Synergistic Urban-Transport prediction model Using Support Vector Machine; in Support of Sustainable Development Goals Achievement: A Case of State of Qatar

Aya Hasan AlKhereibi
Qatar Transportation and Traffic Safety Center,
College of Engineering,
Qatar University
Doha, Qatar
aalkhereibi@qu.edu.qa

Murat Kucukvar
Department of Mechanical and Industrial Engineering
College of Engineering,
Qatar University
Doha, Qatar
mkucukvar@qu.edu.qa

Nuri Onat
Qatar Transportation and Traffic Safety Center,
College of Engineering,
Qatar University
Doha, Qatar
onat@qu.edu.qa

Jaser Mohammad Saleh Mustafa
Department of Mechanical and Industrial Engineering
College of Engineering,
Qatar University
Doha, Qatar
js2000624@qu.edu.qa

Abstract

Predicting metro ridership is an essential requirement for efficient metro operation and management. The dependence of Metro ridership on the land-use densities is extremely logical and entails a need for a model that predicts the Metro ridership using land use densities. The objective of this research study is to develop a model to predict the metro station ridership utilizing the land use densities in the vicinity of metro stations. A Support Vector Machine (SVM) model was developed as a classifier in order to predict ridership patterns. The ridership data was obtained from Qatar Rail, and the land use data were obtained from the Ministry of Municipality and Environment (MME), State of Qatar. The land use densities in the catchment area of 800m around the metro stations have been considered in this study. The nonlinear relationship between the metro ridership and land use densities has been captured through the SVM model. The resultant performance of models prediction showed a rational accuracy in which the variance of predicted ridership and actual data didn't exceed 0.14, the proposed prediction model can be utilized by both Urban and transport planners in their processes to plan the land use around metro stations and predict the transit demand from those plans and achieve the optimal use of the transit system i.e. Transit-Oriented Developments (TOD).

Keywords
Metro Ridership, Machine Learning, Transportation Planning, Support Vector Machine, Land use.

1. Introduction
1.1 Overview
In the presence of rapid urban growth and economic evolution, the successful operations and efficiency of transportation infrastructure are, to a major extent, a key indicator of a country's strength. In such a context, the need for transportation engineers and planners becomes eminent in order to predict the transportation demand using the attributes related to urban growth. In addition, achieving the integration and sustainability of public transportation and land-use planning around transport nodes plays a significant role (Shao et al., 2020). The transport planners must always choose the best scenario of development for various land uses considering the various impacts of the development. One of the major concerns is providing enough accessible public transport facilities to encourage residents to use sustainable modes of transport (He et al., 2020). A solution for such a situation can be provided by implementing an intelligent computing tool that assures the best utilization for land use and operation of public transport. Some developing countries are witnessing the introduction of the metro system as a new mode of transport. ridership prediction models are essential to plan metro rail operations and assess the sufficiency of metro stations. In the case of the State of Qatar, Qatar Rail has officially started metro operations in November 2019. To plan an efficient metro schedule, it is necessary to develop models to predict ridership, which could help in decision-making and support for metro ridership improvement (Wang et al., 2018; Lam & Toan, 2008; David, 2003). Since no previous models are available such as this research study attempts to fill this gap by developing a predictive model that estimates station-based monthly metro ridership with respect to the land use densities around rail stations for the State of Qatar. To the best of our knowledge, no previous research has proposed a Support Vector Machine (SVM) ridership prediction model using land-use densities around metro stations. By intelligently incorporating the ridership data into the SVM projected space, the quantification of various scenarios have been achieved which, in turn, has enabled a better insight into the real-world challenges associated with the ridership prediction. The manuscript is organized as follows; the first section provides a review of the literature with a brief theoretical background. The second section presents the detailed methodology followed in this study along with details of the SVM model and collected data. along with the model development SVM . Subsequently, the model results are explained and discussed, and finally, Section 6 presents the conclusions, applications, and limitations of this study.

1.2 Research objectives and motivation
The nature of metro transportation differs from any other mode of transportation in its unique characteristics such as its non-linear nature (Sun et al., 2015). Although rail transport and land use planning have been widely studied, other models that take land use into account in forecasting rail use are needed. To the extent of the authors' knowledge, no previous research has investigated the use of aggregate Support Vector Machine model to predict metro ridership taking into account land use in the catchment area around stations. several key factors influencing the forecast for metro ridership. To this end, this study examines the use of the Support Vector Machine (SVM) model to predict metro ridership for the state of Qatar. In addition, this study proposed ML predictive model for a unique case of the State of Qatar, where rail transit is recent transit, and its use requires sophisticated and reliable prediction and planning tools.

2. Literature Review
The main concern of urban planners and transit operators is to assure the highest efficiency of land use and rail operation (Berawi et al., 2020). The basic strategies for such a purpose are illustrated by Cervero & Kockelman, (1997); basically, the transit ridership should be raised by increasing the land use densities around the stations. Moreover, increase the passengers' convenience by diversifying the land use types in the station catchment area, and increase the rail transit use through following strategies that enhance accessibility, for instance by creating a pedestrian-oriented environment. The pedestrian-oriented environment could result in the high mixed use of the land around the station, which allows citizens to satisfy all their desired activities in walkable and transit-friendly access. Planning the environment around the station with a high-dense-diverse development should be a crucial apprehension for urban planners and transit operators (Wey et al., 2016). Metro ridership forecasting has been a vital process over the last two decades. Generally, long-term and short-term ridership prediction models are developed to aid transport planners. The traditional methods of ridership prediction for long-term predictions are mainly based on the patterns of trip generation/ attraction and users' demand (Bar-Gera & Boyce, 2003; Jovicic & Hansen, 2003). This type of model is significant during the stages of planning and construction of the rail, yet it couldn't satisfy the needs of the

operational aspect (Chen & Wei, 2011). The urban fabric and land use densities characteristics around the metro station are vital for the planning and operation of the transport systems as well as ridership demand estimation; unlike the vehicle traffic flow which focuses mostly on operational aspects (Yin & Shang, 2016; Sundaram et al., 2011)
The short-term ridership prediction models can be divided into two types of prediction: theoretical method and data-driven models. The theoretical method is mainly based on the ridership flow within the station and investigates its dynamics. On the other hand, the data-driven models depend on the collection and observation of datasets and construct their prediction features based on the data. (Yin & Shang, 2016; Sundaram et al., 2011). Data-driven models can further be divided into three types; linear, non-linear, and hybrid prediction methods. The linear methods depend on stationary time series datasets and linearized formulations. The non-linear models have many types, such as nonparametric regression, Gaussian maximum likelihood model, Neural Network (NN), and SVM (Yanwen Li et al., 2017; Smith et al., 2002; Guan et al., 2020). For better accuracy and reliability, the hybrid models tend to combine more than two methods for more accurate forecasting results.

Several existing studies in the broader literature have concluded that the NN method is the most popular non-linear model (Lam & Toan, 2008; Wang et al., 2018). The NN method is most suitable for short-term prediction since it trains the neurons based on a historical dataset, which generates a complex nonlinear link between the inputs and outputs. The main shortcoming of the NN method is that it does not minimize the risk expectations which leads to overfitting, (Vlahogianni et al., 2005; Vapnik, 2000). Compared to the NN method, the SVM is also suitable for short-term prediction models as it satisfies the general prediction requirements and can be generalized for data-specific dynamics. This ability is mainly based on the principles of risk minimization and is obtained using a kernel function. In addition, the SVM has the ability to solve and predict outcomes of a small nonlinear sample with the least overfitting probability (Zhao et al., 2020). Support Vector Machine is one of the classic machine learning techniques that can still help solve big data classification problems. However, the Support Vector Machine is mathematically complex and computationally intensive (Suthaharan, 2001). This section deals with attempts to clarify the theory of SVM in order to successfully implement it in the problem under discussion. The main idea of the SVM implementation is that it maps the inputs into a space Z of high dimensional characteristics by utilizing some nonlinear mapping tools. The concern for such a method is a balanced compromise between how the model could be generalized (conceptual), and how it could be implemented (technical). In order to realize the conceptual part, Suthaharan, (2001) constructed soft separating hyperplane and margin separating hyperplane. On the other hand, to solve the technical aspect, it is not necessary to consider the feature space in its explicit form but to calculate the values within the support vectors and its feature space.
Moreover, problems caused by nonlinear and big samples can be evaded by using SVM. It shows a fundamental role when performing recognition, classification, and regression prediction models (Gao et al., 2022). For prediction and classification problems, SVM could be classified as support vector classification (SVC) and support vector regression (SVR). To predict the metro ridership based on land use densities, SVM was used in this study, and the theory is shown below.

$\{(x\_i, y\_i)\ = 1,2, \ldots\ldots, n, x\_i \in R\hat{}n, y\_i \in R$ is a given dataset, where X_i is the input training sample, model input trained dataset, and Y_i is the output trained detatset. Given that the general linear regression equation of SVR is constructed as shown in Eq. (1)

$$f(x) = w^T \cdot x + b \tag{1}$$

where *w* represents the weight vector, where w represents the weight vector, which is the coefficient of *x*, *b* represents a constant. x_i can be substituted into *x* in Eq. (1), for the calculation and with yi, w and b are selected according to the SRM principle. f(x) refers to the output value, which could result in a compared error to improve generalizability, Eq. (2) represents the process

$$\min \tfrac{1}{2}\|w\|^2 + C \sum_{i=1}^{m}(\xi_i + \xi_i^*) \tag{2}$$

$$s.t. \begin{cases} y_i - f(x_i \le \varepsilon + \xi i) \\ f(x_i) - y_i(\le \varepsilon + \xi_i^*) \\ \xi_i \ge 0, \xi_i^* \ge 0 \end{cases}$$

Where ε represents the loss function, ξi , ξ_i^* shows the relaxation variables that have different values C represents the penalty coefficient, and m is the number of trained sample. A Lagrangian function could be established as shown in Eq. (3):

$$L(w, b, \propto, \propto^*, \xi, \xi^*, \mu, \mu^*) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}(\xi_i + \xi_i^*) + \sum_{i=1}^{m} \propto_i (f(x_i) - y_i - \varepsilon - \xi) - \sum_{i=1}^{m} \mu_i \xi_i - \sum_{i=1}^{m} \mu_i^* \xi_i^* + \sum_{i=1}^{m} \propto_i^* (y_i - f(x_i) - \varepsilon - \xi_i^*) \quad (3)$$

where $\propto, \propto^*$, μ and μ^* are Lagrangian multipliers, which are bigger than zero with dissimilar values. The partial derivatives *w, b*, ξ and $\xi^*$are obtained given Eq (4):

$$w = \sum_{i=1}^{m}(\propto_i^* - \propto_i)\, x_i \quad (4)$$

$$f(x_i) = \sum_{i=1}^{m}(\propto_i^* - \propto_i) \quad (5)$$

The metro ridership is subject to many land use density input features. To avoid the results of the multi-dimensional characteristics, the kernel function Eq. (6-10) is used for the purpose of presenting the data in a high dimensional space

$$w(\propto_i, \propto_i^*) = \sum_{i=1}^{m}(\propto_i - \propto_i^*)\, y_i - \varepsilon \sum_{i=1}^{m}(\propto_i - \propto_i^*) - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\propto_i - \propto_i^*)(\propto_j - \propto_j^*)\, K(x_i, x_j) \quad (6)$$

$$s.t. \sum_{j=1}^{m}(\propto_i - \propto_i^*) = 0 \quad (7)$$

$$w = \sum_{i=1}^{m}(\propto_i^* - \propto_i)\varphi(x_i) \quad (8)$$

$$f(x) = \sum_{i=1}^{m}(\propto_i^* - \propto_i)\left(\varphi(x_i)\varphi(x)\right) + b \quad (9)$$

$$f(x) = \sum_{i=1}^{m}(\propto_i^* - \propto_i)\, K(x_i, x) + b \quad (10)$$

where $K(x_i, x)$ represents the kernel function.

$$K(x_i, x) = \exp\left(-\frac{\|x_i - x\|^2}{2\sigma^2}\right) \quad (11)$$

where σ in Eq. (11) is the bandwidth of the Gaussian kernel and is the key parameter in the kernel function.

The nature of metro ridership differs from any other transportation mode due to its unique characteristics, (Sun et al., 2015). After reviewing and comparing the modeling methods that deal with nonlinear problems, we have selected the SVM method as the most suitable method that fits the short-term, nonlinear and periodic type metro ridership prediction. On that basis, the research study proposes a data-driven SVM model that uses actual metro ridership data as model inputs to train the model and subsequently test the model using the testing dataset which is previously unseen by the training model, in order to provide the ridership class as output considering land use density around metro stations in the State of Qatar. Although rail ridership and land use planning have been extensively studied, further models that consider the land use in rail ridership prediction are required. The significance and contributions of this study are as follows: I) A new model has been proposed that predicts short-term metro ridership concerning the land use densities around the stations using the nonlinearity characteristics of the input data of ridership attributed by date on a monthly basis. II) The developed model used four weeks of data (5627 observations on an hourly scale) to construct the prediction function using SVM. III) In general, ridership prediction represents the forecasting for a pre-defined time interval. Moreover, we can assure that ridership input data has a periodic nonlinear nature resulting from the differences in ridership from weekdays, weekends, and vacations. Thus, and according to what (Wang et al., 2018) discussed, some proposed models in the literature worked well for the weekdays, yet fail to give satisfactory predictions for the weekends. In order to prove the applicability of the model for different ridership observations, both weekdays and weekend data were used in the training and testing of the models.

3. Methods

As presented in the literature several scholars have investigated rail transit ridership using ML models, yet to the extent to the authors' knowledge that no previous research has investigated the use of SVM to predict rail ridership considering land use in the catchment area around the stations. Moreover, this applies the proposed model to a uniques case of the State of Qatar, where the rail transit is newly transit, in which its use needs sophisticated and applicable prediction and planning tools This section discusses the machine learning model development, the data used to fit the models, data training, model validation, and model testing. As illustrated in Figure .1, the structure of the presented study went through several stages; starting from the first stage of problem identification and literature review, followed by data collection and analysis. The third stage included data filtering and preprocessing, then the SVM model building using the RStudio software package. This included training on 80% of the data, followed by testing on the remaining 20% of the data. The constructed SVM model was tested using three performance indicators namely, MAE, RMSE, and R^2. Finally, the model was run to generate the predicted values of ridership to present the tuned model and assess its performance. The following section provides a theoretical background for the SVM model and later presents the data collection process.

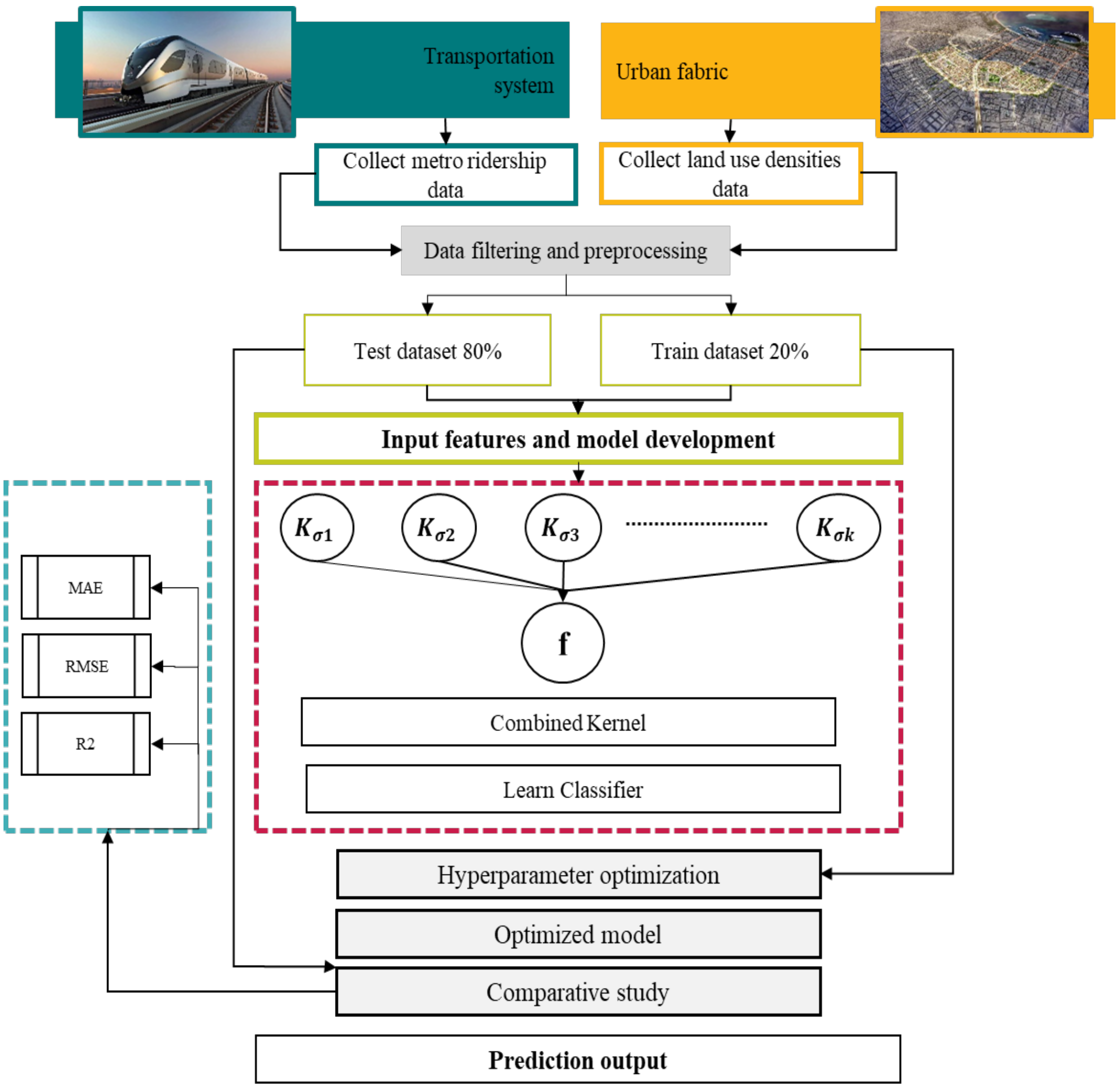


Figure .1 Research design flowchart

4. Data Collection

State of Qatar incorporates the Qatar Integrated Railways Program (QIRP) in establishing the Doha Metro Project, which features a 300 km network intended to link all significant districts to the city's international airport, Olympic stadia, ports, and the urban villages. Qatar Rail is responsible for developing and operating the Doha metro project. At its ultimate stage, the project will consist of four lines serving 98 stations. Figure .2 shows the map for Doha metro project. The whole project is designed in two phases. The first phase comprises the red, golden, and green lines, with 37 stations and 75 km of revenue lines. This phase has started its operation fully in 2019. The next phase is expected to add 60 more stations, and 130 km revenue lines to fully revolutionize people's movement around the city of Doha, and its suburbs. The establishment also creates access to job opportunities through a reliable and sustainable transport system (Qatar Rail, 2020).

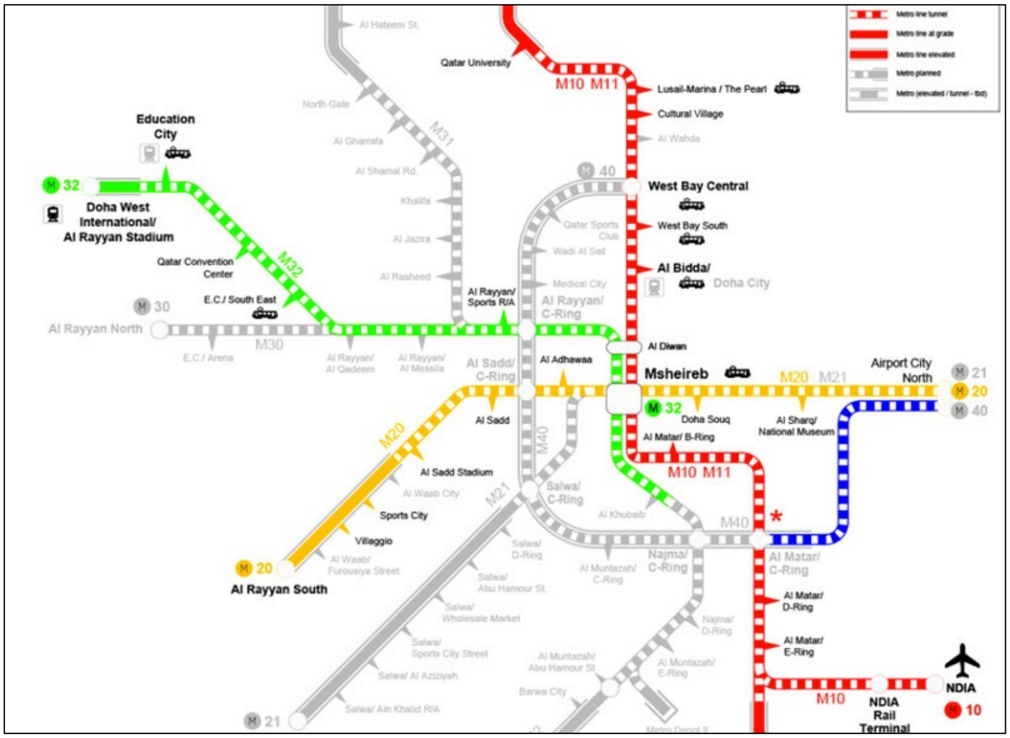


Figure .2 Doha Metro Lines (Source: Qatar Rail)

Ridership data for Phase I of the metro operations were collected from the concerned authorities at the transport planning department from Qatar Rail. The following section provides an overview of the data. The collected dataset is for 37 rail stations with a total of 5627 trip records on an hourly basis taken from February 1st, 2020 until, February 29th, 2020. Figure .3 and Figure .4 represent the relationship between the number of trips attracted to the station and the trips generated from the station and the time of the day at each station. As shown in Figure .3 the trip attraction is highest at 07:00 am at DECC station with the maximum trips of 1918 and a mean of 523.75 trips and a standard deviation of 566.069 trips. Also, DECC has the highest trips generation at 07:00 to 08:00 with a total of 12021 trips and a mean of 961.72 trips, and a standard deviation of 2353.320 trips. On Thursdays, DECC also has the highest trip attraction at 17:00 to 18:00 with several trips of 12021 trips, and a mean of 961.72 trips with a standard deviation of 2353.320 trips. It has the highest trips generation from 7:00 to 08:00 with maximum trips of 14046 having a mean of 1123.52 trips and a standard deviation of 2738.412 trips, as per Figure .4. During Fridays, National Museum Station has the highest trips attraction at 18:00 to 19:00 with a number of trips 8862 in a mean of 708.96 trips and standard deviation of 1773.013 trips, DECC and National Museum Station almost similarly have the highest trips generation from 17:00 to 18:00 with maximum trips of 8784 trips having a mean of 702.72 trips and standard deviation of 1763.974 trips as shown in Figure .4. Table .1 shows the general statistical characteristics for the dependent variables, (o) trips generation, (d) trips attraction, and total ridership (o+d). Table 2 shows the descriptive statistics for various land uses for 37 stations under consideration.

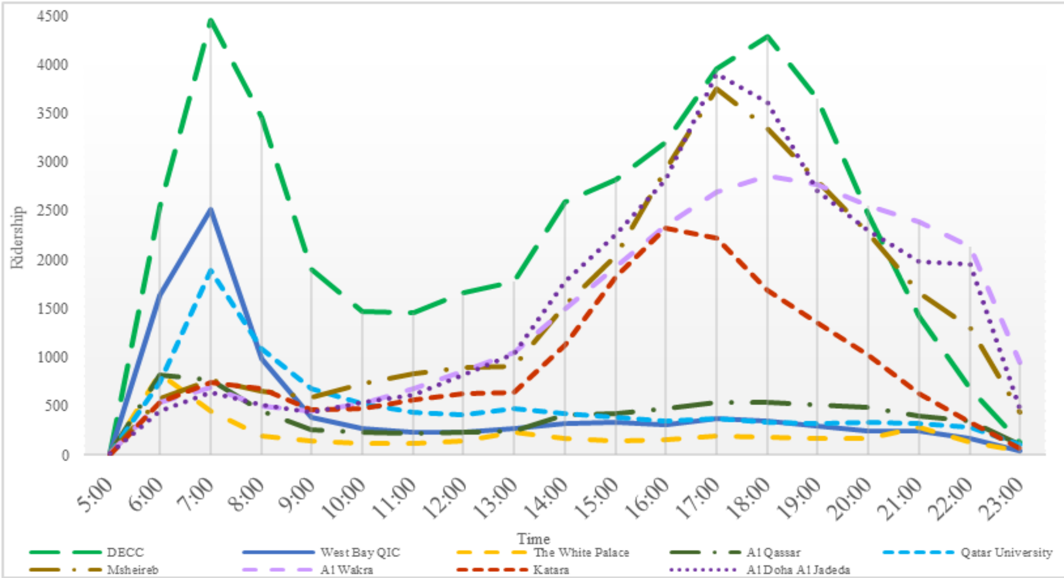


Figure .3 Stations Trips Generation

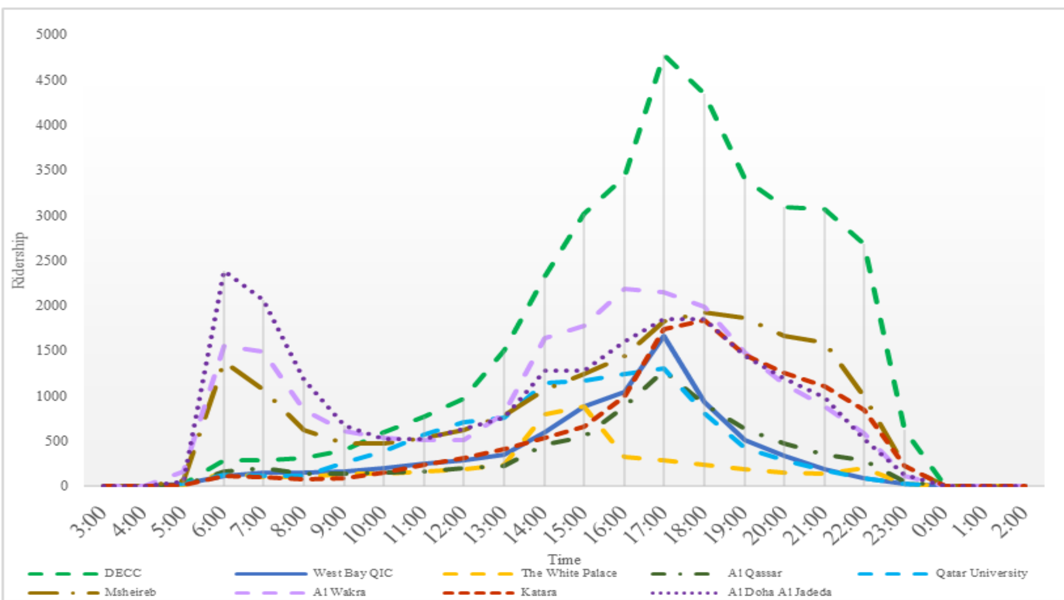


Figure .4 Stations Trips Attracti

Table .1 Descriptive statistics of input features

Statistics	Count	Mean	Std	Min	Max
Educational Facilities Land use density (ER)	2738	0.07	0.19	0	0.99
Governmental Land use density (GR)	2738	0.11	0.16	0	0.58
Mixed-Use density (MR)	2738	0.04	0.06	0	0.21
Residential land use densityRR)	2738	0.32	0.27	0	0.92
Open Space & Recreation (Indoor-Outdoor) density (OR)	2738	0.1	0.12	0	0.41
Religious Facilities land use density (RiR)	2738	0.01	0.01	0	0.04
Retail / Commercial Land use density (CR)	2738	0.04	0.12	0	0.68
Special Use density (SR)	2738	0.16	0.19	0	0.6
Transportation Land use density (TR)	2738	0.12	0.29	0	1
Day_mod	2738	1.49	1.12	0	3
Station_mod	2738	18	10.68	0	36
Ridership	2738	336.4	397.47	0	2881

Table (1) shows the general statistical characteristics for the dependent variables, (o) trips generation, (d) trips attraction, and total ridership (o+d). Table 2 shows the descriptive statistics for various land uses for 37 stations under consideration. Table (3) represents the ridership derived from the actual ridership dataset and land use densities data, (ER) refers to Educational Facilities Land use ridership, (GR) Governmental Land use ridership, (MR) Mixed Use ridership, (RR) Residential land use ridership, (OR) Open Space & Recreation (Indoor-Outdoor) ridership, (RiR) Religious Facilities land use ridership, (CR) Retail / Commercial Land use ridership, (SR) Special Use ridership, (TR) Transportation Land use ridership.

5. Results and Discussion

A very critical step in the SVM modeling is the identification of input features., As previously mentioned in the literature review, rail ridership has a periodic, nonlinear nature. In that regards successive actual values of ridership are needed to gather as inputs, such that the model will be able to predict the projected values accordingly. As an eventual performance-optimized parameter we need to get the optimum value of the penalty coefficient, loss function, and corresponding parameters in kernel function. The following equation represents the objective function used in this research;

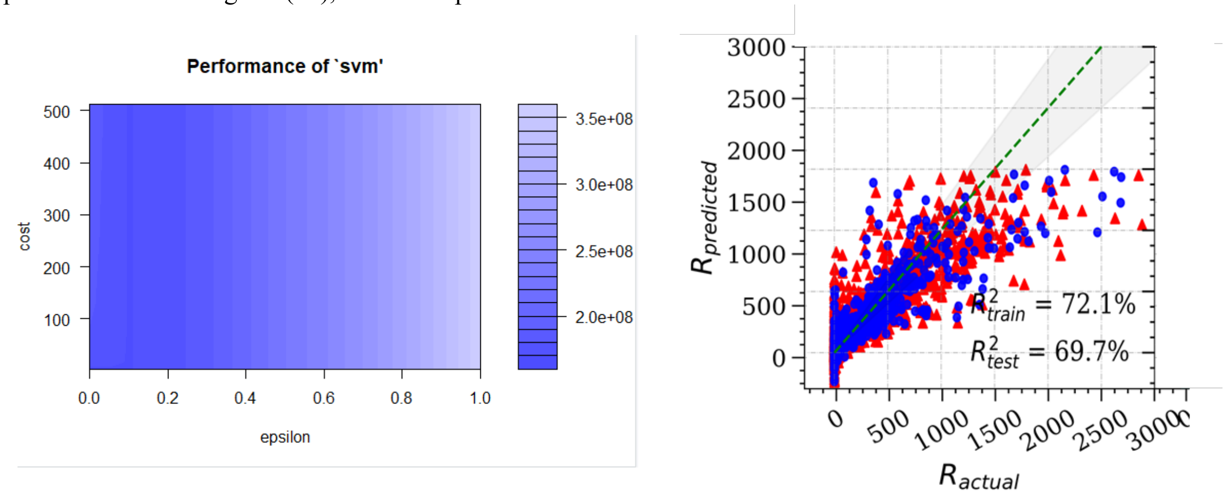
$$\text{Ridership per Land use} = \sum_{=1}^{n} \sum_{k \in K} \left(O_{s_i} + d_{s_i}\right) D_K \tag{12}$$

where, (O) refers to trip generation for ith station, (d) for trips attraction for ith station, and (D) is the land use density for land use (k). In this paper, land use was clustered into nine types; Educational Facilities Land use, Governmental Land use, Mixed Use, Residential land use, Open Space & Recreation, Religious Facilities land use, Retail / Commercial Land use, Special Use, and Transportation Land use.

After splitting data into training and testing subsets, the training data were fitted with the SVM, as a regression model with two tuning parameters that need to be set with cost penalty and epsilon parameters of the radial basis function (Kuhn, 2008). Nine different cost penalties were chosen to test the behavior of the trained function for the dataset, besides one single sigma value using the "kernlab" package. The sigma value is assumed to be constant in this model. To broaden the calculation for the cost penalties the values will be ranged between 1 and 9. The next stage was to tune the model, in which the aim is to select one cost penalty value among all candidate values to give the best performance of the model. Following the model tuning summary of results, the textual output shows the general features of the model indicating that the SVM type is eps-regression, radial kernel, with a gamma value of 0.0909 and an epsilon of 0.1, and the model gives its optimum performance with the cost 4 with a number of support vectors to be 16.
According to Kuhn (2008) for regression models, 80-20 training-testing data should be conducted. Accordingly, in this paper, the authors conducted 80% of all the data as a training set and 20% for testing. The training and testing datasets were created using create Data Partition in the caret package.

5.1 Model performance

To tune the designed model, an epsilon- regression should be performed in order to find the optimum epsilon, i.e., the cost penalty values for the model using hyperparameter optimization in which the model will train different couples of values for epsilon and cost to select the optimum value. After running the model, the best performance for the model is at epsilon 0 and cost 8 as shown in Figure .(5a) where the darkest blue color shows the area where the model performs the best. Figure (5b), shows the predicted and actual of the model.



(a) (b)

Figure .5 SVM performance

5.2 Model Validation

In order to evaluate the behavior of the developed model, three performance indicators were investigated; Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R Squared. MAE is an indicator used to compare the difference between actual and predicted ridership values for each case. Here, xi is the actual ridership, and Yi represents predicted ridership. MAE can be calculated as follows:

$$\text{MAE} = \sum_{i=1}^{n} \frac{|y_i - x_i|}{n} = \sum_{i=1}^{n} \frac{|e_i|}{n} \left(\hat{Y}_i - Y_i\right)^2 \tag{13}$$

For each value of the dataset xi the model predicts a mirror value of Yˆ_i, so in order to test how good the designed model is. The MSE indicators are used to compute how many errors the model makes, thus the value of Y_i is compared to the predicted value Yˆ_i, so the value of the error is reflected by Yˆ_i-Y_i, which means the lower value we have the fewer error values the model predicts For the model designed in this paper, the MAE absolute value was found to be 117 which indicates a small error margin that the model produces with reflects a good behavior of it. The RMSE value is a common indicator used to measure the error in the machine learning models, which is calculated by taking the root of the Mean Squared Error (MSE) (i.e. RMSE=√MSE), for the derived model we had an RMSE absolute value of 204.

$$\text{MSE} = \frac{1}{n} \sum_{i=1} \left(\hat{Y}_i - Y_i\right)^2 \tag{14}$$

The R-squared measure is mostly used to calculate the portion of variance for the dependent variable that could be explained by the independent variables; the following formula is used to calculate the R2:

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}} \tag{15}$$

For the developed model, R^2 value is 0.721, which shows a satisfactory model of ridership variability, the model validation indicators are presented in Table (2).

Table.2 Model Validation indicators values

Indicator	Trained Dataset	Test Dataset
MAE	117	131
RMSE	204	235
R Squared	72.1	69.7

5.3 Model implication for Sustainabel Development Goals SDGs Acheivement

According to the 2030 Agenda for Sustainable Development, paragraph 27 states that "sustainable transport systems, coupled with universal access to affordable, reliable, sustainable and modern energy services, quality and resilient infrastructure and other policies that increase productive capacity, create solid economic foundations for all countries". However, twelve objectives are included in this section which is associated with the transport sector, and five of them are directly associated with the transport sector, while seven of them are indirectly associated with it. Furthermore, the five targets directly associated with transport are: road safety (target 3.6); energy efficiency (target 7.3); sustainable infrastructure (target 9.1), urban access (target 11.2) and fossil fuel subsidies (target 12.c). However, it underlines that sustainable transport is important for accelerating the achievement of the SDGs. On the other hand, the seven goals that are indirectly associated with transport are: agricultural productivity (Target 2.3), air pollution (Target 3.9), access to drinking water (Target 6.1), sustainable cities (Target 11.6), food loss reduction (Target 12.3), adaptation to change (Target 13.1) and climate change mitigation (Target 13.2), Figure .6 maps the urban-transport direct and indirect related SDGs. The authors stress that the appropriate predictive models that incorporate transportation systems and urban planning could be of great help to achieve urban-transport SDGs. Since it gives better data-based insights and predictions for planners and authorities for their decision-making process.

6. Conclusion

This paper presents the authors' investigation on the application of ML model, namely SVM to predict rail transit ridership concerning land use planning in the catchment area around the stations. The prediction performance of the model with actual data showed that SVM significantly performs considering the land use densities around the stations. The study proposes a new SVM model to predict a short-term station-based metro rail ridership considering the land use densities to capture the nonlinear, periodic ridership characteristics using a current metro ridership dataset in the state of Qatar. In this research work, the authors employed a machine learning SVM model, first 80% of the dataset was used for model training, then the model was validated using three indicators namely, MAE, RMSE, and R2. Afterwise 20% of the data were tested and prediction outcomes of the models were compared to the actual ridership records, the difference was slightly satisfactory. However, the appropriate operation of Metro is mainly based on the changes that happened in the ridership nature, where the maximum utilization of the transportation system resources and to timely adjust the operation and management strategies are highly desired. Nevertheless, the proposed model used the total monthly ridership and did not consider the special circumstances such as the weekends, hourly variation, and special events. A broader investigation is needed for special circumstances that affect rail ridership in future work. The proposed model can be used to predict ridership for newly planned developments based on land use densities. Model uses only one variable to predict ridership accurately, (advantage of using a nonlinear and data-driven model) so application to various cases is feasible with the possibility of obtaining close-to-real results. In this paper, we have introduced a new emergence of transportation and urban planning perspectives using Machine Learning techniques, which could be used in different scenarios and situations. Moreover, the paper stresses other previous studies' findings on the efficiency of SVM predictive performance and generalizes its accurate prediction. However, the model needs more improvement in future work represented in examining the accuracy of the predictions of the model on a daily or hourly basis. Further work could be done by examining other ML models for the ridership prediction considering land use. The proposed model can be used by planning and operation authorities in their processes to plan the land use around metro stations predict the transit demand from those plans and achieve the optimal use of the transit system. Results can be applied directly in practice.

Figure .6 Urban-Transport direct and indirect related SDGs

References

Bar-Gera, H., & Boyce, D. (2003). Origin-based algorithms for combined travel forecasting models. *Transportation Research Part B: Methodological*, *37*(5), 405–422. https://doi.org/10.1016/S0191-2615(02)00020-6

Berawi, M. A., Saroji, G., Iskandar, F. A., Ibrahim, B. E., Miraj, P., & Sari, M. (2020). Optimizing land use allocation of transit-oriented development (TOD) to generate maximum ridership. *Sustainability (Switzerland)*, *12*(9), 1–20. https://doi.org/10.3390/su12093798

Cervero, R., & Kockelman, K. (1997). Travel demand and the 3Ds: Density, diversity, and design. *Transportation Research Part D: Transport and Environment*, *2*(3), 199–219. https://doi.org/10.1016/S1361-9209(97)00009-6

Chen, M. C., & Wei, Y. (2011). Exploring time variants for short-term passenger flow. *Journal of Transport Geography*, *19*(4), 488–498. https://doi.org/10.1016/j.jtrangeo.2010.04.003

David, M. J. T. (2003). *ONLINE SVM LEARNING : F R O M CLASSIFICATION TO DATA DESCRIPTION*.

Gao, B., Yang, H., Lin, H.-C., Wang, Z., Zhang, W., & Li, H. (2022). A Hybrid Improved Whale Optimization Algorithm with Support Vector Machine for Short-Term Photovoltaic Power Prediction. *Applied Artificial Intelligence*, *00*(00), 1–33. https://doi.org/10.1080/08839514.2021.2014187

Guan, J., Zhang, K., & Chen, Y. (2020). *How is public transit in the megacity peripheral relocatees ' area in China ? Captive transit rider and dynamic modal accessibility gap analytics in a peripheral large-scale residential area in Shanghai ,*. 1–21.

He, Y., Zhao, Y., & Tsui, K. L. (2020). An adapted geographically weighted LASSO (Ada-GWL) model for predicting subway ridership. In *Transportation* (Issue 0123456789). Springer US. https://doi.org/10.1007/s11116-020-10091-2

Jovicic, G., & Hansen, C. O. (2003). A passenger travel demand model for Copenhagen. *Transportation Research Part A: Policy and Practice*, *37*(4), 333–349. https://doi.org/10.1016/S0965-8564(02)00019-8

Lam, S. H. M., & Toan, T. D. (2008). Short-Term Travel Time Prediction Using Support Vector Regression. *Transportation Research Board 87th Annual Meeting*, *February*, 16p. http://pubsindex.trb.org/paperorderform.pdf

Li, Y., Wang, C., & Gong, J. (2017). A multi-model probability SOC fusion estimation approach using an improved adaptive unscented Kalman filter technique. *Energy*, *141*, 1402–1415. https://doi.org/10.1016/j.energy.2017.11.079

N. Vapnik, V. (2000). *Vladimir-Vapnik-the-Nature-of-Statistical-Learning-Springer-2010.Pdf* (p. 334).

Shao, Q., Zhang, W., Cao, X., Yang, J., & Yin, J. (2020). Threshold and moderating effects of land use on metro ridership in Shenzhen: Implications for TOD planning. *Journal of Transport Geography*, *89*(October), 102878. https://doi.org/10.1016/j.jtrangeo.2020.102878

Smith, B. L., Williams, B. M., & Keith Oswald, R. (2002). Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C: Emerging Technologies*, *10*(4), 303–321. https://doi.org/10.1016/S0968-090X(02)00009-8

Sun, Y., Leng, B., & Guan, W. (2015). A novel wavelet-SVM short-time passenger flow prediction in Beijing subway system. *Neurocomputing*, *166*, 109–121. https://doi.org/10.1016/j.neucom.2015.03.085

Sundaram, S., Koutsopoulos, H. N., Ben-Akiva, M., Antoniou, C., & Balakrishna, R. (2011). Simulation-based dynamic traffic assignment for short-term planning applications. *Simulation Modelling Practice and Theory*, *19*(1), 450–462. https://doi.org/10.1016/j.simpat.2010.08.004

Suthaharan, S. (2001). Support Vector Machineの多値分類問題への適用法について. 情報処理学会研究報告. 自然言語処理研究会報告, *2001*(112), 33–38. https://doi.org/10.1007/978-1-4899-7641-3

Vlahogianni, E. I., Karlaftis, M. G., & Golias, J. C. (2005). Optimized and meta-optimized neural networks for short-term traffic flow prediction: A genetic approach. *Transportation Research Part C: Emerging Technologies*, *13*(3), 211–234. https://doi.org/10.1016/j.trc.2005.04.007

Wang, X., Zhang, N., Zhang, Y., & Shi, Z. (2018). Forecasting of Short-Term Metro Ridership with Support Vector Machine Online Model. *Journal of Advanced Transportation*, *2018*. https://doi.org/10.1155/2018/3189238

Wey, W. M., Zhang, H., & Chang, Y. J. (2016). Alternative transit-oriented development evaluation in sustainable built environment planning. *Habitat International*, *55*, 109–123. https://doi.org/10.1016/j.habitatint.2016.03.003

Yin, Y., & Shang, P. (2016). Forecasting traffic time series with multivariate predicting method. *Applied Mathematics and Computation*, *291*, 266–278. https://doi.org/10.1016/j.amc.2016.07.017

Zhang, N., Zhang, Y., & Wang, X. (2013). Forecasting of short-term urban rail transit passenger flow with support vector machine hybrid online model. *Transportation Research Board 92nd Annual Meeting Compendium of Papers*, 16p.

Zhao, X., Yan, X., Yu, A., & Van Hentenryck, P. (2020). Prediction and behavioral analysis of travel mode choice: A comparison of machine learning and logit models. *Travel Behaviour and Society*, *20*(January), 22–35. https://doi.org/10.1016/j.tbs.2020.02.003

Biographies

Aya Hasan Alkhereibi. is a Ph.D Candidate in the Department of Industrial and Mechanical Engineering at Qatar University. Serves as a Graduate Assistant at Qatar Transportation and Traffic Safety Center (QTTSC) in College of Engineering. Eng. Aya holds a bachelor of Architecture from IUG, Palestine, and a master's degree in Engineering Management from IUG, Palestine. Eng. Aya performs diverse research in Transportation and Urban Planning, Smart Mobility, Machine learning, and sustainability. Moreover, she worked in construction projects management as a technical office engineer since 2013 before joining academia and research work. In addition to research, Miss Aya is interested in Entrepreneurship and continuous improvement, she gained a graduate diploma in Entrepreneurship and Business Application and has been awarded as a top10 project at the Arab Innovation Academy, Science and Technology Park (QSTP), Qatar 2019. Besides professional and research work, miss Aya is highly interested in sustainable development and environment-friendly solutions that can be in help of improving livability in cities.

Murat Kucukvar, Dr. Murat Kucukvar serves as an Associate Professor in the Department of Industrial and Systems Engineering at Qatar University. Dr. Kucukvar is an expert in Sustainable Cities and Societies, Sustainable Operations Management, Supply Chain Management, and Transportation, Modeling and Simulation, Multi-Criteria Decision Making and Optimization.

Nuri Onat, Dr. Nuri C. Onat serves as an Assistant Professor at the Qatar Transportation and Traffic Safety Center (QTTSC) at Qatar University. He is also the Co-founder and Co-director of Sustainable Systems & Solutions Lab (S3-Lab). Prior to joining Qatar University, he worked as an Assistant Professor in the Department of Industrial Engineering at the City University of Istanbul. He is an Emeritus Walton Postdoctoral Research Fellow in Julie Ann Global Sustainability Institute at the Arizona State University as part of the Walton Sustainability Fellowship Program. Dr. Onat received his Ph.D. from the Civil, Environmental, and Construction Engineering Department at the University of Central Florida (UCF). Onat completed his M.S. in the Department of Civil & Coastal Engineering, under the Engineering School of Sustainable Infrastructure & Environment (ESSIE) at the University of Florida (UF), where he also received a Certificate of Engineering Entrepreneurship from Engineering Innovation Institute. He also has a second master's degree from the Department of Industrial Engineering & Management Systems at UCF. Prior to starting his academic studies, he worked as a Civil Engineer in several engineering firms in Turkey, specializing in renewable power generation. He completed his B.S. in the Department of Civil Engineering at Gazi University, Turkey.

Jaser Mohammad Saleh Mustafa. is a Master Candidate in the Department of Industrial and Mechanical Engineering at Qatar University. Serves as an executive manager at Seven Dimensions Consultant W.L.L. Eng. Jaser holds a bachelor of Architectural Engineering from JU, Jordan. Eng. Jaser Manages diverse range of projects in Qatar through the use of the Building Information Modeling Technology in the field of Design and Construction. Moreover, he works as an instructor of many engineering projects from software to project management professional (PMP,PMI) to BIM Management Professional since 2012. In addition to education certificates, Mr. Jaser holds some international recognized certifications, like but not limited to; AUTODESK certified professional, AUTODESK certified instructor, AUTODESK authorized developer, author and published, RICS certificate in BIM project management, and Project Management Professional (PMP, PMI), besides the professional construction and engineering practice Eng. Jaser is interested in Research in the fields of latest technologies, BIM, Sustainability, and Circular Economy, in addition to his interest in continuous improvement and skills development and volunteer works.