2. Research Problem

1.3. Research Objectives

1.4. Research Significance

Chapter One
Introduction to the Study

1.1. Introduction and Background

The transportation systems are developing and improving year after another in parallel with evolving smart technologies to facilitate the best traffic flow on the road. The smart and intelligent technologies that are applied in the transportations field are referred to as "Intelligent Transportation Systems" (ITS). ITS has been initiated by a group of transportation professionals in the 80s, when they started observing the impact and the efficiency of computing and communications revolution and the information power on the transportation systems. Moreover, ITS was established in the beginning to be used for vehicles, which is used to increase the efficiency and the safety not only for people, but also for goods (Weiland & Purser, 2000). ITS is the umbrella that can cover all the technological tools and automated systems used in the transportation sector. The aforementioned became a necessity with the fact of having wasteful aspects cause gigantic misfortunes of time, diminish within the level of security for both vehicles and pedestrians, high contamination, and the corruption of quality of life. Therefore, ITS is meant to maximize the transportation efficiency in handling all vehicles traffic conditions in different and various manners in normal and abnormal days, in real time or traffic patterns, and also used for centralizing the legacy traffic management and assessment systems (Dimitrakopoulos & Demestichas, 2010).

1.1.1. Transportation Traffic Overview

Traffic is known to be the accumulation of vehicles (motorized and non-motorized) that use common infrastructure to get from one location to another. Cars driving down roads and highways, planes flying in and out of airports, ships sailing following defined itineraries, and so on (Designing Buildings, 2022).

The traffic congestion is neither an issue for the planes flying in the sky nor an issue for the ships cruising the waters. Nonetheless, the terrestrial traffic congestion is a pressing issue. Table (1) illustrates the general causes leading to traffic congestions (Wenzel, 2015).

Table (1) Reasons for Traffic Congestions

Traffic Congestion Causes	Description	Percentage
Bottleneck	When motorists cannot efficiently use the road	40%
Traffic Accidents	Car accidents, damaged automobiles, and the unexpected presence of hazardous objects are examples of these	25%
Force Majeure	Such as negative weather conditions (snow or hail, extreme wind, etc.)	15%

Human and Pedestrian Factors	Human factors influencing pedestrian walking and crossing behaviour include mobility characteristics and travel motivations, risk perception and time value, attitudes toward walking and related preferences, walking and crossing behaviour and adherence to traffic rules, self-assessment, and driver opinion.	20%

As the traffic congestions issue aggravated over time, a sub discipline namely transportation engineering has emerged under the civil engineering to address this problem (Institute of Transportation Engineers, n.d.).

1.1.2. Intelligent Transportation Systems

ITS in the smart cities are an essential part of urban planning and future smart cities that contributes towards enhanced road safety, transportation, and transit efficiency, as well as to reduced environmental pollution. It uses different technical tools to enhance the transportation process. It basically refers to the utilization of information and communication technologies in transport systems. The technical tools are used for many applications and are not limited to the congestion, such as accident analysis, social analysis, commercial analysis, etc. ITS improves mobility and safety along with the fuel consumption. This is done without the need to expand the physical infrastructure.

As the Information and Communication Technology (ICT) sector evolves, ITS took the advantage of such evolution. Searching, exchanging and retrieving information is a matter of a click with the exponential growth

on the internet that is adequately complemented by the sound wireless infrastructure.

Despite the technical developments in ITS, it is very early to announce the maturity of such systems. ITS is of substantial complexity, hence there is great room for further development and research in this field. Professionals still lack the holistic understanding of how the traffic behaves overtime, how to best gather data and present insightful information, how to optimize traffic with certain budgetary and social constraints, etc. (Yokota et al., 2005).

1.1.3. Research's Machine Learning Approaches

In this research, several of Machine Learning (ML) approaches are used to empower the findings of this study. These ML approaches are Support-Vector Machines (SVM), K-Nearest Neighbours (KNN), Decision Tree (DT), Random Forest (RF) and Multi-Layer-Perceptron (MLP).

SVM is considered as a supervised model which works with classification problems. The data points in the dataset are classified by a hyperplane built as n-dimensional space to split the data points classes. The support vector points to increase the distance between classes to the maximum. SVM has the ability of handling the non-linearly separable problems using kernel function. This function is converting the problem to input linearly separable problem through turning the input space into higher dimensional space, which is a significant benefit SVM has, to improve the classification outcomes. Alternatively, SVM performance decreases when deals with large datasets (Support Vector Machine Algorithm, 2021).

Moreover, KNN is a non-parametric model that is frequently utilized in both classification and regression issues. It does not immediately execute the classification; instead, it saves the data (Pandey & Jain, 2017). KNN selects the number (k) of the closest neighbours that will be taken into account while classifying a particular point. Based on the shortest distance between the provided place and those neighbours, those neighbours will be chosen. The Euclidean Distance function is used to determine this distance. The most frequent class of the selected neighbours is then given the provided point. This algorithm's disadvantage is that, as k is the only hyper-parameter, it is significantly influenced by the value that is selected. However, it works well with sizable datasets (JavaTpoint, 2011).

Furthermore, the decision tree method is intuitive because it mimics how people make decisions. Whether one has discrete or continuous data as input, it can solve in both cases. Basically, it operates by deciding when to split a node using a variety of techniques. It can generate a sub-node that is nearer the target variable by splitting the node. Such a method doesn't call for any feature transformation. It doesn't rely on the spatial distribution assumptions or the structure of the classifier because it is a non-parametric method. This method's disadvantage is that it is quite vulnerable to overfitting. Large datasets are also not a good fit because they lead to complex models (Harsh, 2018).

Another ML model called Random Forest can be used to address classification problems as well as regression problems. It operates by selecting random samples with replacement and aggregating numerous base classifiers (Decision Trees). The benefit of this method is that it lowers the correlation between the predictors, which in turn lowers the variance among the trees. For classification prediction it uses majority vote or average are used for regression prediction. The non-parametric nature of RF allows for the avoidance of incorrect data assumptions. It also

detects outliers and requires less data cleaning. If there are a lot of trees, it can take a very long time (Donges, 2021).

In addition, the term "Neural Network Approach" is also used to refer to the Multilayer Perceptron (MLP) model. Input layer, hidden layer(s), and output layer are the components of this model. The weights located between the layers affect the value that will be passed to the following layer. This model is intended to address issues that linear regression is unable to address, and as a result, it has the advantage of being adaptable enough even to resolve problems with non-linear forms. After training, it can also analyse huge datasets with quick predictions. On the other hand, MLP is regarded as a "black box" when it comes to figuring out how the prediction processing works. There are numerous subtypes of MLP, such as the Convolutional Neural Network and Radial Basis Functional Neural Network (IBM Cloud Education, 2020)

1.2. Research Problem

The main problem that ITS is meant to solve is the traffic congestion problem. When a traffic problem happens due to some factors such as the increase in the population density and urbanization, the traffic engineers need to monitor and collect a lot of information from different places and they might also need a site physical visit to analyse this information. Thereafter, the engineers put set of proposed solutions to solve the congestion problem. This process might take a quite long time which may take years in some projects to implement the suitable solution for the main problem that causes the traffic congestion.

This research is focusing on finding the best solution for the junction that suffers from a traffic congestion using ML techniques. This research

has covered all the junctions in Kingdom of Bahrain which are suffering from the high values of delay that leads to traffic congestion. The best solution is the constriction decision to be applied on the junction based on the important factors fed to the model. This research is seeking to find the best solution using several ML approaches.

1.3. Research Objectives

The main objectives of this research are:

1. Collect the data of all the junctions in the Kingdom of Bahrain with standardized variables that affect the traffic congestions the most and common for all the junction's types.
2. Apply ML approaches on the junctions' data to help the traffic administration to take the proper decision to solve the possible traffic congestions in shorter time.
3. The ability to determine the future enhancements needed by predicting the suitable solutions from the forecasted inputs.

1.4. Research Significance

This research is focusing on finding the best decisions to solve traffic congestions based on the junction's information like the number of vehicles, number of lanes, the amount of the delay in all the three peaks with its control type. Other researches were mostly focusing on predicting the traffic congestions but not on its solution from traffic engineering perspective. The data used in this study is a primary data which is collected by the author from the traffic engineering projects reports and maps. To the knowledge of the researcher, this kind of study has not been done before. This research is meant to support decision making in this field and to narrow down the time spent by the traffic engineers on congestion relieving projects.

1.5. Research Structure

This research has five primary chapters, and each one explains a certain section in great detail as follows:

- **Chapter 1:** This work's introduction chapter provides a background on the ITS, its benefits, and the rationale for applying ML on dealing with the traffic congestion challenges. The research problem's definition, the research objectives, and its significance.
- **Chapter 2:** This chapter titled as Literature Review, provides a review of ML applications used in ITS field generally and the ML approaches used in traffic congestion problem along with the datasets used, followed by the contribution of this research work.
- **Chapter 3:** This is the Methodology chapter, which mentioned the study's population with the tools used, and explains in details the data collection and procedures which includes the data and variables extraction and data preparation. It ends with the ML approaches used in this study with the evolution criteria.
- **Chapter 4:** This chapter, which is entitled Results and Discussion, presents the experimentation results obtained from different scenarios with using three oversampling methods and five ML algorithms. In the discussion, the results are compared and discussed.
- **Chapter 5:** The Conclusion chapter, which is the final chapter, which summarizes the work done in this research work, the limitation faced due the research, and the recommendations and future work.

Chapter Two

Literature Review

2.1. ML Applications in ITS Section

2.2. ML for Traffic Congestion Prediction

Chapter Two

Literature Review

The following section briefly demonstrates an overview of articles and resources that are nearly as similar nature to this research. This literature review touches upon two themes of ML used in ITS and the ML used specially in traffic congestion.

The ML algorithms and techniques have recently played a major role in empowering ITS. Number of ITS related publications that are based on ML techniques has exploded starting 2017 as shown in Fig. 1. This section touches upon how the ML is being applied to wide range of ITS services.

Deep learning and reinforcement learning are two ML techniques that have been beneficial in exploring patterns and underlying structures in big data sets for prediction and precise decision making (Yuan et al., 2021).

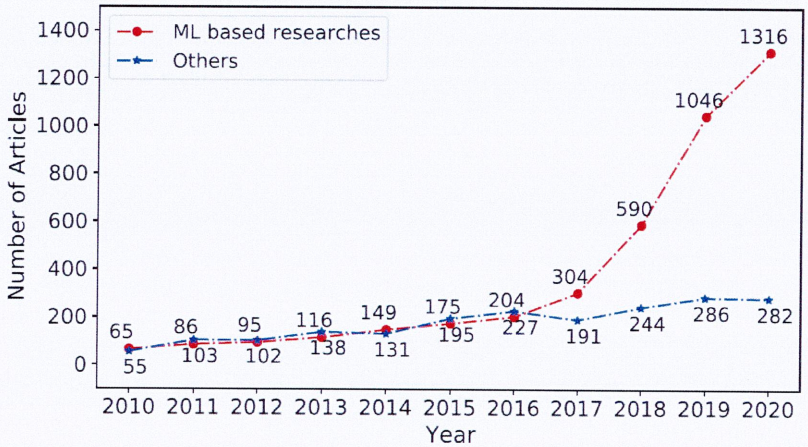


Fig. 1: Number Of Publications On ITS, Including ML-Based Techniques (2010 To 2020) (Yuan et al., 2021)

There is an unlimited list of applications used ML in ITS, and the traffic congestion prediction is one of most applications used different types and combinations of ML algorithms which is mentioned in details in the next section.

2.1. ML for Traffic Congestion Prediction

The effectiveness of using ML in predicting the traffic congestion is very high. However, the ML methods and algorithms used are differ, even the type of the data, there is a variety of datasets used in the ML models. In this section, the types of datasets and the ML algorithms are discussed in details.

2.1.1. The Dataset Used for Traffic Congestion Prediction

In the field of traffic congestion prediction, there is a plenty of research around the world, which makes a variety of datasets used as well as the methods. The datasets are varying in its availability, there are confidential and private datasets, free public datasets and payable public datasets.

The private dataset is owned by a governmental or private organization. Likewise, a congestion prediction used for smart sustainable cities by applying ML approaches with IoT, this model has been applied on a private dataset about vehicle speed measurements were owned by Matlock County Hall (Majumdar et al., 2020).

Alternatively, the public datasets are available for all the researches to study it. GPS trajectory data is one of the forms of datasets publicly published and used for traffic prediction (Sun et al., 2019). Similarly, Kamble & Kounte (2020) used GPS trajectory data for traffic congestion monitoring system using ML approaches in Internet of Vehicles. There are vehicle trajectories used for ML Approach to Short-Term Traffic Congestion Prediction, these trajectories are available through connected vehicles technology (Elfar et al., 2018).

The most common payable public datasets, are the ones extracted using Google Maps API, it is extracted based on the required information using queries (Google Maps Platform, 2022). Similarly, Zafar & Ul Haq (2020) build a model to predict the traffic congestion based on arrival estimated time and the traffic data which is collected from Google Map API.

In general, the traffic datasets are including stationary and probe data, can be broadly categorized into two categories for the purposes of various investigations. Sensor data and fixed cameras are additional categories for stationary data. On the other hand, the studies' probe data consisted of GPS information from mounted vehicles (Akhtar & Moridpour, 2021).

The Performance Measurement System (PeMS), is categorized as stationary data which collects highway data on traffic flow, sensor occupancy, and travel speed in real-time across all significant metropolitan areas of the State of California, was the most widely utilized dataset. Bai et al., 2020, used PeMS dataset in traffic congestion prediction for the smart cities with relative position congestion tensor. Similarly, Kishk et al. (2021) also utilized the PeMS dataset for Intelligent Traffic Management Systems (ITMSs), which take the place of conventional systems and aim to increase traffic efficiency and lessen congestion, can be made possible by real-time accurate traffic congestion prediction. Other researchers, used their suggested DNN-BTF model (Deep Neural Networks – Based Traffic Flow) and tested it using PeMS on a long-term horizon prediction challenge. The strategy developed by the authors outperforms state-of-the-art approaches, according to experimental results (Wu et al., 2018).

There are other factors that may affect the traffic congestion, like the weather. Hou et al. (2021) used Deep Learning algorithms and data fusion to predict the short-term traffic flow with weather conditions. There are also other datasets used rather than the weather, which can be the data sent

from IoT devices, for real-time traffic congestion prediction (Kishk et al., 2021).

2.1.2. ML Algorithms Used for Traffic Congestion

With enormous number of researches on traffic congestion using ML algorithms, there is a wide range of ML approaches had been covered. This section providing the list of these approaches briefly.

Majumdar et al. (2020) aimed at analysing the flow of traffic in the UK to predict the propagation of congestion through deep learning models. Scholars utilized the vehicle speed data that was provided by the Matlock County Hall, to train the univariate and multivariate LSTM models. The univariate LSTM model used the observed speed limit as input, and the multivariate considers the traffic flow rate and vehicle headway as parameters in the prediction. After testing the performance of the model, analysis showed that both models gave accuracy of 84%-95%, depending on the road layout.

Researchers aimed at analysing the performance of deep learning and traditional ML models to accurately predict the traffic congestion level based on processed GPS trajectory. The following models were trained on the taxi GPS trajectory data, CNN, RNN, LSTM, GRU, ARIMA, and SVM. Authors concluded that deep learning models obtained higher accuracy than the traditional ML algorithms (Sun et al., 2019).

Kamble & Kounte (2020) were interested to identify congestion based on live traffic prediction in real-time based on ML models. Scholars implemented the Gaussian technique in the ML models to predict short-term traffic using historical data. Researchers concluded that ML can

provide reliable prediction of traffic speed, in addition the use of Gaussian process in the ML models enhanced their performance further.

Elfar et al. (2018) aimed at predicting the traffic congestion environment on a Short-Term horizon. In addition, scholars further expanded the experiment to understand the relationship between perturbations and shockwave formation, to enhance the prediction of congestion formation. Using the vehicle trajectories dataset which were made available through connected vehicles technology, authors utilized logistic regression, random forests, and neural network to predict the congestion of traffic. Authors experimented these models firstly on historical data and are trained again if there is any significant change occurs in the system, and another time using the historical data which are updated on real-time information on prevailing traffic conditions obtained through V2V/V2I communications. Analysis showed that the accuracy of the models reached to 97% accuracy, which indicate that can be further enhanced when integration with traffic control systems.

Bai et al. (2020) proposed a model which can predict the traffic congestion in smart cities using the relative position congestion. Since authors notes that in past literature, most of the models that predict the traffic congestion are based on specific location. However, this is not a flexible approach since the model need to be updated every time the location changes. Accordingly, researchers proposed to build a Position Congestion Tensor for traffic. They used dataset which included historical speed series and real-time speed data to calculate the congestion index, from which it used to merge the road network map to construct the congestion metrics. Authors used the relative locations for road nodes to convert matrices into three-dimensional spatial-temporal tensors. After

which several ML models were trained and tested their performance. After analysing the results, it was concluded that the proposed approach and model can accurately capture the temporal and spatial characteristics of traffic. Further the LSTM model outperforms the rest of the other ML models in all locations.

Kishk et al. (2021) proposed a new strategy based on the edge layer called Traffic Congestion Prediction Strategy (TCPS). It is based on the collected data in the edge's cache server at this layer. The first stage of the strategy's three stages involved real-time congestion prediction (RCP), which used a fuzzy inference system to predict traffic congestion based on the reasons of congestion in the hotspot. Second, the congestion direction detection (CD2) stage uses the Optimal Weighted Naïve Bayes (OWNB) approach to determine the direction of congestion based on predictions from the RCP if there is congestion. After determining the direction of congestion in CD2, the width change decision (WCD) step is used to adjust the width of changeable routes (CR) in order to avoid congestion occurring. The results of the experiments have demonstrated that the proposed TCPS works better than other contemporary techniques. The highest accuracy, precision, and recall are offered by TCPS. Additionally, it offers the least amount of error, with values of 95%, 74%, 75%, and 5%, respectively.

Hou et al. (2021) used to predict traffic flow, a combined framework of stacked auto-encoders (SAE) and radial basis functions (RBF) neural networks that is proposed in this research. This framework efficiently captures the temporal correlation, periodicity, and disturbance of meteorological conditions in traffic flow data. To begin, SAE is used to process traffic flow data in various time slices in order to obtain a preliminary prediction. The relationship between weather disturbance and

traffic flow periodicity is captured using RBF to produce a new prediction. In order to combine the two predictions mentioned above at the decision level and produce a rebuilt forecast with greater accuracy and robustness, a second RBF is needed.

Lv et al., (2014) proposed a unique deep learning-based traffic flow prediction approach that naturally takes into account the spatial and temporal correlations. Generic traffic flow features are learned using a stacked auto-encoder model that is trained greedily layer-by-layer. Experiments also show that the proposed strategy for predicting traffic flow performs better.

To anticipate traffic flows, Polson & Sokolov (2017) created a deep learning model whose architecture incorporates a linear model fitted with l_1 regularizations and a series of tanh layers. They demonstrated how these nonlinear spatial-temporal effects can be captured by deep learning frameworks. Predictor relationships are identified in the top layer, which also models nonlinear relationships. They used data from road sensors to demonstrate their methodology and forecast traffic on two special occasions. Both studies involved abrupt, fast shifts in the traffic flow regime, and they demonstrated how deep learning can produce accurate predictions of the short-term traffic flow.

It is crucial to create traffic control models that can manage large amounts of traffic data and synchronize traffic lights in an urban network in real time without interfering with other activities. Borges et al. (2021) suggested a deep reinforcement learning-based approach for synchronizing traffic signals in a network of two intersections in an urban area. Utilizing actual traffic data gathered at each intersection's approach, this model's

calibration and neural network training were carried out. The simulations' outcomes were quite encouraging, producing notable gains in metrics measured in comparison to the network's baseline circumstances. In times of high demand, the model decreased delays and queue lengths by up to 50% and could handle a wide range of traffic flows. The model could handle a wide range of traffic patterns and, during times of high demand, cut back on delays and queue lengths by more than 28% and 42%, respectively.

Online open data now contain a wealth of information on traffic. Official websites for traffic management and operations, web-based map services like websites for weather forecasts, websites for local events (sporting events, musical performances, etc.) and Google Maps are examples of these resources. Yisheng et al. (2016) used publicly available web data to provide statistics on traffic. To show the viability, traffic data gathered from web-based map services is employed. To learn and forecast the patterns of traffic conditions, a type of deep architecture called the stacking long short-term memory model is used. The suggested model for predicting traffic conditions outperforms multilayer perceptron models, decision tree models, and support vector machine models, according to experimental results. Similarly, Zafar & Ul Haq (2020) used Google Maps API dataset and conducted a model based on ML approached and the results showed that, with a prediction accuracy of 92%, the random forest classification method is the most accurate, followed by XGBoost and KNN.

On the basis of data mining and ML, this research provides a methodology for tackling the problem of accurate trip time estimation based on dynamic traffic data. Kumarage, et al. (2017) suggested a model integrates dynamic and concurrent traffic data gathering on various links

based on Google Maps Distance Matrix Application Programming Interface rather than relying on the aforementioned reference vehicle journey time data. A thorough estimation of travel time and the establishment of bottlenecks is determined using streaming data given into ML model. Within an experimental workbench, the suggested approach functioned well, and the model using SVM is continually improving as it gains experience, producing better computations over time.

Another study proposes a sort of fully-connected Deep Neural Network called Supervised Deep Learning Based Traffic Flow Prediction (SDLTFP) (FC-DNN). Reliable traffic flow prediction is also greatly hampered by timely predictions. A deep network's training process could take a while, and overfitting could occur, especially when feeding the deep architecture with little amounts of data. The network learns everything well during training, but when tested with fresh data, it might not be able to generalize the model. The authors use the Dropout and Batch Normalization (BN) approaches to aid in network training. The weight is updated using SGD and momentum. then used publicly available data, such as past traffic statistics, to forecast the future. According to their experiments, the traffic flow prediction's Mean Absolute Percentage Error (MAPE) is less than 5% when utilizing sample data and between 15% and 20% when using data that is not in the sample (Tampubolon & Hsiung, 2018)

Table (2) is summering the previous studies and providing the dataset used with the ML approach they used in their models.

Table (2) Summary of Literature Review

Research	Dataset Used	ML Approach
Majumdar et al. (2020)	Vehicle speed measurements were owned by Matlock County Hall	univariate and multivariate LSTM models
Sun et al. (2019)	GPS trajectory data	CNN, RNN, LSTM, GRU, ARIMA, and SVM.
Elfar et al. (2018)	GPS trajectory data	logistic regression, random forests, and neural network
Kamble & Kounte (2020)	GPS trajectory data	Gaussian technique
Zafar & Ul Haq (2020)	Google Maps API	random forest classification, XGBoost and KNN.
Kishk et al. (2021)	The Performance Measurement System (PeMS) & IoT devices data	fuzzy inference system , Naïve Bayes,
Wu et al. (2018)	The Performance Measurement System (PeMS),	DNN-BTF model (Deep Neural Networks – Based Traffic Flow)
Hou et al., (2021)	Weather data with traffic private data	Deep Learning algorithms and data fusion; stacked auto-encoders (SAE) and radial basis functions (RBF) neural networks

Lv et al. (2014)	Traffic flow dataset	Deep Learning algorithms
Polson & Sokolov, (2017)	Data from road sensors	Deep Learning algorithms
Borges et al. (2021)	Traffic signals in a network dataset	a deep reinforcement learning-based
Yisheng et al. (2016)	Google Maps API	multilayer perceptron models, decision tree models, and support vector machine models
Bai et al. (2020)	The Performance Measurement System (PeMS)	LSTM
Kumarage, et al. (2017)	Google Maps Distance Matrix	SVM
Tampubolon & Hsiung (2018)	Traffic data and the weather data from Open Data Taipei City	Fully-Connected Deep Neural Network (FC-DNN)

A different summary Table (3) presented more than 30 researches using ML methods in traffic congestion and showing the model name and the ML methods used. It is listed that most of the researches used LSTM as the first place with more than 15 researches and in the second place CNN with 15 papers (Razali et al., 2021).

Table (3) Review for ML models

Research	Model Name	ML Approach
Tu et al. (2021)	SG-CNN	CNN
Zhang et al. (2021)	Multi-city Traffic Flow Forecasting Network (MTN)	CNN, LSTM
Hou et al. (2021)	SAE-RBF Framework	Sparse Autoencoder (SAE),) Radial Basis Function (RBF)
Xia et al. (2021)	WND-LSTM	LSTM
Romo et al. (2020a)	Traffic Speed Prediction Framework	CNN, LSTM, XGBoost
Abdelwahab et al. (2020b)	Traffic Congestion Classification based on compact image representation	CNN
Abdellah and Koucheryavy (2020)	IoT traffic prediction	LSTM
Wang et al. (2020)	Multitask Deep Learning Model (MTGCN)	Graph Convolution Network (GCN)
Lin, Wang, et al. (2020)	Traffic flow prediction model(LSTM_SPLSTM)	LSTM
Qu et al. (2020)	Local and Global Spatial Temporal Network (LGSTN)	LSTM
Wang et al. (2020)	Improved traffic state identification	Graph Convolution Network (GCN)

Shin et al. (2020)	LSTM-based traffic flow prediction	SVM, Rand Fuzzy C-Means m Subspace, Fuzzy C-Means
Ranjan et al. (2020)	Hybrid neural network for the purpose of spatial and temporal information extraction	LSTM
Liu et al. (2020)	Deep Learning Network	CNN, LSTM
Elleuch et al. (2020)	Intelligent Traffic Conges- tion Prediction System using Floating Car Data (FCD)	Transpose CNN
Jingjuan Wang and Chen Varying spatiotemporal (2020)	graph-based convolution model (VSTGC)	LSTM, Graph Convolution Network (GCN)
Essien (2020)	Deep learning urban traffic prediction model combined with tweet information	CNN
Ren and Xie (2019)	Transfer Knowledge Graph Neural Network (TKGNN)	ANN, Deep Neural Network (DNN),Recurrent Neural Network (RNN)
Chou et al. (2019)	Deep Ensemble Stacked Long Short-Term Memory (DE-LSTM)	Graph Convolution Network (GCN)

Yi and Bui (2019))	Vehicle Detection System (VDS	LSTM
Xu et al. (2019)	End-to-end neural network named C-LSTM	Sparse Auto-encoder (SAE)
Jingyuan Wang et al. (2019)	Deep urban traffic flow prediction (DST) based on spatial temporal features	CNN
Yang et al. (2019)	CNN-based multi-feature predictive model (MF-CNN)	LSTM
Chen et al. (2019)	Multiple residual recurrent graph neural networks (Mres-RGNN)	LSTM, Recurrent Neural Network (RNN)
Bartlett et al. (2019)	ML method comparative study	CNN, LSTM
Xu et al. (2018a)	Treating network status as a video for prediction of congestion level	CNN, LSTM
Shirazi and Morris (2018)	Feature collection system	CNN
Tampubolon and Hsiung (2018)	Supervised Deep Learning Based Traffic Flow Prediction (SDLTFP)	CNN
Jin et al. (2018)	Spatio Temporal Recurrent Convolutional Network (STRCN)	Graph Convolution Network (GCN), Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU)

Duan et al. (2018)	Deep hybrid neural network	ANN,MR
Chen et al. (2018)	Fuzzy Deep convolutional Network (FDCN)	LM
Kong et al. (2018)	Intelligent Traffic Recom-mendation System	CNN
Tian et al. (2018)	Traffic flow forecasting	Deep Neural Network (DNN)
Khan et al. (2017)	Framework that integrates CV with AI	Deep Neural Network (DNN)
Lawe and Wang (2016)	Deep-learning neural network for traffic flow optimization	CNN, LSTM
Wang et al. (2016)	Traffic Condition Estimation Integrated with GPS and tweet	CNN, LSTM

To the best of our knowledge, most of the researches are focusing on the traffic congestion prediction using ML, while there is no research done for predicting the solution to mitigate the traffic congestion from construction perspective for the junctions and roads. Furthermore, this study is conducted for the first time in Kingdom of Bahrain and the dataset used is primary data and collected by the author.

Chapter Three

Methods and Procedures

3.1 Methodology

3.2 Research Study's Population

3.3 Research Study's Tools

3.4 Data Collection and Procedures

3.5 ML Approaches

Chapter Three
Methods and Procedures

3.1. Methodology

This research has adopted an exploratory research design using ML approaches in which it assesses the traffic management system in Bahrain and helps the traffic engineers with treating the frequent traffic congestions in specific types of roads and junctions.

This research is built based on the process of studying the traffic congestion in the traffic studies directorate in the Ministry of Works, Bahrain (MOW). Hence, all the procedures here are extracted from what is happening in the real field and based on the traffic engineers practice, rules and experience. The following sections the study population, study tools, followed by explaining the rules, the full data collection process and ends with the processing models using ML approaches used in this study.

3.2. Research Study Population

The population considered in this study is all the junctions in the Kingdom of Bahrain. The Ministry of Works and specifically the roads sector is responsible of developing the roads in Bahrain. There are 56 junctions in Bahrain, these junctions are varying in the traffic control (Un-signalized, Signalized, Roundabout and Grade-Separated). The ministry is working on the junctions' development as projects. This is applied on the most critical areas with high traffic that includes specific junctions network. These projects cover the areas fully for more than one year and ends with set of solutions. There are also annual reports that covered all the

junctions in Bahrain with the final solution applied on them. This study covered 6 projects reports and 2 annual reports. These reports pertain to the period between January 2014 and November 2020. It is important to mention that these reports represent all the available reports in the ministry until the date of this study. The reports are not limited to the existing condition, but it also includes the forecasted values related to the junctions for the year 2030. These values are already included in this study as well.

3.3. Research Study Tools

There are set of tools used in this research, which are:

1. Microsoft Excel: used for data entry initially, then saved the full data as CSV format for processing.
2. Python in Google Colab: used for data cleaning and oversampling.
3. Google Earth Pro: used for viewing the junctions, counting the lanes and identify the control type. It has the feature of viewing the map on satellite mode chronologically.
4. WEKA v.3.8.4: used for classification ML models.
5. Adobe PDF reader: used for reading the engineering reports.

3.4. Data Collection and Procedures

The data in this research is primary data as it has been collected and gathered directly by the author. The process of collecting the data has been approved by the traffic engineers in MOW. This section is explaining, in details, all the data collection procedures and steps.

Fig. 2 illustrates the steps of the implemented methodology that is began with (Data Extraction), then (Data Preparation) which consist of (Attribute Selection) then (Data Cleaning) and followed by (Data Oversampling). Data oversampling is done using three methods and the fourth one is the combination of the three. After that, the resulted datasets with the original dataset passed through (Model Building) using five ML approaches and finally, the results evaluation.

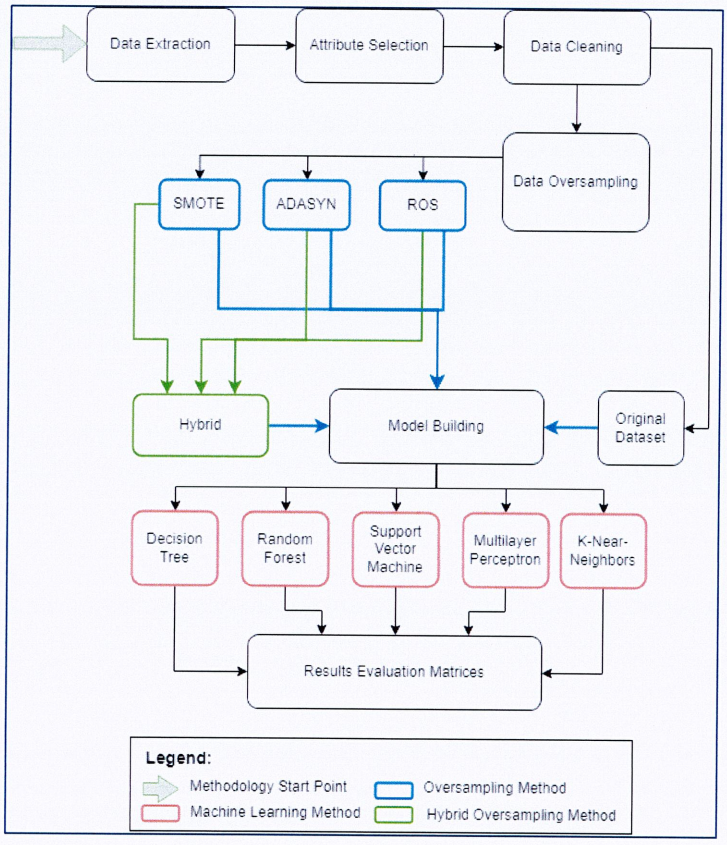


Fig. 2: Schematic Diagram of implemented methodology

3.4.1. Data and Variable Extraction

After several discussions with the civil/traffic engineers in MOW about this kind of study, they were concerned about which features to be included in this study as there are many aspects need to be considered when a traffic congestion problem occurs and most of them are treated as observations which are not recorded. For instance, the junction location, there is crucial difference when the junction in an educational area, business/working area or commercial area. Moreover, when a traffic congestion occurs on set of junctions connected as network, the engineers study it as a full project for the existing year only, hence it does not depend on cumulative data. When the data collected for a particular junction in a certain year, in practice this means it is collected in one day only. This day must be a normal day. A normal day is a weekday from Sunday to Wednesday, not an official holiday, not during summer holiday nor spring holiday, not in Ramadhan, no big events or special occasions conducted in this day. This normal day will show the real causes of the congestions. All these factors are limiting the amount of the data that can be extracted.

The dataset initially consists of 10 variables including the output as shown in Table (4). It has been collected and extracted from the reports using different ways, each attribute has been extracted differently. This section explains how each attribute was extracted.

Table (4): Variables Description

Variable	Description
Junction Name	The name of the both roads that are connected in the junction
Current Control Type	Junction's current traffic control type, whether un-signalized, signalized, roundabout, grade-separated.
Volume AM	The total number of vehicles travelled on the junction in the morning peak.
Volume LP	The total number of vehicles travelled on the junction in the Lunch peak.
Volume PM	The total number of vehicles travelled on the junction in the evening peak.
Current No# Lanes	The number of lanes in the existing junction
Delay AM	The time in seconds when the vehicles are delayed in the morning peak.
Delay LP	The time in seconds when the vehicles are delayed in the Lunch peak.
Delay PM	The time in seconds when the vehicles are delayed in the evening peak.
Decision	The contraction decision on the junction traffic control.

3.4.1.1. Junction Name Extraction

The name of the junction is represented with the roads/avenues connected together like Muharraq Avenue/16th December Avenue Junction, or with a landmark near the junction like Bahrain Gas Junction. The junctions' names are usually listed in the beginning of the project

report along with their detailed description. Although the junctions' names are not included in model, it is very important to collect them with the rest of the data in order to identify all the related details to that junction and to avoid any replication that might be included in other projects' reports as it is a connected network.

3.4.1.2. Volumes Extraction

The volumes are extracted from different places in the report which are the maps, Field counts, Synchro and Sidra. All the volumes are collected three times in the three peaks which are morning peak (6:45-7:45), lunch peak (13:30-14:30) and evening peak (16:00-17:00).

Fig. 3 and Fig. 4 are showing samples of a map with the volumes on each junction. The map is showing the volumes based on the movements and representing each peak with different colour (Morning peak = blue, Lunch peak = orange, Evening peak = green). The movement of the vehicles means in which direction the vehicle moves (right turn, straight, left turn and U-turn) for all the lanes of the junction. The values in each colour are summed together to represent the volume of the peak and link it with the name of the junction on the map. In some cases, the junction is not clear, therefore there is a need to view it from Google Earth Pro for better accuracy.

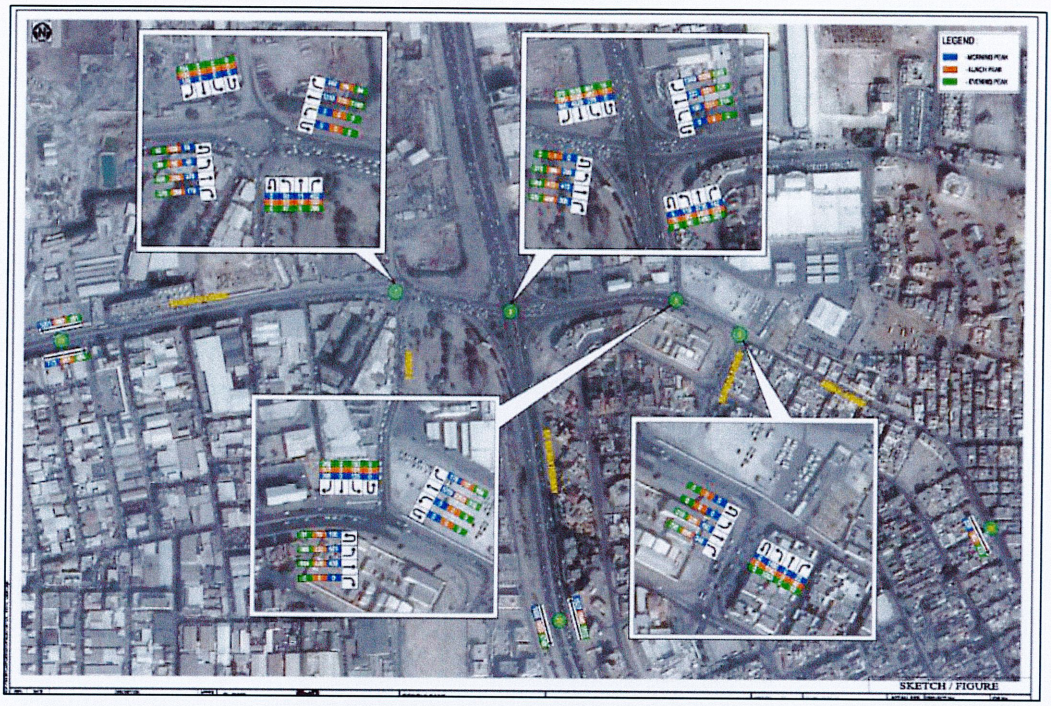


Fig. 3: Map Sample of Four Junction



Fig. 4: Map Sample of One Junction

There is another method of volumes collection, which is Field Movement Count as shown in Fig. 5. The field movement count is done by an engineer or a technician who goes to the site and count the vehicles one by one in the peak hours. The peaks are represented with different colours (Morning peak = blue, Lunch peak = orange, Evening peak = green). Similar to the maps method, the values in each colour are summed together to represent the volume in the peak hour. However, for each peak for the

field movement count method, there are two reports, one for Light Goods Vehicles (LGV) and another one for Heavy Goods Vehicles (HGV). The total number of volumes for each peak is the summation of both reports LGV and HGV.

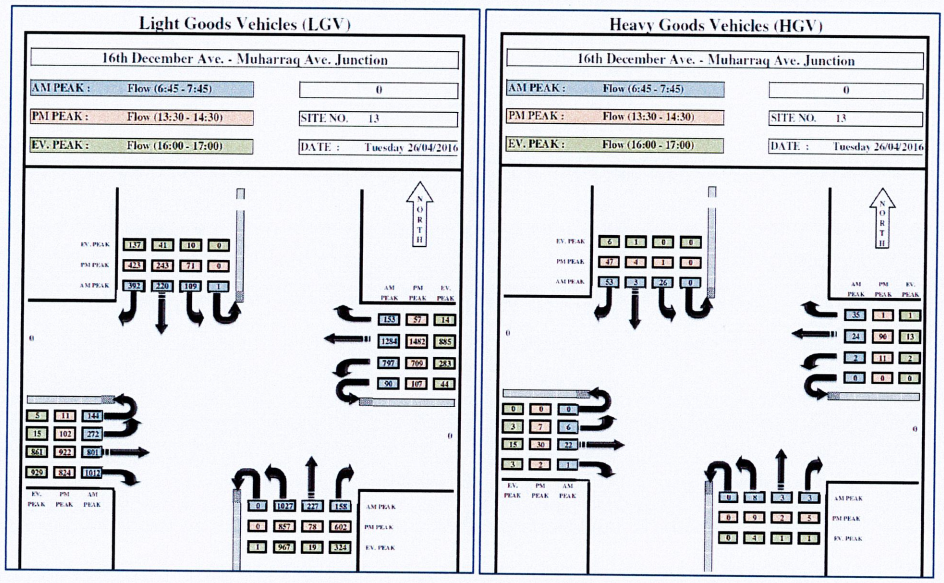
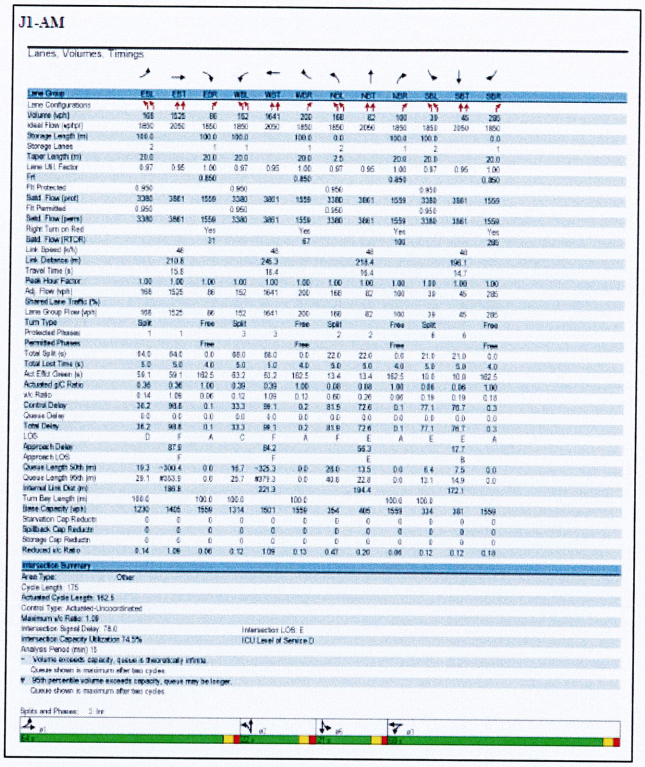


Fig. 5: Field Movement Count

Synchro is a software used for construction management (Bentley Systems, 2022) Its results are reported in details as shown in Fig. 6. Synchro is giving the details for the three peaks (AM-MD-PM) in separated report pages as presented in Fig. 7 for Muharraq Ave/SZH Junction, where MD means Mid-Day which is equivalent to LP (Lunch Peak). The volumes are illustrated in the second row under name (Traffic Volume (vph)). For each lane, to calculate the volumes for each peak, all the values in the row (Traffic Volume (vph)) must be summed manually by the author. This is repeated for each peak report for all the junctions.



J1-AM

Lanes, Volumes, Timings

Lane Group	EBL	EBT	EBR	WBL	WBT	WBR	NBL	NBT	NBR	SBL	SBT	SBR
Lane Configurations												
Volume (vph)	168	1525	88	152	1641	200	168	82	100	39	45	286
Ideal Flow (vphpl)	1850	2050	1850	1850	2050	1850	1850	2050	1850	1850	2050	1850
Storage Length (m)	100.0		100.0	100.0		100.0	0.0		100.0	100.0		0.0
Storage Lanes	2		1	1		1	2		1	2		1
Taper Length (m)	20.0		20.0	20.0		20.0	2.5		20.0	20.0		20.0
Lane Util. Factor	0.97	0.95	1.00	0.97	0.95	1.00	0.97	0.95	1.00	0.97	0.95	1.00
Frt			0.850			0.850			0.850			0.850
Flt Protected	0.950			0.950			0.950			0.950		
Satd. Flow (prot)	3380	3861	1559	3380	3861	1559	3380	3861	1559	3380	3861	1559
Flt Permitted	0.950			0.950			0.950			0.950		
Satd. Flow (perm)	3380	3861	1559	3380	3861	1559	3380	3861	1559	3380	3861	1559
Right Turn on Red			Yes			Yes			Yes			Yes
Satd. Flow (RTOR)			21			67			100			286
Link Speed (kph)		48			48			48			48	
Link Distance (m)		210.8			246.3			218.4			196.1	
Travel Time (s)		15.8			18.4			16.4			14.7	
Peak Hour Factor	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
Adj. Flow (vph)	168	1525	88	152	1641	200	168	82	100	39	45	286
Shared Lane Traffic (%)												
Lane Group Flow (vph)	168	1525	88	152	1641	200	168	82	100	39	45	286
Turn Type	Split		Free	Split		Free	Split		Free	Split		Free
Protected Phases	1	1		3	3		2	2		6	6	
Permitted Phases			Free			Free			Free			Free
Total Split (s)	64.0	64.0	0.0	98.0	98.0	0.0	22.0	22.0	0.0	21.0	21.0	0.0
Total Lost Time (s)	5.0	5.0	4.0	5.0	5.0	4.0	5.0	5.0	4.0	5.0	5.0	4.0
Act Effct Green (s)	59.1	59.1	162.5	93.2	93.2	162.5	13.4	13.4	162.5	10.0	10.0	162.5
Actuated g/C Ratio	0.36	0.36	1.00	0.56	0.56	1.00	0.08	0.08	1.00	0.06	0.06	1.00
v/c Ratio	0.14	1.09	0.06	0.12	1.09	0.13	0.60	0.26	0.06	0.19	0.19	0.18
Control Delay	36.2	90.6	0.1	33.3	99.1	0.2	81.9	72.6	0.1	77.1	76.7	0.3
Queue Delay	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total Delay	36.2	90.6	0.1	33.3	99.1	0.2	81.9	72.6	0.1	77.1	76.7	0.3
LOS	D	F	A	C	F	A	F	E	A	E	E	A
Approach Delay		87.9			84.2			56.3			17.7	
Approach LOS		F			F			E			B	
Queue Length 50th (m)	19.3	~200.4	0.0	16.7	~325.3	0.0	26.0	13.5	0.0	6.4	7.5	0.0
Queue Length 95th (m)	29.1	#263.9	0.0	25.7	#379.2	0.0	40.6	22.8	0.0	12.1	14.9	0.0
Internal Link Dist (m)		190.8			221.3			194.4			172.1	
Turn Bay Length (m)	100.0		100.0	100.0		100.0			100.0	100.0		
Base Capacity (vph)	1230	1405	1559	1314	1501	1559	354	405	1559	334	381	1559
Starvation Cap Reductn	0	0	0	0	0	0	0	0	0	0	0	0
Spillback Cap Reductn	0	0	0	0	0	0	0	0	0	0	0	0
Storage Cap Reductn	0	0	0	0	0	0	0	0	0	0	0	0
Reduced v/c Ratio	0.14	1.09	0.06	0.12	1.09	0.13	0.47	0.20	0.06	0.12	0.12	0.18

Intersection Summary

Area Type: Other

Cycle Length: 175

Actuated Cycle Length: 162.5

Control Type: Actuated-Uncoordinated

Maximum v/c Ratio: 1.09

Intersection Signal Delay: 78.0 Intersection LOS: E

Intersection Capacity Utilization 74.5% ICU Level of Service D

Analysis Period (min) 15

~ Volume exceeds capacity, queue is theoretically infinite.

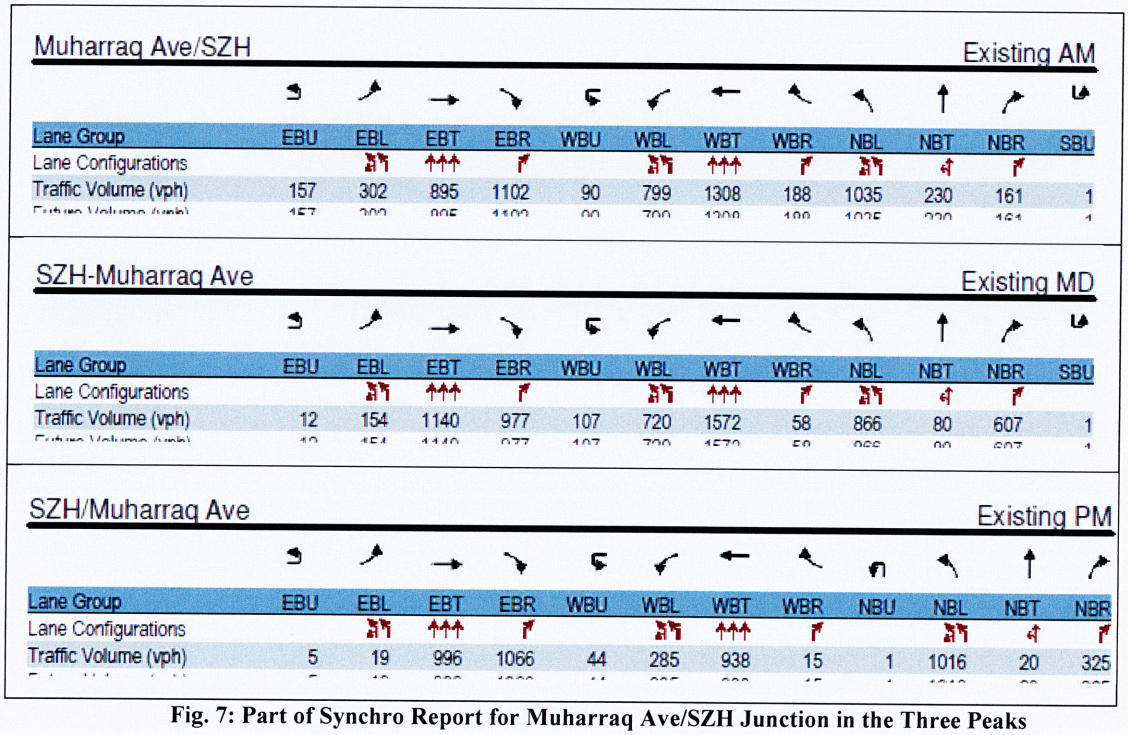
Queue shown is maximum after two cycles.

95th percentile volume exceeds capacity, queue may be longer.

Queue shown is maximum after two cycles.

Splits and Phases: 3: Int

Fig. 6: Full Report of J1-AM in Synchro



Muharraq Ave/SZH — Existing AM

Lane Group	EBU	EBL	EBT	EBR	WBU	WBL	WBT	WBR	NBL	NBT	NBR	SBU
Lane Configurations												
Traffic Volume (vph)	157	302	895	1102	90	799	1308	188	1035	230	161	1

SZH-Muharraq Ave — Existing MD

Lane Group	EBU	EBL	EBT	EBR	WBU	WBL	WBT	WBR	NBL	NBT	NBR	SBU
Lane Configurations												
Traffic Volume (vph)	12	154	1140	977	107	720	1572	58	866	80	607	1

SZH/Muharraq Ave — Existing PM

Lane Group	EBU	EBL	EBT	EBR	WBU	WBL	WBT	WBR	NBU	NBL	NBT	NBR
Lane Configurations												
Traffic Volume (vph)	5	19	996	1066	44	285	938	15	1	1016	20	325

Fig. 7: Part of Synchro Report for Muharraq Ave/SZH Junction in the Three Peaks

Synchro does not give accurate data for some junctions, e.g. the roundabout junctions. Therefore, the traffic engineers are using another software called Sidra. Sidra is micro-analytical platform that engineers use to design and evaluate a single and networks of intersections (SIDRA SOLUTIONS, 2022). Unlike Synchro, Sidra is collecting the details of the junctions based on the movement. The volumes already summed in the last row with the first column. Sidra is similar to Synchro in terms of producing separated report for each peak, which means that the total number of the vehicles must be extracted from each peak individually. Fig. 8 is showing a report sample of Sidra.

MOVEMENT SUMMARY

Site: Salmabad Bypass & Salmabad Ave

New Site
Roundabout

Movement Performance - Vehicles

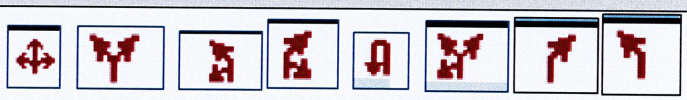
Mov ID	OD Mov	Demand Flows Total veh/h	HV %	Deg. Satn v/c	Average Delay sec
South: Salmabad Bypass					
3u	U	1	0.0	0.764	22.6
3	L2	157	10.0	0.764	22.6
8	T1	720	6.0	0.764	22.5
18	R2	262	10.0	0.764	22.3
Approach		1139	7.5	0.764	22.4
East: Salmabad Ave					
1u	U	1	3.0	0.989	67.1
1	L2	483	17.0	0.989	67.1
6	T1	113	16.0	0.241	11.3
16	R2	482	22.0	0.365	0.1
Approach		1078	19.1	0.989	31.3
North: Salmabad Bypass					
7u	U	1	0.0	1.010	67.3
7	L2	373	3.0	1.010	67.3
4	T1	778	5.0	1.010	66.7
14	R2	8	50.0	1.010	66.5
Approach		1160	4.6	1.010	66.9
West: Salmabad Ave					
5u	U	1	0.0	0.300	23.7
5	L2	24	42.0	0.300	23.7
2	T1	45	26.0	0.300	23.7
12	R2	68	44.0	0.300	23.8
Approach		138	37.5	0.300	23.7
All Vehicles		3515	11.3	1.010	39.9

Fig. 8: Part of Sidra reports

3.4.1.3. Number of Lanes Extraction

The number of lanes in the junctions is an important factor in the study of development any junction as it affects the final decision in the project. Extracting the number of lanes is done either via Synchro or Google Earth Pro. Extracting the junction's lanes using Synchro is accomplished by counting the red arrows in the first row under name (Lane configuration) in Fig. 7, each arrow shape is defining how many lanes. Table (5) is showing each arrow shape and its equivalent number of lanes to be counted. The number of lanes is counted in one of the peaks only as it is the same for all. If there is a difference in the number of lanes in the three peaks, the higher is the approved one.

Table (5) Lane Configuration Arrows Equivalence

No# Lanes	Equivalent Arrow Shape
One lane	
Two lanes	
Three lanes	
Four lanes	
Six lanes	

When Synchro didn't capture that junction then the number of lanes is counted using Google Earth Pro which has the ability to show the junction before few years as shown in Fig. 9. This ability can show the junction with the number of lanes before applying the decision. Note that the lanes counted only toward (into) the junction and not the ones exiting it.

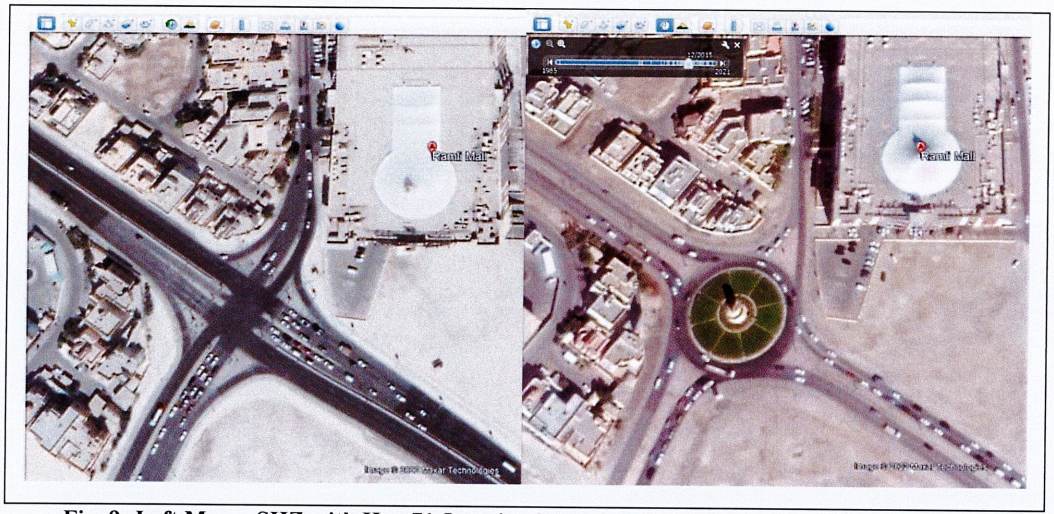


Fig. 9: Left Map – SHZ with Hwy71 Junction latest Satellite Map, Right Map - SHZ with Hwy71 Junction Satellite Map on 2015

3.4.1.4. The Delay in the Junction

The delay in each junction is the first reason to study this junction and always the aim of any junction development project is to decrease the delay value as much as possible. The only way to extract the delay is using Synchro. In the end of Synchro report of each peak there is an Intersection Summary as shown in Fig. 10. In the 7th line of the summary, there is (Intersection Signal Delay), this value is extracted from the report in each peak time for every junction. When the junction data extracted using Sidra, the delay is located in the same report presented in Fig. 8, the delay value is under "Average Delay" column located in the figure as 6th column, the bottom of the table is showing the total value, the value is meant to be extracted.

Intersection Summary		
Area Type:	Other	
Cycle Length: 140		
Actuated Cycle Length: 140		
Natural Cycle: 140		
Control Type: Actuated-Uncoordinated		
Maximum v/c Ratio: 1.09		
Intersection Signal Delay: 99.8		Intersection LOS: F
Intersection Capacity Utilization 87.0%		ICU Level of Service E
Analysis Period (min) 15		
~ Volume exceeds capacity, queue is theoretically infinite.		
Queue shown is maximum after two cycles.		
# 95th percentile volume exceeds capacity, queue may be longer.		
Queue shown is maximum after two cycles.		

Fig. 10: Part of Synchro Report - Intersection Summary

3.4.1.5. The Traffic Control Type of the Junction

The traffic control type of the junction means it is either one of these (un-signalized, signalized, roundabout, grade-separated). The type of the control mostly mentioned with the junction description in the project report. However, if not found, it can be referred to Google Earth Pro and use the feature of viewing the image of the junction before applying the solution as shown in Fig. 9.

3.4.1.6. The Decision Extraction

The main purpose of the substantial projects' reports for the junction development is the decision, which is the type of the solution that would be applied to enhance the junction in question. The extraction of these decisions in this study is made in two phases. Firstly, extract the best option of the suggested solutions for all the junctions, then classify the extracted solutions.

To start with extracting these decisions, first it must be understood that these decisions are chosen based the best suggested option. It is very often that the junction has more than option, like in Alfateh project, there

were 29 options proposed for only one junction. These options are based on many traffic engineering factors listed in the report along with their details. At the end of each junction's options list, there is an overall ranking for all options as shown in Fig. 11 knowing that it is relaying mostly on the delay in seconds caused by each option. The option with ranking of 1 is the best option, hence in Fig. 11 case, option 5 is the best one.

Table 5.2 Al Fateh Highway Workshop – 1 & 2 Overall Ranking

Option No.	Scenario	Total Point	Overall Ranking
1	Elevated Round-abouts (Awal & Shk. Daij)	102	6
2	Dumbbell Interchange (Awal)	113	4
3	Elevated U-turn Ramps (Awal)	107	5
4	1 DDI (Awal)	120	2
5	2 DDI (Awal & Shk. Daij)	122	1
6	Grade Separation along Al Fateh	117	3

Fig. 11: A sample of an overall ranking for the solution options to enhance the junction

Extracting the decision is the most critical phase, as this needs an engineering background to understand the types of the solution made, especially with the fact that the extracted solution is text-based purely written in traffic engineering terminologies and each solution has its own different details. Therefore, all extracted solutions need to be classified and standardized. After the discussions made with the engineers, all the solution has been classified into categories. There are four categories, At-Grade-Signal, Grade-Separated, Additional-Lanes and Do-Nothing. At-Grade means that the junction is on one level, in another words, no bridges, flyover or tunnels. This category has sub-category which is signal that is represented as At-Grade-Signal. On other hand, Grade-Separated means that the junction is on more than one level which means it may have bridges, flyover or tunnels. Grade-Separated has also sub-categories which

are the number of levels and in Bahrain we have until three levels only. Thus, the sub-categories are Grade-Separated-2 and Grade-Separated-3. Additional-Lanes means that the junction traffic control will stay but the number of lanes will increase. Lastly, Do-Nothing solution is the engineering term that tells no need to change this junction. Table (6) demonstrates the list of solutions after the classification.

Table (6) Decisions Description

Classified Decision	Description
At-Grade-Signal	To convert the junction from its current status to traffic signal based. This solution is on one level.
Grade-Separated-2	To convert the junction from its current status to have a flyover or a tunnel. This solution is on two levels.
Grade-Separated-3	To convert the junction from its current status to have a flyover and a tunnel alone with the at-grade level. This solution is on three levels.
Additional-Lane	To keep the junction's traffic control the same, but adding more lanes to the junction.
Do-Nothing	To keep the junction as is.

3.4.2. Data Preparation

Upon extracting the data from the report, data were prepared for the model. Data preparation in this research was broken down into three phases, namely, attribute selection, data cleaning and data over-sampling.

3.4.2.1. Attribute Selection

The extracted data is not the final data to be processed in the model as there are some attributes that are not included because they do not affect the decision directly. Thus, the Junction Name is partitioned from the extracted data. As well as the volumes in each peak, it must be divided over the current number of lanes. In other words, the volumes in each peak are replaced with volume/lane per peak.

3.4.2.2. Data Cleaning

After selecting the important attributes, data cleaning started by removing the instances with duplicated values and missing values. Then, verifying the suitability of each variable's datatype. All the variables were numbers (volumes, number of lanes: int64 and delay: float64), but the current control type was string as classes. Thus, for more efficient processing, control type has been changed to numerical value reflecting each type. Table (7) is showing the variables after attributes selection and data cleaning with their datatype. The data cleaning is done using Python in Google Colab. By the end of this stage, the dataset instances were 259 and the attributes are 9.

Table (7) Variables Datatype After Data Cleaning

Variable	Datatype
Current Control Type	int64, constant values (1-2-3-4)
Volume/Lane AM	int64
Volume/Lane LP	int64
Volume/Lane PM	int64
Current No# Lanes	int64
Delay AM	float64
Delay LP	float64

Delay PM	float64
Decision (class)	Object, constant values (5 classes: AG-S,AL,DN,GS-2,GS-3)

3.4.2.3. Data Oversampling

When the data is visualized as in Fig. 12, it was clear that the data is highly imbalanced. Therefore, more than oversampling method were applied on the data in order to have accurate predictions. First, the Synthetic Minority Oversampling Technique (SMOTE). This method works by replicating the minority classes multiple times to equate it with the dominant/majority classes, which is increased the variety as well. Second, Adaptive Synthetic (ADASYN) algorithm. This method is similar to SMOTE but it generates different number of samples depending on an estimate of the local distribution of the class to be oversampled, and last, Random Oversampling (ROM), which is selecting examples randomly from the minority class with replacement, and adding them to the training dataset. In this research, the models are applied on the dataset before oversampling, after each oversampling methods and on the merged data between the three methods.

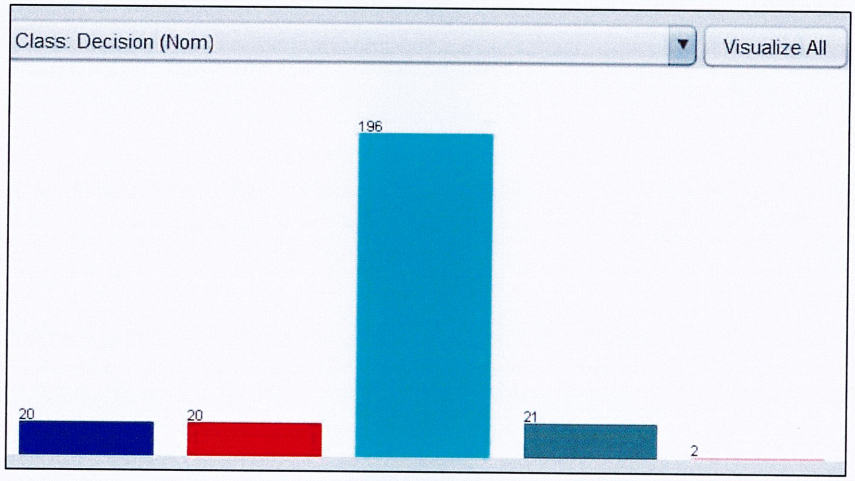


Fig. 12: Visualising the Imbalanced Classes in the Dataset

3.5. ML Approaches

This section illustrates the ML Approaches that are implemented in this research and the evaluation criteria.

3.5.1. ML Models

The model problem is classification as to predict the correct decision. In this research, three main scenarios were considered for generating and assessing the models. The first scenario is to fit the models on the dataset after the data cleaning and before applying any oversampling method. The second scenario the models are fitted in three phases, in each phase the dataset used is after applying an oversampling methods, means each phase is dealing with one of the oversampling methods only. The last scenario is to merge and append the three datasets resulted from all oversampling methods used to have the hybrid dataset then fit the model. These scenarios are meant for comparing the results of all the three oversampling methods. In addition, the hybrid dataset is providing better accuracy and higher chance to avoid the misleading outcomes from the resulted prediction made, because of the imbalanced dataset.

The ML used in this study are Support-Vector Machines (SVM), K-Nearest Neighbours (KNN), Decision Tree (DT), Random Forest (RF) and Multi-Layer-Perceptron (MLP). All the models used the default settings which are provided in Table (8), note that the batch size is 100 for all the models.

Table (8) Models' Parameters

Model	Parameters
SVM	SVM Type: C-SVC (classification) Kernel Type: Radial Base Function: exp(-gamma*|u-v|^2) Cache Size: 40.0
KNN	K=1 , Nearest Neighbor Search Algorithm: LinearNNSearch
DT	Confidence Factor: 0.25, Min Num Obj: 2, Num Folds: 3
RF	Bag Size Percent: 100, Max Depth: 0, Iterations: 100
MLP	Hidden layers: a, Learning Rate: 0.3, Momentum: 0.2

WEKA built-in ML models were used to train and test the aforementioned models. The cross-validation approach (k=10) was adopted to split the datasets into training and testing folds. Such approach is adequate for the datasets of limited size; hence it was followed. Moreover, in this approach, each record is used for training and testing. This technique helps to enhance the generalization capabilities of the models.

3.5.2. Evaluation Criteria

In order to compare the results of all the models on the three phases, there are several evaluation matrices used, such as the accuracy, Recall, Precision and F-measure. Although RMSE is the most evaluation criteria used in traffic congestion prediction (Razali et al., 2021), it is not relevant to our classification problem.

Chapter Four

Results and Discussion

4.1. Results

4.2. Discussion

Chapter Four

Results and Discussion

4.1. Results

This section presents the obtained results through applying the set of ML methods on the datasets through the three scenarios as mentioned earlier in section 3.5.1.

4.1.1. Scenario 1: Results Before Oversampling

In this scenario, the models are fitted using the dataset before applying any oversampling methods with the imbalanced classes as shown in Fig. 11.

Table (9) illustrates the results of all the models in this scenario. The accuracy is ranging between (%76.0618 – %87.2587), the highest used Random Forest model and the lowest using SVM. For Precision and F-Measure, it was not available for all the models except KNN with values 0.857, 0.858 respectively. The reason behind the unavailability of the Precision and F-Measure value is, there are some classes are not predicted correctly at all, because of the imbalance classes and the average weight between all the classed cannot define any value in WEKA this is case.

Table (9) Results of Fitted Models in Scenario 1

	DT	RF	KNN	MLP	SVM
Accuracy	83.0116	87.2587	86.1004	84.556	76.0618
Precision	Not Available	Not Available	0.857	Not Available	Not Available
Recall	0.830	0.873	0.861	0.846	0.761
F-Measure	Not Available	Not Available	0.858	Not Available	Not Available

4.1.2. Scenario 2: Results After Oversampling

In the second scenario, the models are fitted on three phases, each phase is using the dataset after applying an oversampling methods SMOTE, ADASYN and ROS.

4.1.2.1. Oversampling Using SMOTE

In the first phase of the second scenario, SMOTE is applied on the dataset. Therefore, the instances became from 259 to 980 with 196 instances for each class as shown in Fig. 13 and Fig. 14 respectively.

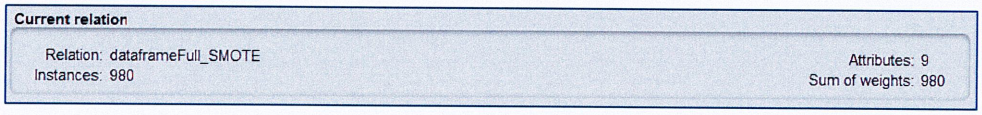


Fig. 13: The Dataset's Instances and Attributes After SMOTE

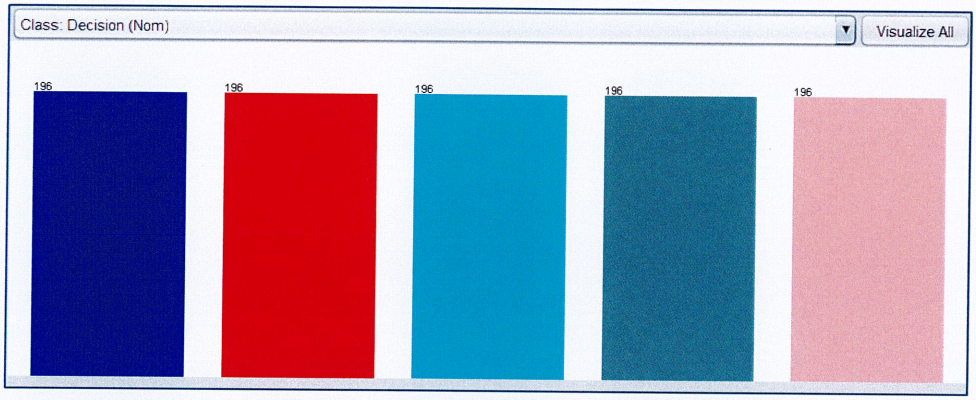


Fig. 14: Visualising Dataset's Classes After SMOTE

Table (10) emphasize the results of all the models in this phase. The accuracy is ranging approximately between (%73.9796 – %96.9388), the highest used KNN model and the lowest using SVM. The measures Precision, Recall and F-Measure were ranging almost between (0.747-0.969), and KNN was having the highest value among them.

Table (10) Results of Fitted Models in Scenario 2 - SMOTE

	DT	RF	KNN	MLP	SVM
Accuracy	95.2041	96.5306	96.9388	90.4082	73.9796
Precision	0.952	0.965	0.969	0.905	0.887
Recall	0.952	0.965	0.969	0.904	0.740
F-Measure	0.951	0.965	0.969	0.900	0.756

4.1.2.2. Oversampling Using ADASYN

In this phase of the existing scenario, ADASYN is applied on the dataset. Accordingly, the instances number 997 as shown in Fig. 15, and they are distributed among the classes differently between 196-202 instances as shown in Fig. 16.



Fig. 15: The Dataset's Instances and Attributes After ADASYN

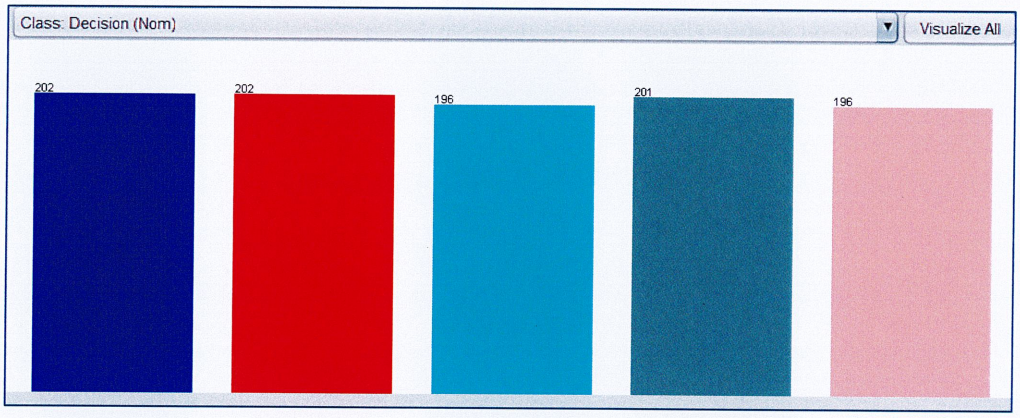


Fig. 16: Visualising Dataset's Classes After ADASYN

The accuracy is ranging approximately between (%79.5386 – %97.2919), the highest used KNN model and the lowest using SVM. The measures Precision, Recall and F-Measure were very close mostly in the range of (0.795 -0.973) and KNN was having the highest value among them. These results are illustrated in Table (11).

Table (11) Results of Fitted Models in Scenario 2 - ADASYN

	DT	RF	KNN	MLP	SVM
Accuracy	94.4835	96.7904	97.2919	90.2708	79.5386
Precision	0.944	0.968	0.973	0.903	0.900
Recall	0.945	0.968	0.973	0.903	0.795
F-Measure	0.944	0.967	0.972	0.898	0.810

4.1.2.3. Oversampling Using ROS

In the last phase of this scenario, ROS method is applied on the dataset. Appropriately, the instances number is 980 with 196 instances for each class as Fig. 17 and Fig. 18 show.

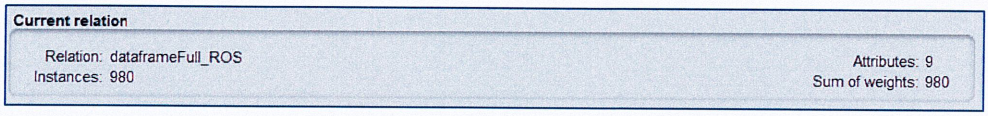


Fig. 17: The Dataset's Instances and Attributes After ROS

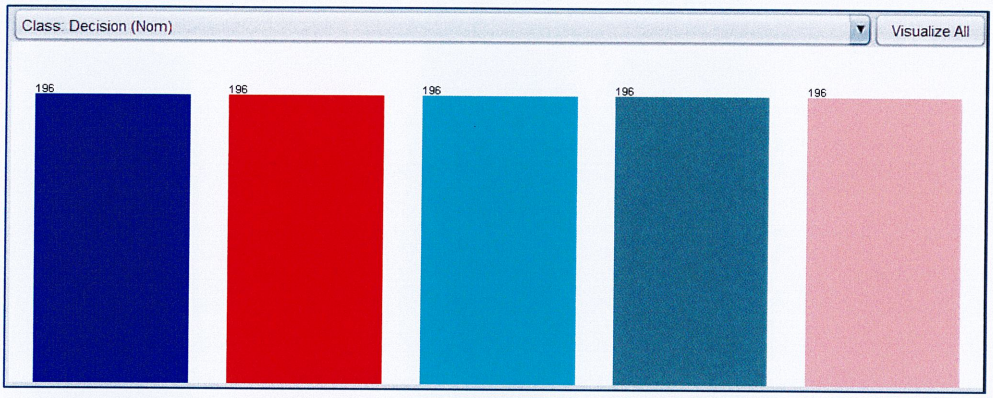


Fig. 18: Visualising Dataset's Classes After ROS

As reported in Table (12), the accuracy the highest is resulted by SVM model with %100 and the lowest using MLP %89. 6939. SVM was the highest value of all the measures Precision, Recall and F-Measure with value of 1.000. The rest of the models were ranged between (0.894-0.990).

Table (12) Results of Fitted Models in Scenario 2 - ROS

	DT	RF	KNN	MLP	SVM
Accuracy	97.0408	98.9796	98.5714	89.6939	100
Precision	0.972	0.990	0.986	0.896	1.000
Recall	0.970	0.990	0.986	0.897	1.000
F-Measure	0.970	0.990	0.986	0.894	1.000

4.1.3. Scenario 3: Using the Hybrid Dataset

In the last scenario, the dataset is a merge between the three oversampling method datasets and provided a hybrid dataset, there was no duplication while merging the three datasets. This dataset consists of 2957 unique instances as seen in Fig. 19 and the classes are distributed in a range of (588-594) as Fig. 20 shows.

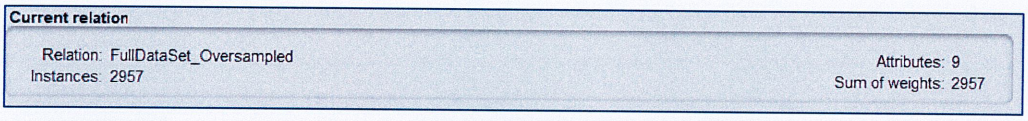


Fig. 19: The Hybrid Dataset's Instances and Attributes

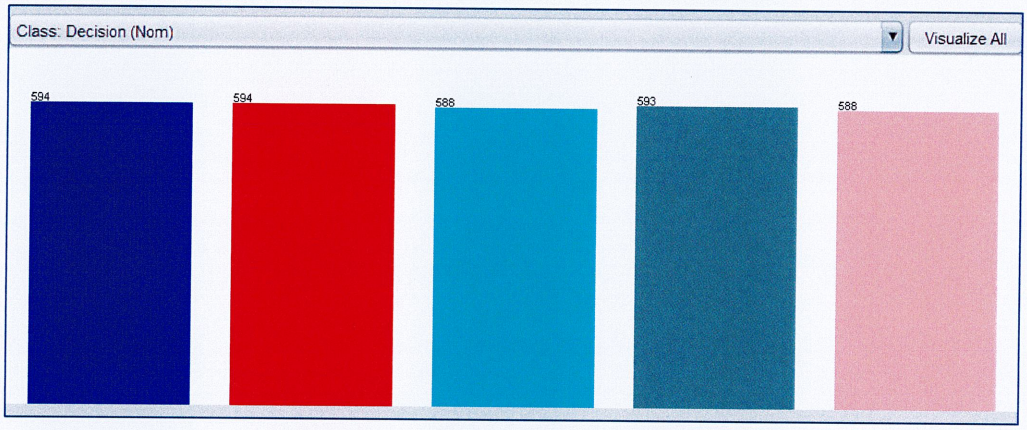


Fig. 20: The Hybrid Dataset's Classes Visualizing

As presented in Table (13), all the models provided an accuracy higher than %91.7146, reaching to the maximum %100 with KNN. For the rest of the measures all of the models resulted more than 0.90, and the optimal result was for KNN with value of 1.00.

Table (13) Results of Fitted Models in Scenario 3

	DT	RF	KNN	MLP	SVM
Accuracy	99.4927	99.9324	100	91.7146	96.28
Precision	0.995	0.999	1.000	0.916	0.969
Recall	0.995	0.999	1.000	0.917	0.963
F-Measure	0.995	0.999	1.000	0.915	0.964

4.1.4. Summary of the Findings

As a summary of the accuracy started with lower values and increased with the other datasets after oversampling. For DT, the accuracy was the lowest with %83.0 before oversampling and increases with %95.2, %94.5, %97.0 and %99.5 after oversampling methods SMOTE, ADASYN, ROS and Hybrid method respectively. Similarly, the accuracy of RT without oversampling was %87.3 and it rose to %96.5, %96.8, %99 and %99.9 and after oversampling methods. By the same token, KNN resulted an accuracy of %100 used the Hybrid method compared the accuracy without oversampling (%86.1), which is higher than the rest of oversampling methods, SMOTE (%96.9), ADASYN (%97.3) and ROS (98.6%). Last but not least, the accuracy of MLP results was also the lowest with %84.6 before oversampling and increases with %90.4, %90.3, %89.7 and %91.7 after oversampling methods following the same order. On the slip side, SVM when used SMOTE method, it gives a lower accuracy (%74) from the original dataset (%76.1), but increased in the others oversampling methods, ADASYN (%79.5), Hybrid (%96.3) and reached the maximum accuracy when used ROS method (%100) as illustrated in Fig. 21.



Fig. 21: The Accuracy of the Models Using all the Datasets

4.2. Discussion

From the results obtained, it has been noticed that all experimented models - using five forms of the dataset - are performing better when the oversampling methods. This is evidenced by the increment in the accuracy. Except the SVM when SMOTE method is used, a lower accuracy is obtained from the original dataset but increased in the other oversampling methods and reached the maximum accuracy when ROS method is used.

Moreover, the models using the ML approaches perform excellently using the Hybrid method which accumulates all the oversampling methods in this study with 100% value of accuracy using KNN model. Likewise, DT and RF, as both has tree model nature, their accuracy using the Hybrid dataset was very close to KNN, as DT has 99.4927% and RF has 99.9324%, knowing that both of these ML approaches are non-parametric models. SVM has an accuracy of %96.28 when the Hybrid dataset is used, which is lower than the maximum accuracy when ROS method is adopted.

It is still higher than the accuracy of MLP using the same dataset (91.77146%).

When the models fitted using the original dataset, the precision, which is the proportion of the correct identification of the positive cases, were unavailable in all the models except KNN. Similarly, with the F-Measure, as the precision is a part of its calculation. After oversampling, it started producing values in very acceptable range. For the Recall, it started with the original dataset with lower value, and it kept increasing when the oversampling methods was used like the others. For the three of them; Precision, Recall and F-measures, they reached the maximum value of 1, when SVM used ROS oversampling method and when KNN used the Hybrid method. This improvement in the Precision, Recall and F-Measures is an indication of the positive impact of study's approach.

Overall, KNN and RF were the best mostly in all the dataset and especially using the Hybrid of oversampling methods. While it is challenging to determine - for exact - the reasons behind the outstanding performance of the KNN and RF, in the following paragraph an attempt is made to interpret such observation. Nonetheless, caution need to be taken not to generalize the forthcoming to other datasets as each has its own characteristics.

The KNN model does not have any assumption on the data distribution, moreover it predicts multiclass problems quite easily. It also performs well when the data is non-linear. Furthermore, KNN performs well when the dimension of the dataset is relatively small.

As for the RF model, it uses the bagging technique which randomly selects data sample with replacement. This helps the models to have better prediction capabilities.

This study proves that there is a pattern of the solution process that the ML approaches are capturing and learning from it with and excellent prediction results. Nevertheless, it is acknowledged that the produced results might hint overfitting of the models. In other words, models have captured the trends too well to the extent that they become less flexible to predict new dataset with similar accuracy. However, mitigation measures have been considered to overcome the overfitting phenomenon. Firstly, the dataset size has been increased via the oversampling methods, secondly, the cross validation approached was adopted as opposed to the conventional data split techniques. This is done to ensure that the models attain the generalization capabilities.

Unlike the previous studies presented in the Literature Review, the most common ML approaches were CNN and LSTM, while it does not suit the nature of the dataset used in this study as most of the prior research are used on spatial-temporal data.

Chapter Five

Conclusion

5.1. Conclusion

5.2. Limitations

5.3. Recommendations

Chapter Five

Conclusion

5.1. Conclusion

Conducting the suitable solution for the traffic congestion is considered as a challenge for the traffic and civil engineers. This procedure is taking usually years of study, in this research, with the use of ML methods it showed the possibility to predict the solution from construction perspective with accuracy reached to %100.

As an essential component of this study, understanding the field from the experts and in this case, the traffic and civil engineer in the Ministry of Works. The knowledge has been built though multiple meetings with them along with the professional specialized reports, documentations and presentations. Moreover, several studies were focusing in the field of ITS and mostly aimed for traffic congestion prediction, they covered many and various types of ML methods with different combination, the previous studies were varying also in the sort and dimensionality of the dataset used. Unlike the datasets used in the prior studies were secondary datasets, this study used a primary dataset extracted and specialized for the junctions in Kingdom of Bahrain.

The data collection in this study, passed through several phases, started with data extraction and ended with data preparation. The data extraction process went through to each attribute level and treated separately and differently. After that, the data preparation, which firstly starts with attribute selection, then data cleaning and followed by the data oversampling. With the classification nature of the dataset, it faced the

dilemma of imbalance classes, to overcome this problem, three oversampling methods applied on the original dataset; SMOTE, ADASYN ROS and the Hybrid. This maximized the efficiency of the model process in the next stage.

The processing models used based on ML approaches were using five methods; DT, RF, KNN, MLP and SVM. These models were applied on three scenarios; before oversampling, after each oversampling method and Hybrid between the three oversampling methods.

The experimentation results of this study showed interesting outcomes. All the models performed with high accuracy value when used the hybrid method, the highest used by KNN %100, RF %99.9 and DT %99.5, while SVM and MLP were %96.3 and %91.7 respectively.

Based on the obtained results, the proposed approach of using Hybrid method which merges the three oversampling methods with ML approaches can be considered suitable for predicting the solution to mitigate the traffic congestion from a construction based perspective.

5.2. Limitations

Despite the promising outcomes of the study, there were many obstacles faced in each stage. The toughest among them is the data collection as this task need lots of time spent and effort to understand the reports as they are written in using complex engineering terminologies which are not unified. Similar to the data extraction, not all the reports following a unified structure as there are different consultation offices conducing these reports. This understanding is limited to the engineers' time table in order to clarify the solutions and valid all the extracted data to maintain its integrity, however there is still a place for human errors as the data is primary collected, and also software errors; in this scenario Synchro, Sidra and Google Earth Pro. In spite of the effort spent on the data collection, the amount of data collected is considered a bit small which leaded to use the oversampling methods, to insure of the quality of the results conducted. Also, there is no similar prior studies that used to focus on predicting the solution from engineering perspective, which limits the resources to depend on their experience.

5.3. Recommendations and Future Work

With the results obtained in the research, it is recommended to use this kind of prediction in the construction side to mitigate the traffic congestion, which might enhance the procedure of studying any traffic problem in the future and empower the decisions taken using ML approaches. Moreover, it is highly recommended to create a centralized database using any tool for database representation and manipulation for all the data collected and not limit it to the text based reports. This will allow the researchers, data analysts and data scientists to come up with new methods and techniques to support this field.

For future work, it is always better to have a wider range of data that includes more instances and more attributes. This will contribute to the robustness of the ML models. The traffic engineers, also highlighted many factors that can be considered as attributes which is how in reality happens; like and not limited to the pedestrians count and the vehicles queue length.

In the research, five ML approaches used, and three oversampling methods tested. Thus, more and different ML approaches and oversampling methods are interestingly worth to test and evaluate.

This study could be the spark to initiating a database hub for all the projects owned by Ministry of Works, which is definitely is meant to be futuristic planed.

References

Akhtar, M., & Moridpour, S. (2021). A Review of Traffic Congestion Prediction Using Artificial Intelligence. *Journal of Advanced Transportation, 2021*, e8878011. https://doi.org/10.1155/2021/8878011

Bai, M., Lin, Y., Ma, M., Wang, P., & Duan, L. (2020). PrePCT: Traffic congestion prediction in smart cities with relative position congestion tensor. *Neurocomputing*. https://doi.org/10.1016/j.neucom.2020.08.075

Bentley Systems. (2022). *SYNCHRO Resources*. Www.bentley.com. https://www.bentley.com/en/products/brands/synchro/resources

Borges, F. de S. P., Fonseca, A. P., & Garcia, R. C. (2021). Deep Reinforcement Learning Model to Mitigate Congestion in Real-Time Traffic Light Networks. *Infrastructures*, *6*(10), 138. https://doi.org/10.3390/infrastructures6100138

Chen, C., Li, K., Teo, S. G., Zou, X., Wang, K., Wang, J., & Zeng, Z. (2019). Gated Residual Recurrent Graph Neural Networks for Traffic Prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*, 485–492. https://doi.org/10.1609/aaai.v33i01.3301485

Crivellari, A., & Beinat, E. (2020). LSTM-Based Deep Learning Model for Predicting Individual Mobility Traces of Short-Term Foreign Tourists. *Sustainability*, *12*(1), 349. https://doi.org/10.3390/su12010349

Defining congestion. (2019). https://www.transport.govt.nz/assets/Uploads/Paper/DefiningCongestion.pdf

Designing Buildings. (2022, April 7). *Traffic v transport*. Www.designingbuildings.co.uk; Designing Buildings. https://www.designingbuildings.co.uk/wiki/Traffic_v_transport

Dimitrakopoulos, G., & Demestichas, P. (2010). Intelligent Transportation Systems. *IEEE Vehicular Technology Magazine*, *5*(1), 77–84. https://doi.org/10.1109/mvt.2009.935537

Donges, N. (2021, July 22). *A Complete Guide to the Random Forest Algorithm*. Built In. https://builtin.com/data-science/random-forest-algorithm

Elfar, A., Talebpour, A., & Mahmassani, H. S. (2018). Machine Learning Approach to Short-Term Traffic Congestion Prediction in a Connected Environment. *Transportation Research Record: Journal of the Transportation Research Board*, *2672*(45), 185–195. https://doi.org/10.1177/0361198118795010

Fan, X., Liu, J., Wang, Z., Jiang, Y., & Liu, X. (2017). Crowdsourced Road Navigation: Concept, Design, and Implementation. *IEEE Communications Magazine*, *55*(6), 126–128. https://doi.org/10.1109/mcom.2017.1600738

Google Maps Platform. (2022). *Google Maps Platform - Location and Mapping Solutions*. Google Maps Platform. https://mapsplatform.google.com/

Holst, A. (2021, June 7). *Data created worldwide 2010-2025 | Statista*. Statista; Statista. https://www.statista.com/statistics/871513/worldwide-data-created/

Hou, Y., Deng, Z., & Cui, H. (2021). Short-Term Traffic Flow Prediction with Weather Conditions: Based on Deep Learning Algorithms and Data Fusion. *Complexity*, *2021*, 1–14. https://doi.org/10.1155/2021/6662959

IBM Cloud Education. (2020, August 17). *What are Neural Networks?* Www.ibm.com; IBM. https://www.ibm.com/cloud/learn/neural-networks

INRIX. (2021, December 7). *INRIX: Americans Lost 3.4 Billion Hours Due to Congestion in 2021, 42% Below Pre-COVID*. Inrix. https://inrix.com/press-releases/2021-traffic-scorecard/

Institute of Transportation Engineers. (n.d.). *History*. Institute of Transportation Engineers. Retrieved February 10, 2022, from https://www.ite.org/about-ite/history/

JavaTpoint. (2011). *K-Nearest Neighbor(KNN) Algorithm for Machine Learning - Javatpoint*. Www.javatpoint.com. https://www.javatpoint.com/k-nearest-neighbor-algorithm-for-machine-learning

Kamble, S. J., & Kounte, M. R. (2020). Machine Learning Approach on Traffic Congestion Monitoring System in Internet of Vehicles. *Procedia Computer Science*, *171*, 2235–2241. https://doi.org/10.1016/j.procs.2020.04.241

Khan, S. M., Dey, K. C., & Chowdhury, M. (2017). Real-Time Traffic State Estimation With Connected Vehicles. *IEEE Transactions on Intelligent Transportation Systems*, *18*(7), 1687–1699. https://doi.org/10.1109/tits.2017.2658664

Kishk, A. M., Badawy, M., Ali, H. A., & Saleh, A. I. (2021). A new traffic congestion prediction strategy (TCPS) based on edge computing. *Cluster Computing*, *25*(1), 49–75. https://doi.org/10.1007/s10586-021-03377-2

Kumarage, S. P., De Silva, D., & Bandara, J. M. S. J. (2017). Travel time estimation based on dynamic traffic data and machine learning principles. *IESE Annual Sessions, 2017*, 1135–1142.

Lay, M. (2012). An historical review of the assessment and management of congestion. *Road and Transport Research*, *21*, 32–45.

Lv, Y., Duan, Y., Kang, W., Li, Z., & Wang, F.-Y. (2014). Traffic Flow Prediction With Big Data: A Deep Learning Approach. *IEEE Transactions on Intelligent Transportation Systems*, *16*(2), 1–9. https://doi.org/10.1109/tits.2014.2345663

Majumdar, S., Subhani, M. M., Roullier, B., Anjum, A., & Zhu, R. (2020). Congestion Prediction for Smart Sustainable Cities using IoT and Machine Learning Approaches. *Sustainable Cities and Society*, 102500. https://doi.org/10.1016/j.scs.2020.102500

Pandey, A., & Jain, A. (2017). Comparative Analysis of KNN Algorithm using Various Normalization Techniques. *International Journal of Computer Network and Information Security*, *9*(11), 36–42. https://doi.org/10.5815/ijcnis.2017.11.04

Polson, N. G., & Sokolov, V. O. (2017). Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, *79*, 1–17. https://doi.org/10.1016/j.trc.2017.02.024

Razali, N. A. M., Shamsaimon, N., Ishak, K. K., Ramli, S., Amran, M. F. M., & Sukardi, S. (2021). Gap, techniques and evaluation: traffic flow prediction using machine learning and deep learning. *Journal of Big Data*, *8*(1). https://doi.org/10.1186/s40537-021-00542-7

SIDRA SOLUTIONS. (2022). *Powerful Traffic Engineering Software | SIDRA SOLUTIONS*. Www.sidrasolutions.com. https://www.sidrasolutions.com/

Sun, S., Chen, J., & Sun, J. (2019). Traffic congestion prediction based on GPS trajectory data. *International Journal of Distributed Sensor Networks*, *15*(5), 155014771984744. https://doi.org/10.1177/1550147719847440

Support Vector Machine Algorithm. (2021, January 20). GeeksforGeeks. https://www.geeksforgeeks.org/support-vector-machine-algorithm/

Tampubolon, H., & Hsiung, P.-A. (2018). Supervised Deep Learning Based for Traffic Flow Prediction. *ICSGTEIS 2018*.

Wang, Z. (2018). Analysis and Prediction of Urban Traffic Congestion Based on Big Data. *International Journal on Data Science and Technology*, *4*(3), 100. https://doi.org/10.11648/j.ijdst.20180403.14

Wang, Z., Chu, R., Zhang, M., Wang, X., & Luan, S. (2020). An Improved Selective Ensemble Learning Method for Highway Traffic Flow State Identification. *IEEE Access*, *8*, 212623–212634. https://doi.org/10.1109/access.2020.3038801

Weiland, R. J., & Purser, L. B. (2000). INTELLIGENT TRANSPORTATION SYSTEMS. *Transportation in the New Millennium*. https://trid.trb.org/view/639268

Wenzel, E. (2015, June 30). *Road Congestion Relief: How Engineers are Fighting Traffic | Greenbiz*. Www.greenbiz.com; GreenBiz Group. https://www.greenbiz.com/article/road-congestion-relief-how-engineers-are-fighting-traffic#:~:text=Among%20the%20top%20reasons%20for

Wu, Y., Tan, H., Qin, L., Ran, B., & Jiang, Z. (2018). A hybrid deep learning based traffic flow prediction method and its understanding. *Transportation Research Part C: Emerging Technologies*, *90*, 166–180. https://doi.org/10.1016/j.trc.2018.03.001

YANG, D., LI, S., PENG, Z., WANG, P., WANG, J., & YANG, H. (2019). MF-CNN: Traffic Flow Prediction Using Convolutional Neural Network and Multi-Features Fusion. *IEICE Transactions on Information and Systems*, *E102.D*(8), 1526–1536. https://doi.org/10.1587/transinf.2018edp7330

Yisheng, L., Yuan-yuan, C., Zhenjiang, L., & Fei-Yue, W. (2016). Long Short-Term Memory Model for Traffic Congestion Prediction with Online Open Data. *2016 IEEE 19th International Conference on Intelligent Transportation Systems (ITSC)*.

Yokota, T., Weiland, R. J., & Yamagata, H. (2005). Approaches for Introducing Intelligent Transportation Systems into Developing Countries. *Transportation Research Record: Journal of the*

Transportation Research Board, *1910*(1), 72–81. https://doi.org/10.1177/0361198105191000109

Yuan, T., Rocha, W., Rothenberg, C. E., Obraczka, K., Barakat, C., & Turletti, T. (2021). Machine learning for next-generation intelligent transportation systems: A survey. *Transactions on Emerging Telecommunications Technologies*. https://doi.org/10.1002/ett.4427

Zafar, N., & Ul Haq, I. (2020). Traffic congestion prediction based on Estimated Time of Arrival. *PLOS ONE*, *15*(12), e0238200. https://doi.org/10.1371/journal.pone.0238200

Zhang, K., & Batterman, S. (2013). Air pollution and health risks due to vehicle traffic. *Science of the Total Environment*, *450-451*, 307–316. https://doi.org/10.1016/j.scitotenv.2013.01.074

Appendices

الملخص

على الرغم من فوائد أنظمة النقل الذكية في تعزيز أداء آلية النقل وتسهيل الإيقاع السريع لنمط حياة السكان، إلا أن مشكلة الازدحام المروري لازالت موجودة. ومع التزايد السكاني والعمراني، يزداد الازدحام المروري أيضًا. لذلك، أجريت هذه الدراسة للتخفيف من الازدحام المروري باستخدام مناهج التعلم الآلة والذي يهدف إلى تخفيف الازدحام المروري في مملكة البحرين من خلال التنبؤ بالحل المناسب من منظور هندسة المرور والهندسة المدنية. والجدير بالذكر بأن هذا البحث هو الأول من نوعه في مملكة البحرين. استخدم هذا البحث بيانات أولية جمعها المؤلف من التقارير والوثائق التي قدمتها وزارة الأشغال المسؤولة عن إنشاء الطرق في البحرين. كانت مجموعة البيانات المستخرجة غير متوازنة التصنيف، لذلك تم تطبيق ثلاث طرق لأخذ العينات على مجموعة البيانات. تنتج ثلاث مجموعات من البيانات من كل طريقة من طرق أخذ العينات الزائدة ؛ تقنية الإفراط في أخذ العينات للأقليات الاصطناعية (SMOTE) ، والتركيب التكيفي (ADASYN) ، والاختزال العشوائي (ROS) والأسلوب الرابع هو مزيج من مجموعات البيانات الثلاث. النماذج مبنية باستخدام خمسة مناهج تعلم الآلة. تم تطبيق جميع النماذج خمس مرات؛ الأولى قبل تطبيق الاختزال المفرط والآخريات لكل طريقة من طرق زيادة العينات بما في ذلك الطريقة الهجينة. أنتجت مجموعة البيانات الهجينة دقة عالية في جميع نماذج تعلم الآلة .نتج عن k-Nearest Neighboursو Random Forest أفضل دقة بنسبة٪ 100 و٪ 99.9 على التوالي .أثبتت هذه الدراسة أن مناهج تعلم الآلة تعمل على تمكين اتخاذ القرار في اختيار الحل المناسب لتحسين الطرق في التقاطعات للتخفيف من الازدحام المروري.

جامعة البحرين

كلية العلوم

قسم برامج الدراسات العليا

استخدام أساليب تعلُم الآلة للتقليل من الإزدحامات المرورية

إطروحة مقدمة كجزء من متطلبات الحصول على درجة الماجستير
في علوم وتحليل البيانات الضخمة

إعداد

عائشة خالد مبارك

الرقم الجامعي: ٢٠٠٩٢٩٠٥

إشراف

د. وائل المدني
أستاذ مشارك
جامعة البحرين

أ.د. نبيل حويحي
أستاذ
جامعة البحرين

مـملكة البـــحرين
سبتمبر ٢٠٢٢م