International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 10, pp 25-31, October-2024

https://doi.org/10.47001/IRJIET/2024.810005

Urban Traffic Congestion Prediction Using GTFS Data and Advanced Machine Learning Models

Ali Atta Gheni

Department of Computer and Communications, Faculty of Engineering, Islamic University of Lebanon, Wardanieh, Lebanon

Abstract **- Urban traffic congestion represents a complex challenge influenced by many dynamic factors. Peak periods typically exacerbate congestion, while bad weather can slow vehicle movements and increase travel times. Accidents and road closures cause sudden and unexpected disruptions, making traffic management a constant challenge. Using a dataset of over 66,000 GTFS records with machine learning classifiers like Random Forest, XGBoost, CatBoost, and Decision Tree models, the study seeks to forecast traffic conditions. SMOTE is used to ensure greater representation of minority classes in order to solve the dataset's intrinsic imbalance, and feature scaling enhances model convergence. With an accuracy of 98.8%, Random Forest was the most accurate model for this challenge. The outcomes demonstrate that the system is able to precisely forecast traffic in real-time, which aids in route planning, traffic control, and enhancing urban mobility.**

Keywords: Machine Learning, GTFS, Random Forest, Cross-Validation, Real-Time Prediction, Congestion Detection.

I. INTRODUCTION

One of the most important issues that cities throughout the world are dealing with is urban traffic congestion, which has a big influence on daily commutes as well as the general economic and environmental well-being of urban regions. Traffic congestion has become a recurring problem due to population growth and urbanization, which lowers quality of life and increases travel times, fuel consumption, and air pollution [1]. Because of their reliance on static or sparse real-time data, traditional traffic management systems are ill-equipped to handle the complex and dynamic character of contemporary traffic patterns. Due to their incapacity to comprehend and respond to changing conditions, such as traffic incidents, weather variations, or sudden surges in traffic demand during special events, these systems frequently fail to produce accurate, timely predictions of congestion [2]. Consequently, the demand for increasingly sophisticated, data-driven traffic control solutions is rising.

Through the analysis of extensive and varied datasets, machine learning (ML) models are able to identify intricate patterns in traffic situations that conventional approaches frequently miss. With the use of historical and current data, these models can accurately forecast traffic volumes while taking the time of day, the weather, and traffic accidents into consideration. With the use of this predictive ability, commuters can plan their routes with confidence, cutting down on delays and maximizing travel time [3]. These models provide more effective traffic management for municipal traffic authorities, including better resource allocation, dynamic signal control, and enhanced incident response. In the end, machine learning improves urban mobility and traffic flow.

General Transit Feed Specification (GTFS), a widely used standard for data related to public transportation, offers comprehensive details on the routes, stops, schedules, and journeys for transit systems. Because of its uniform format, which makes it possible to collect data consistently and thoroughly across different cities, GTFS is an invaluable tool for developing predictive models [4]. Because machine learning models can be trained on huge, diversified datasets thanks to its widespread use, predictions of traffic congestion are more reliable and accurate. Through the utilization of GTFS, scholars and urban planners can enhance public transportation services, maximize traffic flow, and gain a deeper understanding of urban transit patterns. Because of its richness, this data is crucial for real-time traffic control and smart city applications, supporting sophisticated analytics and machine learning methods.

Real-time traffic forecasts can improve the whole urban transportation experience by assisting travelers in selecting alternate routes, avoiding crowded areas, and cutting down on travel time. Precise estimates of traffic congestion enable traffic authorities to use resources more effectively, resulting in enhanced traffic signals, variable toll rates, and faster emergency response times. With the use of advanced methods for machine learning and GTFS data, this study seeks to develop a reliable traffic congestion forecast model. Urban traffic management systems that are more intelligent, flexible, and condition-responsive will benefit from the study's findings.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 10, pp 25-31, October-2024

https://doi.org/10.47001/IRJIET/2024.810005

II. RELATED WORK

Traffic congestion has a direct and indirect impact on a country's economy and its dwellers' health. According to Ali et al. [5], traffic congestion causes Pak Rs. 1 million every day in terms of opportunity cost and fuel consumption due to traffic congestion. Traffic congestion effects on individual level as well. Time loss, especially during peak hours, mental stress, and the added pollution to the global warming are also some important factors caused due to traffic congestion. The problem of traffic congestion forecasting can be defined as estimating parameters related to future traffic congestion in the short-term using aggregated traffic data. There are usually five parameters to evaluate, including traffic volume, traffic density, occupancy, traffic congestion index, and travel time during traffic congestion monitoring and forecasting [6]. Depending on the nature of the collected data, a variety of Artificial Intelligence (AI) methods are applied to evaluate congestion parameters.

In general, traffic detectors can be classified in two main groups [7]: sensor-based and vision-based monitoring approach. Vision-based techniques use features obtained from images usually acquired vias cameras. In [8], a deep convolutional neural network was devised to count the number of vehicles on a road segment based solely on video images. The methodology does not regard an individual vehicle as an object to be detected separately; rather, it collectively counts the number of vehicles as a human would. Every second counts when working with "soft real-time" systems, such as intelligent traffic. Therefore, the sensors will not work properly if there is too much noise or if the data is delayed. In [9], a new model was introduced to collect data from the intelligent traffic system controller taking advantage of the time series model. Before use, data from nearby signals is pre-processed.

Big data mining and machine learning for smart cities utilization assists in solving production, transport, and traffic management problems in real-time approaches using frameworks and systems that are incorporated and provide data transfer efficiently through apps and stakeholders. In [10], a methodology for predicting traffic congestion was provided. Several machine learning algorithms and approaches are compared to select the most appropriate one. The methodology was implemented using Data Mining and Big Data techniques along with Python, SQL, and GIS technologies. Evaluation and results have shown that data quality and size were the most critical factors towards algorithmic accuracy. Result comparison showed that Decision Trees were more accurate than Logistic Regression. In [11], using machine learning models and image processing an intelligent strategy to mitigate congestion at traffic lights is presented. Data is collected from vehicles and cars at traffic signal intersections, then the quantity and type of vehicles present are determined and the emergency vehicle is distinguished from a regular vehicle. The solution uses image processing and the Yolov3 algorithm to handle the complex traffic scenarios that arise. In addition, traffic lights in this setup coordinate with each other via a wireless medium in order to give priority to emergency vehicles. According to the results of tests and experiments, the system has the ability to reduce the usual time that vehicles wait at traffic intersections by more than 55% and process emergency vehicles as quickly as possible. In [12], long short-term memory networks for the prediction of congestion propagation across a road network were presented. Based on vehicle speed data from traffic sensors at two sites, the model predicts the propagation of congestion across a 5-min period within a busy town. The results show that long short-term memory networks are suitable for predicting congestion propagation on road networks and may form a key component of future traffic modelling approaches for smart and sustainable cities around the world.

In a modern city many different sensors can be used for information collection. Algorithms that are cast-off in machine learning improves the capabilities and intelligence of a system when the amount of data collected increases. In [13], a system model to analyze traffic congestion in the environment of a smart city was presented. The model comprises an ML-enabled IoT-based road traffic congestion control system whereby the occurrence of congestion at a specific point is notified. In [14], traffic flow situations were classified as congested or not congested in Pakistan. A large number of images were created from video recordings taken in Karachi, Pakistan. These images were then used to build the dataset. Generative Adversarial Network (GAN) was used to produce images from the provided dataset and improve the quality of the images. A five-layer Convolutional Neural Network (CNN) model was developed and deployed to differentiate between images containing crowding and those without crowding. The hybrid Xception Support Vector Machine (XPSVM) classification model was used in [15] in order to give estimates of short-term traffic congestion. Xception classifier relies discrete convolution algorithms in order to predict which feature detection will occur within the dataset. The largest marginal separations are used by the Support Vector Machine (SVM) classifier in order to provide more accurate predictions regarding the output. This is achieved through the use of weight regulation and a precisely calibrated dual super-plane mechanism.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 10, pp 25-31, October-2024

https://doi.org/10.47001/IRJIET/2024.810005

III. METHODOLOGY

Urban traffic patterns are constantly changing and influenced by many conditions including time of day, unique events, and unexpected accidents. Accurate congestion forecasting requires a system that can adapt to changing conditions in real time. Integrating real-time data from multiple sources, including cameras and IoT sensors, poses challenging situations in terms of handling massive amounts of information, sorting it out and determining what is most important. Figure 1 shows the proposed method for detecting traffic congestion using machine learning techniques.

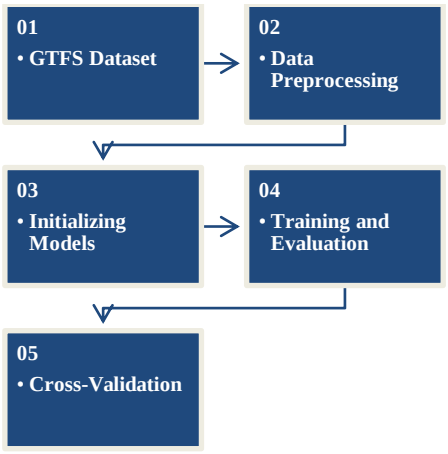

Figure 1: Proposed method for real-time urban traffic congestion prediction

3.1 GTFS Dataset

General Transit Feed Specification (GTFS) is globally widespread and facilitates the development of public transport structures in urban environments. The dataset used is “GTFS Traffic Prediction Dataset” obtained from Kaggle. The collection includes 366 routes, 5624 stops and 21804 trips. The data was classified into four unique categories of 4 degrees of traffic congestion using the Speed Reduction Index (SRI) and organized into a set of 66913 records and 9 columns. These columns are the starting stop number, the arrival stop number, the trip number, the arrival time, the time period required to move between two consecutive destinations, the speed, the number of trips, and “SRI”, which is the speed reduction index that determines the relative speed ratio between the congested flow and the free flow of traffic [16]. Target column is the "Degree_of_congestion", which contains four classifications (“Very smooth”, “Smooth”, “Mild congestion”, and “Heavy congestion”). These classifications are determined based on the SRI value, and the values of this column are distributed as in Figure 2. Statistics reveal that the level of congestion labeled “Very smooth” occurs more frequently while “Heavy congestion” occurs less frequently, indicating that instances of severe traffic congestion are less prevalent and that the data set is unbalanced.

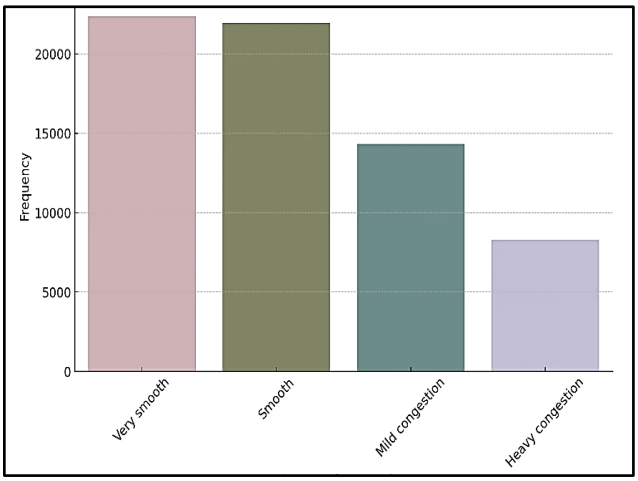

Figure 2: Distribution of "Degree_of_congestion" column

3.2 Data Preprocessing

To process the data set, a number of steps were applied that produced a clean data set ready for modeling:

1. Label encoding for target variable to convert to digital format.
2. Filtering rows based on the number of trips and excluding trips with fewer than 20 instances, ensuring that the dataset is aware of properly represented instances, thus enhancing its robustness.
3. Dropping 'arrival_time' column to simplify the data set because it is less important.
4. Processing of the Time column, including multiplication by thousand and subsequent rounding. This process improves the interpretability of temporal features and aligns values to a mandated time unit, enhancing consistency in subsequent analyses.
5. Handling missing values and assessment for large values and infinity is conducted, with corresponding rows omitted to ensure the dataset's numerical stability and suitability for subsequent modeling endeavors.
6. Feature scaling using Standard Scaler, which enhances convergence and balance.
7. Applying the SMOTE algorithm to address the imbalance. A stratified sampling technique is used to improve minority class representation, enhancing the system's ability to identify patterns across all classes.
8. The dataset was split 70% for training and 30% for testing.

Table 1 shows the final size of the training and test sets after processing.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 8, Issue 10, pp 25-31, October-2024
https://doi.org/10.47001/IRJIET/2024.810005

Table 1: Size of the training and test sets after pre-processing

	Number of Rows	Number of Columns
Training Set	42000	5
Testing Set	18000	5

3.3 Initializing Models

There are a number of factors that decide which machine learning technique is the most appropriate for a traffic congestion prediction system. These factors include the characteristics of the data, the size of the data set, and the needs that are specific to the application. Several machine learning algorithms that are typically utilized in forecasting and projects linked to traffic will be evaluated in order to determine which one is the most effective:

1. Decision Tree classifier constructs a tree-like model by recursively dividing data based on input characteristics. Each internal node represents a decision, and each leaf node signifies a class label or prediction within the constructed tree.
2. Random Forest which randomly sample from the dataset, generates a decision tree for each sample, and use these trees to predict. Assign a vote for each anticipated result. Identify the prediction outcome with the most votes as the final forecast.
3. CatBoost efficiently handles categorical features by transforming data during training, capturing non-linear connections and managing missing values. It iteratively refines predictions using a gradient boosting framework
4. XGBoost combines decision trees through gradient boosting, progressively correcting mistakes for accurate predictions and producing a robust final model.

3.4 Training and Evaluation

Classifiers are trained by fitting each classifier to the training data (and then making predictions on the test to compare classifier performance. To evaluate the effectiveness of a multi-class classification model, a confusion matrix is calculated that provides insight into the model's ability and accuracy to correctly classify instances of each class and allows the precision, recall, and F1 score for each class to be calculated. Figure 3 shows the confusion matrix of the Random Forest classifier.

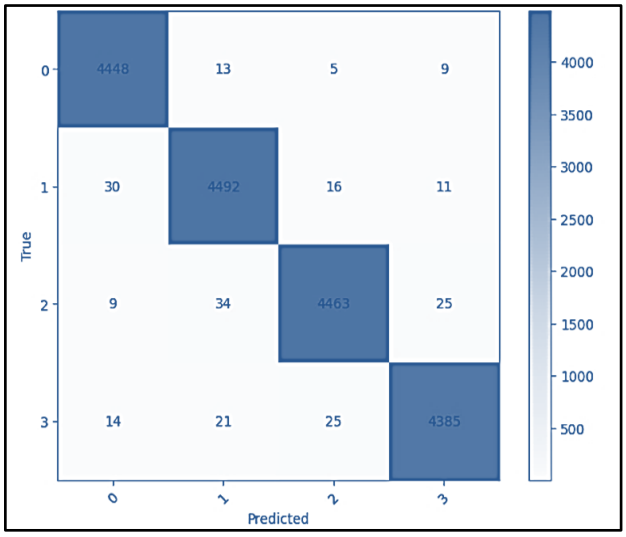

Figure 3: Confusion matrix of Random Forest

Table 2 compares the results of the four models. High accuracy was achieved in all models, indicating great ability for accurate classification. All models show consistently excellent precision and recall results, indicating that they are capable of accurately identifying and classifying positive cases. In general, the random forest model achieved the highest results and will be adopted for the classification task in our study.

Table 2: Comparing classifiers results

Classifier	Accuracy	Precision	Recall
Decision Tree	0.95839	0.95839	0.95839
Random Forest	0.98822	0.98822	0.98822
Cat-Boost	0.96694	0.96694	0.96694
XGBoost	0.96778	0.96778	0.96778

3.5 Cross-Validation

Cross-validation is an approach used in machine learning to evaluate the performance and generality of a predictive model [17]. It reduces the risk of overfitting and provides a more reliable indication of how the model will perform on a different set of data. A subset of the training data is used in each iteration to verify and simulate the overall model performance based on the unseen data. The random forest model underwent 5-fold cross-validation using the available features and the target variable resulting in the results shown in Figure 4. The validation set results, which always exceed 96%, indicate a reliable, efficient, and generalizable model.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 10, pp 25-31, October-2024

https://doi.org/10.47001/IRJIET/2024.810005

```
'Training Accuracy scores': array([0.97877976, 0.97818452, 0.97872024, 0.97785714, 0.97904762]),
'Mean Training Accuracy': 97.85178571428573,
'Training Precision scores': array([0.9787876 , 0.97819101, 0.97871262, 0.97784545, 0.97902993]),
'Mean Training Precision': 0.9785133220455078,
'Training Recall scores': array([0.97877198, 0.97817688, 0.97871382, 0.97784938, 0.97903937]),
'Mean Training Recall': 0.9785102866251624,

'Validation Accuracy scores': array([0.96261905, 0.9652381 , 0.96392857, 0.96619048, 0.96190476]),
'Mean Validation Accuracy': 96.39761904761905,
'Validation Precision scores': array([0.96264726, 0.96527922, 0.96391172, 0.96617953, 0.96186753]),
'Mean Validation Precision': 0.963977054617392,
'Validation Recall scores': array([0.96260033, 0.96524681, 0.96391457, 0.96617244, 0.96187926]),
'Mean Validation Recall': 0.9639626811383636,
```

Figure 4: Cross-validation results of Random Forest

IV. ANALYZE THE RESULTS

All classifiers performed well, but Random Forest's ability to handle large data sets and reduce variance made it the most effective. Precision and recall metrics were consistently high across all models, reflecting their reliability in classifying different crowding levels. The use of SMOTE and cross-validation ensures the robustness of the model and its ability to generalize to unseen data. Table 3 shows a comparative analysis of the results of the Random Forest model used in our study with several previous works. While early works relied on straight image data, later studies explored weather and traffic data, achieving notable improvements in accuracy.

Table 3: Comparing with related work

Ref.	Dataset	Model	Results
Chung & Sohn 2017 [8]	23164 images were labeled manually	Deep Convolutional Neural Network	Accuracy of 93%
Ata et al. 2019 [9]	Dataset from the internet which shows the weather report and traffic speed of M1 junction 37 England at the interval of 10 minutes.	Artificial Neural Networks	Accuracy of 97.56%
Mystakidis & Tjortjis 2020 [10]	Data originating from one of the most problematic streets in Thessalonikiin Greece	Decision Tree Logistic Regression	Best accuracy by Decision Tree: 73%
Faraj & Boskany 2020 [11]	Images and videos from cameras	Yolov3	Accuracy of 89%
Majumdar et al. 2021 [12]	Traffic data of Buxton, UK	Long short-term memory model	Accuracy of 84–95% depending on the road layout
Khan et al. 2021 [13]	Data collected from Data Mill North via the internet consisted of weather and traffic flow at intervals of 10 min	Support Vector Machine	Accuracy of 97.9%
Jilani et al. 2022 [14]	Collected images enhanced and augmented using generative adversarial network	Five-layered CNN model	Accuracy of 98.63%
Anjaneyulu & Kubendiran 2022 [15]	Taking a snapshot using the online service provider Google Maps with the Selenium tool	Hybrid deep learning model	Accuracy of 97.16%
Our Study	GTFS Traffic Prediction Dataset	Random Forest	Accuracy of 98.822%

Research [18] also used GTFS data from the "GTFS Traffic Prediction Dataset". The Decision Tree model outperformed RNN models based on time series such as LSTMs with an accuracy of 81%. Our results outperform those of the previous study as shown in Table 4. These results show that the proposed model is more accurate in producing correct predictions in general.

Table 4: Comparing with a study that used the same data

	Best Model	Accuracy	Precision	Recall
Bannur et al. 2022 [18]	Decision Tree	0.81	0.81	0.81
Our Study	Random Forest	0.98822	0.98822	0.98822

ISSN (online): 2581-3048
Volume 8, Issue 10, pp 25-31, October-2024
https://doi.org/10.47001/IRJIET/2024.810005

V. CONCLUSION

Using advanced machine learning models and GTFS data, this research successfully established a robust system for real-time traffic congestion prediction. The method showed good accuracy across multiple models by classifying traffic conditions into four levels: Very Smooth, Smooth, Mild Congestion, and Heavy Congestion—based on the SRI. With an accuracy of 98.8%, Random Forest fared better than the other examined classifiers. The model's efficiency and generalizability were guaranteed by the use of methods like SMOTE to handle data imbalance and Standard Scaler for feature scaling. Additionally, the 5-fold cross-validation procedure reduced the likelihood of overfitting by further validating the model's dependability. City planners may optimize traffic flow and lessen congestion thanks to this method, which has important implications for urban traffic management. Future research can use real-time IoT and sensor data for dynamic congestion prediction, even if the current method focuses on static GTFS data. Enhancing urban mobility and transportation efficiency, the findings have practical applications in smart city development.

REFERENCES

[1] C. Arti, G. Sharad, K. Pradeep, P. Chinmay, and S. S. Kumar, "Urban traffic congestion: Its causes-consequences-mitigation," *Res J Chem Environ*, vol. 26, no. 12, pp. 164-176, 2022.

[2] T. Ji, Y. Yao, Y. Dou, S. Deng, S. Yu, Y. Zhu, andH. Liao, "The impact of climate change on urban transportation resilience to compound extreme events," *Sustainability*, vol. 14, no. 7, 2022.

[3] S. Majumdar, M. M. Subhani, B. Roullier, A. Anjum, andR. Zhu, "Congestion prediction for smart sustainable cities using IoT and machine learning approaches," *Sustainable Cities and Society*, vol. 64, 2021.

[4] J. P. B. Vieira, R. H. Pereira, and P. R. Andrade, "Estimating public transport emissions from General Transit Feed Specification data," *Transportation Research Part D: Transport and Environment*, vol. 119, 2023.

[5] M. S. Ali, M. Adnan, S. M. Noman, and S. F. A. Baqueri, "Estimation of traffic congestion cost-a case study of a major arterial in Karachi," *Procedia Engineering*, vol. 77, pp. 37-44, 2014.

[6] M. Akhtar, and S. Moridpour, "A review of traffic congestion prediction using artificial intelligence," *Journal of Advanced Transportation*, Jan. 2021.

[7] N. K. Jain, R. K. Saini, and Mittal, "A review on traffic monitoring system techniques," *Soft computing: Theories and applications: Proceedings of SoCTA 2017*, vol. 742, pp. 569-577, Aug. 2019.

[8] J. Chung, and K. Sohn, "Image-based learning to measure traffic density using a deep convolutional neural network,"*IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 5, pp. 1670-1675, Aug. 2018.

[9] A. Ata, M. A. Khan, S. Abbas, G. Ahmad, and Fatima, "Modelling smart road traffic congestion control system using machine learning techniques," *Neural Network World*, vol. 29, no. 2, pp. 99-110, 2019.

[10] A. Mystakidis, and C. Tjortjis, "Big data mining for smart cities: predicting traffic congestion using classification," *In 2020 11th International Conference on Information, Intelligence, Systems and Applications*, Jul. 2020.

[11] M. A. Faraj, andN. W. Boskany, "Intelligent traffic congestion control system using machine learning and wireless network," *UHD Journal of Science and Technology*, vol. 4, no. 2, pp. 123-131, 2020.

[12] S. Majumdar, M. M. Subhani, B. Roullier, A. Anjum, and R. Zhu, "Congestion prediction for smart sustainable cities using IoT and machine learning approaches," *Sustainable Cities and Society*, vol. 64, Jan. 2021.

[13] A. Ata, M. A. Khan, S. Abbas, M. S. Khan, and G. Ahmad, "Adaptive IoT empowered smart road traffic congestion control system using supervised machine learning algorithm," *The Computer Journal*, vol. 64, no. 11, pp. 1672-1679, Nov. 2021.

[14] U. Jilani, M. Asif, M. A. A. Siddique, S. M. Talha, and M. Aamir, "Traffic congestion classification using GAN-Based synthetic data augmentation and a novel 5-layer convolutional neural network model," *Electronics*, vol. 11, no. 15, Jul. 2022.

[15] M. Anjaneyulu, and M. Kubendiran, M. "Short-Term Traffic Congestion Prediction Using Hybrid Deep Learning Technique," *Sustainability*, vol. 15, no. 1, Dec. 2022.

[16] M. Kumar, K. Kumar, and P. Das, "Study on road traffic congestion: A review," *Recent Trends in Communication and Electronics*, pp. 230-240, 2021.

[17] B. Ghojogh, and M. Crowley, "The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial," *arXiv preprint arXiv:1905.12787*, May. 2019.

[18] C. Bannur, C. Bhat, G. Goutham, and H. R. Mamatha, "General Transit Feed Specification Assisted Effective Traffic Congestion Prediction Using Decision Trees and Recurrent Neural Networks," *In 2022 IEEE 1st International Conference on Data, Decision and Systems (ICDDS). IEEE*, pp. 1-6, Dec. 2022.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 8, Issue 10, pp 25-31, October-2024

https://doi.org/10.47001/IRJIET/2024.810005

Citation of this Article:

Ali Atta Gheni. (2024). Urban Traffic Congestion Prediction Using GTFS Data and Advanced Machine Learning Models. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 8(10), 25-31. Article DOI https://doi.org/10.47001/IRJIET/2024.810005

****

