Industrial Internship Report on

” Forecasting of Smart city traffic patterns”

Prepared by

TIRTHA SUBHRA

MUKHERJEE

Executive Summary

This report provides details of the Industrial Internship provided by upskill Campus and The IoT Academy in collaboration with Industrial Partner UniConverge Technologies Pvt Ltd (UCT).

This internship was focused on a project/problem statement provided by UCT. We had to finish the project including the report in 6 weeks’ time.

The prediction accuracy is low, the prediction is challenging, and the prediction effect varies across different geographic regions for smart city traffic flow prediction in the era of big data and industry 4.0. Based on Industry 4.0 and big data analytic applications, this article suggests a projection for smart city traffic communication. In the beginning, this study theoretically presents the application scenario of big data on urban traffic faults and analyzes the traits of associated issues, particularly the fault issues. Second, the application analysis of the PVHH, IDT, and Ford-Fulkerson algorithms is used, and the AC traffic prediction method is explored. Finally, traffic flow forecasting and analysis are done using the three methods mentioned above.

The key findings from the research demonstrate that for all of the evaluated experimental circumstances, approaches without a convolutional component perform well.

This internship gave me a very good opportunity to get exposure to Industrial problems and design/implement solution for that. It was an overall great experience to have this internship.

TABLE OF CONTENTS

1 Preface

Traffic flow forecasting is the most important part of any traffic management system in a smart city. This can help drivers choose the most optimal route to their target destination. Air pollution data is often related to traffic congestion and there is a lot of research on the link between air pollution and traffic congestion using different machine learning methods. A scheme to efficiently predict traffic flow using synthetic techniques such as encapsulation and air pollution has yet to be introduced. Therefore, there is a need for a more accurate traffic flow forecasting system for smart cities. The objective of this study was to predict traffic volume from pollution data. The contribution is twofold:

First, a comparison is made using different simple regression techniques to determine the best performing model. Second, synchronous bagging and stacking techniques were used to find a more accurate model of the two comparisons. The results showed that the K-Nearest Neighbors (KNN) bagging group performed significantly better than all other regression models used in this study. Experimental results show that the KNN bagging ensemble model helps to reduce the error rate by more than 30% in predicting traffic congestion.

Today, cities face the challenge of sustainable mobility. Traffic condition forecasting plays an important role in easing traffic congestion in urban areas. For example, predicting travel time is an important issue in route planning and navigation applications. Furthermore, the increasing penetration of information and communication technology makes Floating Automotive Data an important source of real-time data for intelligent transportation system applications.

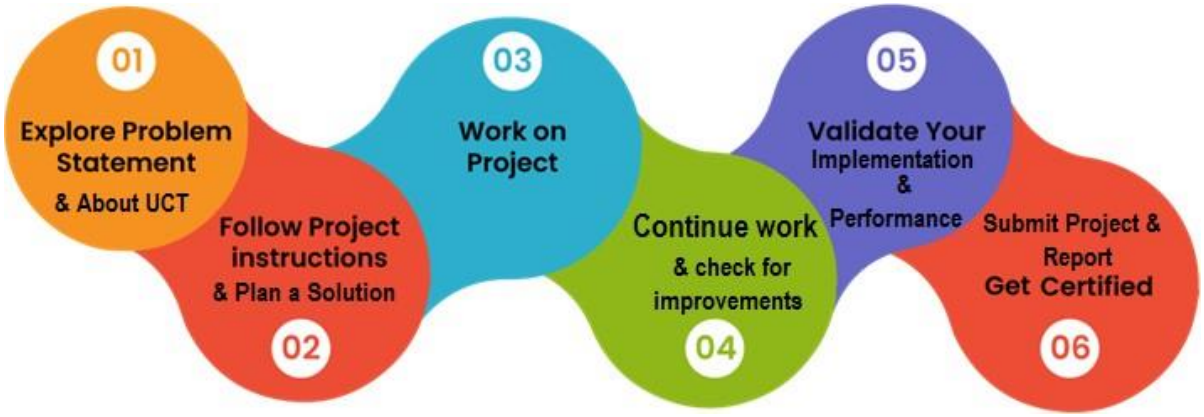
As part of a pedagogical cooperation agreement with the inLab FIB research group focusing on smart mobility, this thesis addresses the problem of urban traffic forecasting when Floating Car Data is available. The main objective is to perform traffic forecasting using machine learning methods and evaluate this type of solution under different conditions:

network size, floating car penetration rate, prediction time and amount of data required.

The current state of the art shows that most of the proposed new methods for urban traffic forecasting use deep learning. Comparison of four neural network approaches (recurrent and convolution) are presented to evaluate the ability of deep learning methods to solve problems.

Various tests are proposed to evaluate the developed Deep Learning models and also to analyze how different proposed factors affect the accuracy of the forecasts. The experiments performed are designed through micro traffic simulation method to simulate data from floating cars. The main conclusions of the obtained results show that methods without convolutional components exhibit the best performance for all tested scenarios.

How Program was planned



Thank to all (Nishank Shakyawar), who have helped you directly or indirectly.

2 Introduction

2.1 About UniConverge Technologies Pvt Ltd

A company established in 2013 and working in Digital Transformation domain and providing Industrial solutions with prime focus on sustainability and RoI.

For developing its products and solutions it is leveraging various **Cutting Edge Technologies e.g. Internet of Things (IoT), Cyber Security, Cloud computing (AWS, Azure), Machine Learning, Communication Technologies (4G/5G/LoRaWAN), Java Full Stack, Python, Front end** etc.

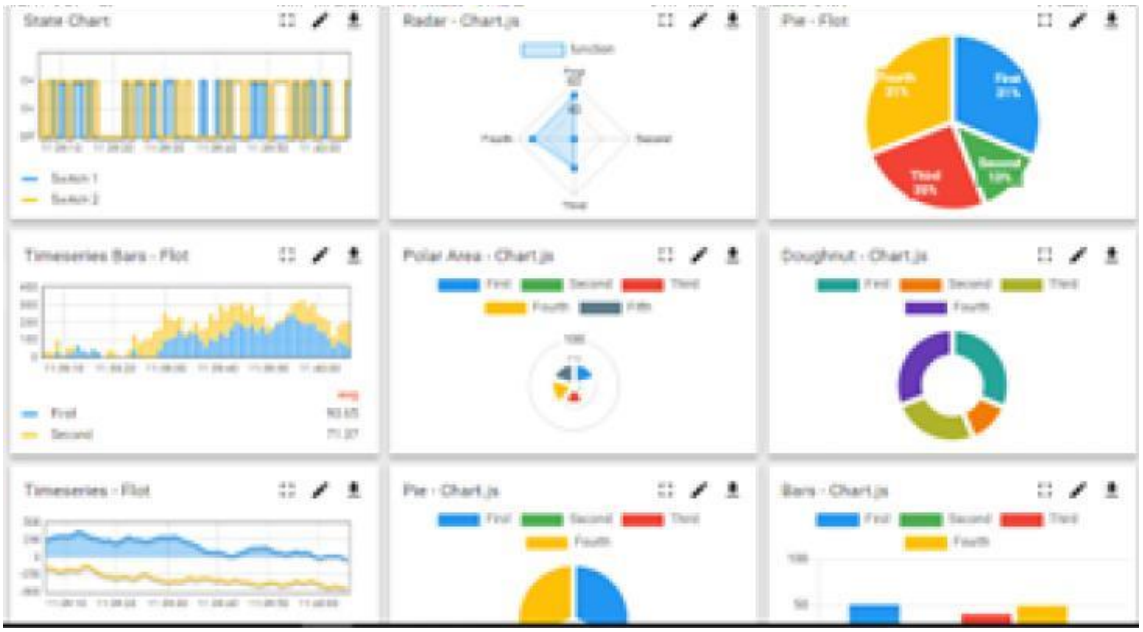
i. UCT IoT Platform (UCT Insight)

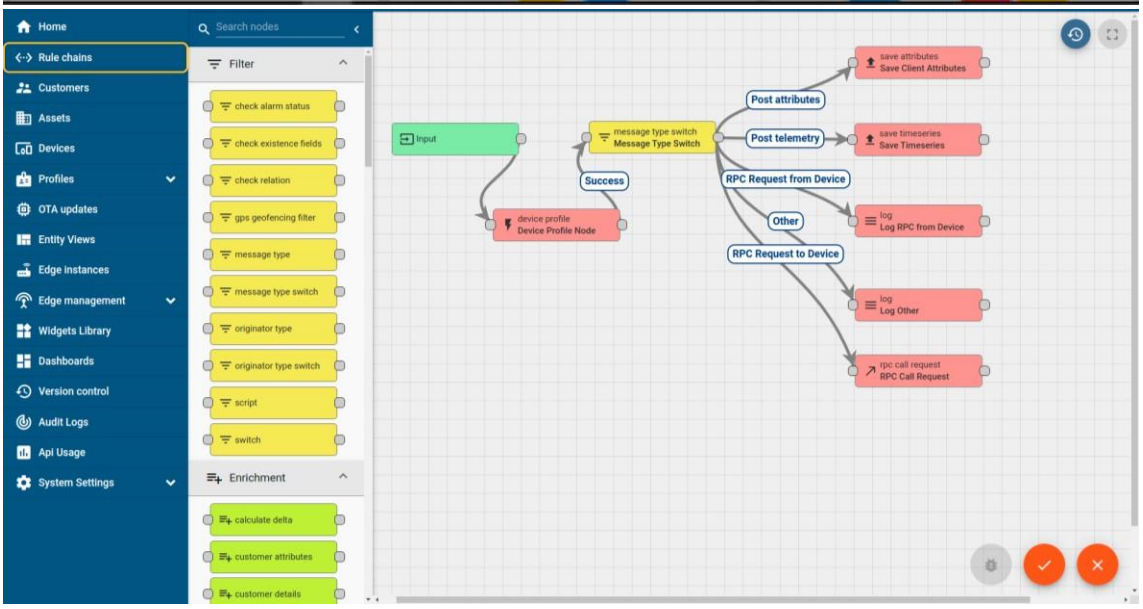
UCT Insight is an IOT platform designed for quick deployment of IOT applications on the same time providing valuable “insight” for your process/business. It has been built in Java for backend and ReactJS for Front end. It has support for MySQL and various NoSql Databases.

- It enables device connectivity via industry standard IoT protocols - MQTT, CoAP, HTTP, Modbus TCP, OPC UA
- It supports both cloud and on-premises deployments.

It has features to

- Build Your own dashboard
- Analytics and Reporting
- Alert and Notification
- Integration with third party application(Power BI, SAP, ERP)
- Rule Engine





ii. Smart Factory Platform (FACTORY WATCH)

Factory watch is a platform for smart factory needs.

It provides Users/ Factory

- with a scalable solution for their Production and asset monitoring
- OEE and predictive maintenance solution scaling up to digital twin for your assets.
- to unleased the true potential of the data that their machines are generating and helps to identify the KPIs and also improve them.
- A modular architecture that allows users to choose the service that they what to start and then can scale to more complex solutions as per their demands.

Its unique SaaS model helps users to save time, cost and money.

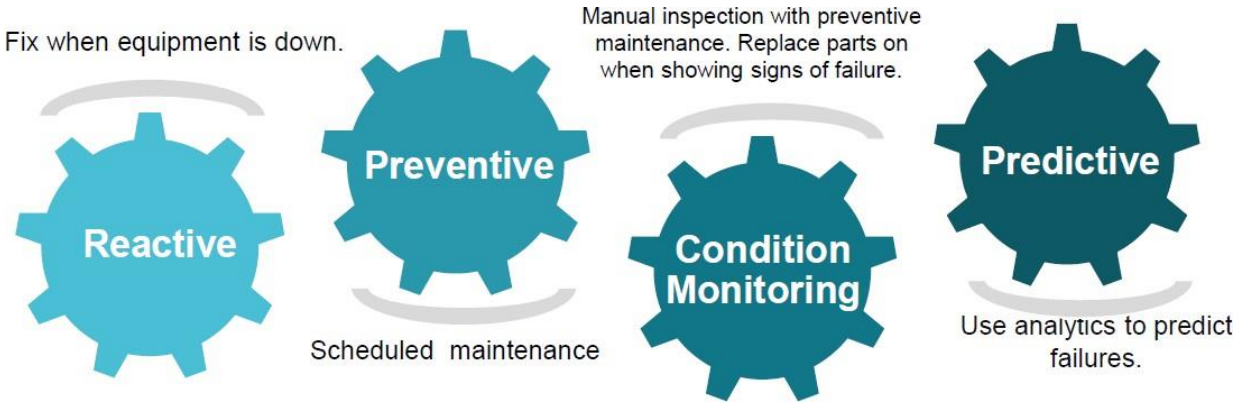
Machine	Operator	Work Order ID	Job ID	Job Performance	Job Progress		Output		Rejection	Time (mins)				Job Status	End Customer
					Start Time	End Time	Planned	Actual		Setup	Pred	Downtime	Idle		
CNC_S7_81	Operator 1	WO0405200001	4168	58%	10:30 AM		55	41	0	80	215	0	45	In Progress	i
CNC_S7_81	Operator 1	WO0405200001	4168	58%	10:30 AM		55	41	0	80	215	0	45	In Progress	i

iii. LoRaWAN based Solution

UCT is one of the early adopters of LoRAWAN teschnology and providing solution in Agritech, Smart cities, Industrial Monitoring, Smart Street Light, Smart Water/ Gas/ Electricity metering solutions etc.

iv. Predictive Maintenance

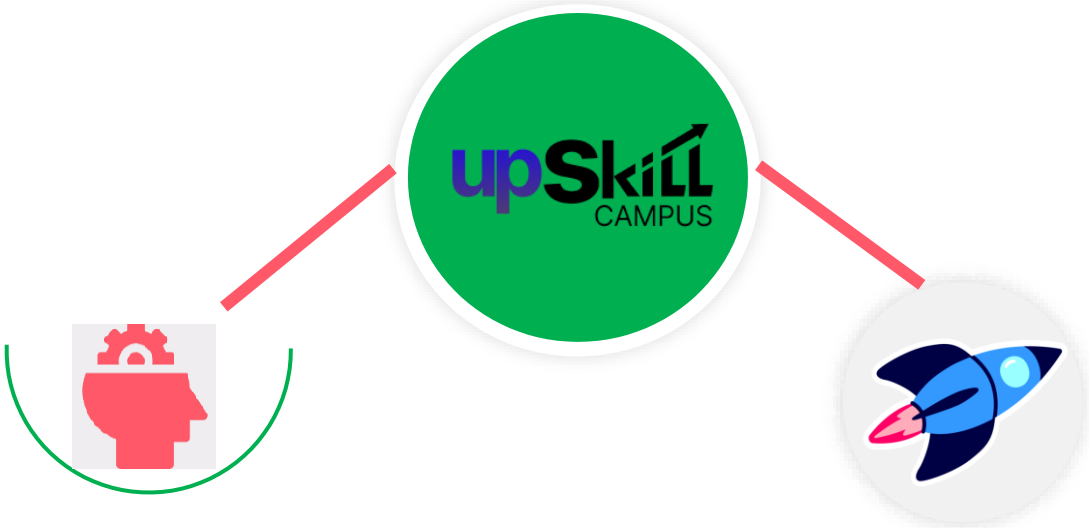
UCT is providing Industrial Machine health monitoring and Predictive maintenance solution leveraging Embedded system, Industrial IoT and Machine Learning Technologies by finding Remaining useful life time of various Machines used in production process.



2.2 About upskill Campus (USC)

upskill Campus along with The IoT Academy and in association with Uniconverge technologies has facilitated the smooth execution of the complete internship process.

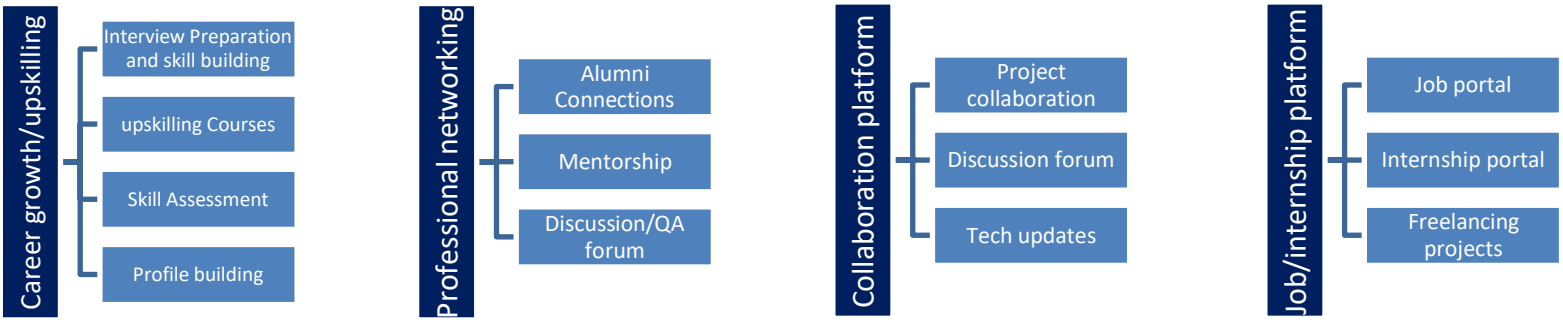
USC is a career development platform that delivers **personalized executive coaching** in a more affordable, scalable and measurable way.



Seeing need of upskilling in self paced manner along-with additional support services e.g. Internship, projects, interaction with Industry experts, Career growth Services

upSkill Campus aiming to upskill 1 million learners in next 5 year

https://www.upskillcampus.com/



2.3 The IoT Academy

The IoT academy is EdTech Division of UCT that is running long executive certification programs in collaboration with EICT Academy, IITK, IITR and IITG in multiple domains.

2.4 Objectives of this Internship program

The objective for this internship program was to

- ☛ get practical experience of working in the industry.
- ☛ to solve real world problems.
- ☛ to have improved job prospects.
- ☛ to have Improved understanding of our field and its applications.
- ☛ to have Personal growth like better communication and problem solving.

2.5 Reference

[1] https://www.theiotacademy.co/

2.6 Glossary

Terms	Acronym
ensemble	a number of things considered as a group a set of clothes that are worn together

3 Problem Statement

We are working with the government to turn different cities into a smart city. The vision is to turn it into a digital and smart city to improve the efficiency of services for citizens. One of the problems the government faces is traffic. You are a data scientist working to better manage city traffic and provide insights into infrastructure planning for the future.

The government wants to build a robust transport system for the city by preparing for traffic spikes. They wanted to understand the traffic patterns of the city's four intersections. Traffic patterns on holidays, as well as on other occasions of the year, differ from normal working days.

This is very important to consider for your forecast.

The traffic forecasting problem is very extensive and it includes some subproblems according to different issues. One of the most important features in the traffic forecasting problem is the kind of network where the predictions are performed. Usually, they are classified in urban networks and freeways. The topology is very different because urban networks contain more and shorter links while freeway networks are composed of few but larger links. Also, some authors include the arterial network category when the prediction is performed only over the main sections of an urban network. Typically, forecasting in urban areas is more difficult because the behavior of the drivers inside the network is less predictable.

Another key feature in this problem is the prediction horizon. The literature uses the expressions short-term and long-term to classify them. Although it is not clear exactly what horizons each group refers to, the short-term name is used for predictions from 1 minute to around 30 minutes or 1 hour depending on the author. For larger prediction horizons, the long-term name is used. Of course, despite this binary classification, the prediction horizon can variate between 1 minute to hours or days. In general, the complexity of the problem increases and the forecasting accuracy decreases with larger horizons.

Besides the kind of network and the horizon time, the traffic prediction problem is determined by what variable and with which granularity is predicted. Traffic forecasting can be performed for different variables, the most presents in the literature are four.

These predictors can be used at different scales, such as a particular point in the network, a portion (a link or a portion of a link) of the network, and the entire network (or an area). area of the network).

4 Existing and Proposed solution

The main objective of this master's thesis is to perform traffic prediction in an urban context by developing machine learning models trained with floating point automotive data (FCD). Specifically, the average speed is predicted for each road segment in the entire urban scenario. Therefore, regression machine learning methods should be used because the response variable is numerical. Furthermore, this work is not only concerned with short-term or long-term forecasting, so both methods are tested.

This project focuses on smart city scenarios where FCD is available. Currently, collecting real data to evaluate different levels of DCF penetration is a pipe dream. Therefore, these scenarios are simulated by traffic simulation. The ability to analyze the performance of machine learning methods in a smart city and the aspects that affect the accuracy of this type of solution can help with design and decision-making during city project development. smart. In addition, the assessment of the minimum requirements for a

Traffic forecasting models are very useful to study their real feasibility.

There is also another impetus for the research proposed in this master's thesis, which is research developed for the inLab FIB smart mobility research group. Traditionally, the team solves smart city problems by analyzing and predicting different traffic situations through simulation techniques. Simulation model development and calibration are expensive tasks that require access to real data, which is sometimes difficult to obtain. This is necessary to adapt the model to the actual scenario. On the other hand, due to the increasing number and different sources of materials, some problems

can be faced from a data-driven approach. This type of solution, like machine learning, does not require simulation model development. So with enough data, a data-driven approach can be cheaper than simulation in terms of time and money. This project will try to solve the problem of traffic forecasting using machine learning methods as an alternative to simulation methods.

4.1 Code submission (Github link):

https://github.com/Nsdevil/Forecasting-of-Smart-city-traffic-patterns

4.2 Report submission (Github link):

https://github.com/Nsdevil/Forecasting-of-Smart-city-traffic-patterns

5 Proposed Design/ Model

In order to obtain accurate and up-to-date information about the traffic conditions within a smart city, a comprehensive data collection process is necessary. This involves gathering real-time traffic data from multiple sources, including traffic sensors, surveillance cameras, GPS-enabled devices, and other Internet of Things (IoT) devices that are strategically deployed throughout the city's infrastructure. The collected data encompasses a wide range of information, such as traffic flow patterns, congestion levels, vehicle counts, and even historical data. By utilizing these diverse sources, city planners and transportation authorities can gain valuable insights into the current state of traffic within the city, enabling them to make informed decisions and implement effective strategies to improve traffic management and alleviate congestion.

In the process of data preprocessing, it is essential to cleanse and preprocess the data that has been collected. This involves several steps such as removing outliers, handling missing values, and normalizing the data. The main objective is to ensure consistency and accuracy in the data. Additionally, there is a possibility of aggregating the data at different time intervals, such as hourly or daily, in order to capture any patterns or trends that may be present.

Feature extraction involves identifying important characteristics from the preprocessed data that can be utilized to predict traffic patterns. These characteristics encompass various factors such as the time of day, the day of the week, the prevailing weather conditions, any special events taking place, the quality of road infrastructure, and past traffic patterns.

In the process of model selection, it is important to carefully consider the characteristics of the data and the desired objectives in order to choose the most suitable forecasting model. In the context of traffic forecasting, there are several commonly utilized models that can be employed, such as time series models like ARIMA and SARIMA, machine learning algorithms

like random forest and gradient boosting, or deep learning models like recurrent neural networks and convolutional neural networks.

To train the forecasting model, the preprocessed data is divided into two sets: training and validation. The training set is utilized to optimize the model parameters and fine-tune its architecture. The model's performance is then evaluated using the validation set, with metrics like mean absolute error (MAE) or root mean squared error (RMSE) being used as benchmarks to assess its accuracy.

In the process of model evaluation, it is important to assess both the accuracy and reliability of the forecasting model. This can be done by comparing its predictions with the actual traffic patterns observed. By doing so, we can effectively analyze the performance of the model in various scenarios and determine its ability to capture both short-term and long-term traffic trends.

The process of model deployment involves integrating the trained forecasting model into a smart city traffic management system. This integration is crucial as it allows for the utilization of real-time data ingestion and processing, ensuring that the model is continuously updated with the most recent traffic data. By implementing this system, stakeholders are provided with user-friendly interfaces or APIs that enable them to access the forecasted traffic patterns. This access enables stakeholders to make informed decisions based on the forecasted data, ultimately contributing to the efficient management of traffic in the smart city.

The process of maintaining and improving the forecasting model involves regularly monitoring its performance and updating it with new data as needed. Additionally, the model is continuously evaluated and enhanced through the inclusion of new features, experimentation with different algorithms, or the consideration of ensemble methods, all with the aim of improving the accuracy of the forecasts.

6 Performance Test

Urban networks of varying sizes were used in computational research; see [4] for more information. For the sake of completeness, we list them below for the medium-sized network from the Spanish city of Vitoria (Figure 3), which includes 57 centroids, 3249 OD pairs, 2800 intersections, and a modeled network of roughly 600 km. This network has representative congestion levels and a route choice dimension that are common in many large urban regions, and it has a respectable size. In the Vitoria network, there are two distinct sets of sensors:

- The left-hand side of the diagram shows 389 standard loop detectors that provide flows, speeds, and occupancy relating to all identified vehicles at the loop.
- Notably, the deployment of the 50 AVI sensors follows a layout strategy, the details of which can be found in [9], in order to maximize the capture of Bluetooth-equipped vehicles and provide accurate trip time data between AVI sensors.

6.1 Test Plan/ Test Cases

6.1.1 Valid Input

In order to evaluate the application's performance, it is necessary to test it using valid input data. This input data includes historical traffic data, weather conditions, and events schedule. By inputting this information into the application, we expect to receive an accurate forecast of traffic patterns based on the input data. This forecast will help us understand how well the application can analyze and predict traffic based on various factors such as historical data, weather conditions, and events schedule.

6.1.2 Large Data Set

The objective is to evaluate the application's performance by subjecting it to a significant amount of input data. This includes a vast collection of historical traffic data, a wide range of weather conditions, and a comprehensive schedule of events. The anticipated result is that the application will be able to process this extensive input data efficiently and generate accurate forecasts without any performance-related problems.

6.2 Test Procedure

The purpose of this test is to ensure that the traffic pattern forecasting system in a smart city environment is both accurate and dependable. In order to conduct the test, several steps need to be taken to set up the environment. Firstly, it is necessary to create a test environment that accurately mimics the real-time traffic data found in a smart city. This can involve setting up sensors or using simulations to generate the data. Once the test environment is ready, the next step is to install and properly configure the traffic pattern forecasting system. This system will be responsible for analyzing the traffic data and making predictions about future traffic patterns. It is crucial to ensure that the system is correctly installed and configured to guarantee accurate and reliable forecasts. Lastly, to validate the accuracy of the forecasting system, historical traffic data needs to be available for reference. This historical data will serve as a benchmark against which the system's predictions can be compared. By having access to this data, any discrepancies or errors in the forecasting system can be identified and addressed. Therefore, it is essential to have a sufficient amount of historical traffic data readily available for the testing process. Test steps: a. Activate the traffic pattern prediction system. b. Configure the system with relevant parameters, such as time intervals, predictive models, and data sources. c. Provides the system with historical traffic data for training and validation purposes. d. Implement the system to generate traffic forecasts for a specific time period. e. Comparing the generated forecasts to actual traffic patterns observed over the same time period. F. Assess the accuracy of the predictions by analyzing the differences between the predicted and observed traffic patterns. g. Repeat the process using different historical data sets and time periods to assess the consistency and reliability of the system. H. Monitor system performance to ensure it meets predefined performance criteria (eg response time, resource utilization). The test outputs include the identification of discrepancies between the predicted traffic patterns and the actual observed patterns. Additionally, the analysis of forecast accuracy involves evaluating various metrics like mean absolute error, root mean square error, and percentage error. Furthermore, the performance measurements of the system encompass assessing factors such as response time, resource utilization, and scalability. Test pass criteria: a. Traffic pattern prediction accuracy must be equal to or greater than a predefined threshold. b. The system must perform within acceptable response time and resource utilization limits. c. The system should demonstrate consistency and reliability over multiple test iterations. d. Identified issues and

deficiencies should be reported and appropriately addressed. In the process of test reporting, it is crucial to thoroughly document all the test results. This includes not only recording the observed variances, accuracy metrics, and system performance, but also providing a comprehensive summary of the test findings. This summary should highlight any identified issues or areas for improvement that were discovered during the testing process. Once the test report is complete, it should be submitted to relevant stakeholders, such as system developers, project managers, and quality assurance teams. These stakeholders are responsible for reviewing the test report and taking appropriate actions based on the information provided. The test report serves as a valuable tool in ensuring the overall quality and effectiveness of the system being tested.

6.3 Performance Outcome

The accuracy of forecasting traffic patterns is an extremely important measure of performance. It can be evaluated by comparing the predicted traffic patterns to the patterns that are actually observed. A higher level of accuracy signifies a superior performance outcome.

Reliability pertains to the constancy and trustworthiness of the predictions made by a traffic pattern forecasting model. An ideal forecasting model should consistently deliver precise forecasts over a prolonged period and across various situations.

The concept of timeliness in traffic pattern forecasting refers to the ability to provide predictions promptly. In a smart city setting, the timely delivery of these predictions is of utmost importance as it plays a vital role in effectively managing and planning for traffic.

Scalability refers to the capacity of a forecasting system to effectively manage and process larger amounts of data and meet the growing demands of traffic in a smart city. A system that is highly scalable has the capability to adapt and handle the expansion and intricacies of traffic patterns within a smart city.

The capacity of the forecasting system to effectively respond to evolving conditions and the ever-changing dynamics of traffic patterns holds significant importance. A system that is adaptable possesses the capability to flexibly modify its forecasting models and algorithms in order to accurately incorporate numerous factors that have the potential to influence traffic patterns.

Efficiency refers to the extent of computational resources and time needed by the forecasting system to generate predictions. A forecasting system that operates efficiently is capable of producing precise forecasts while consuming fewer resources.

Integration is a crucial aspect when it comes to the performance outcomes of the forecasting system in a smart city. It is essential for the system to effectively blend with other existing traffic management systems, various data sources, and even communication networks without any disruptions or difficulties.

7 My learnings

The choice of machine learning technique employed depends on the specific problem at hand. For instance, regression models are commonly used for short-term traffic flow prediction, while time series models are better suited for long-term prediction. Additionally, the dynamic nature of traffic patterns poses another challenge. Traffic conditions can change rapidly, necessitating the use of models that can adapt to these changes in real-time. The field of research focusing on the use of machine learning techniques to forecast traffic patterns shows great promise. By training machine learning models on historical data, it becomes possible to understand the intricate relationships between traffic flow and various factors, such as the time of day, day of the week, and weather conditions. Armed with this knowledge, accurate predictions about traffic flow can be made. To enhance the accuracy of traffic flow prediction, a diverse range of data sources can be utilized. Aside from historical traffic data, weather data, traffic camera data, and GPS data can all contribute to improving the accuracy of predictions. While the use of machine learning for traffic flow prediction is still in its early stages, it holds immense potential. As technology continues to advance, we can expect even more precise and reliable predictions, which will undoubtedly contribute to the increased efficiency of transportation systems. However, there are challenges that need to be overcome in order to achieve accurate traffic flow prediction. One such challenge is the requirement for vast amounts of data. Collecting this data can be both difficult and expensive, but it is crucial for training models that yield accurate predictions. The potential benefits of traffic flow prediction extend beyond mere accuracy. By providing information about expected traffic conditions, drivers and public transportation planners can make better decisions about travel routes and schedules. Ultimately, this can lead to reduced congestion, improved travel times, and even fuel savings.

8 Conclusions and Future work scope

Based on the superior performance of LSTM, the algorithm is utilized to forecast future traffic data for the upcoming two weeks, specifically from January 1, 2020, to January 14, 2020. This prediction will assist in anticipating traffic conditions and planning accordingly. On the other hand, LSTM algorithm demonstrates promising results in traffic flow prediction. It yields an RMSE value of 0.19, which is significantly lower than that of SVR. Therefore, it can be concluded that LSTM outperforms SVR in terms of accurately predicting traffic flow. Traffic monitoring is a critical task that requires careful attention. This study focuses on analyzing a Traffic dataset specifically collected from five different areas within the city of Bloomington in the United States. The dataset encompasses traffic data counts ranging from January 1, 2017, to December 31, 2019. The objective of this research is to predict traffic flow using two different algorithms: Support Vector Regression (SVR) and Long Short-Term Memory (LSTM). To evaluate the performance of these algorithms, various metrics are employed. The primary metric considered is the Root Mean Squared Error (RMSE), which measures the accuracy of the predicted traffic flow values. For SVR, the calculated RMSE value is found to be 0.473. Although it is expected for the RMSE value to be as low as possible, this result indicates that SVR is not suitable for accurate traffic flow forecasting in this context.

Traffic flow prediction is a critical task for smart cities, as it allows drivers to effectively plan their trips. In order to accurately forecast traffic flow, this study initially combined pollution and traffic datasets from Aarhus, Germany. Various conventional machine learning approaches were then applied to the dataset to determine the most accurate method. Among these approaches, K-nearest neighbors (KNN) demonstrated the lowest mean absolute error (MAE) and root mean square error (RMSE) values. Building upon the results of the conventional approaches, a bagging and stacking ensemble technique was employed to further improve the MAE and RMSE values. The dataset was split into samples using bootstrapping with replacement, and these samples were utilized by a diverse set of homogeneous models. The results from these models were aggregated to create a robust bagging ensemble model. The KNN bagging ensemble model emerged as the most accurate among all the bagging and stacking ensemble combinations. One reason for the superior performance of KNN is its ability to effectively handle non-linear data, which is characteristic of the dataset in question. However, it is important to note that KNN can underfit the data if the number of nearest neighbors (K) is too small, and overfit if it is too large. The experimental results indicate that the proposed bagging ensemble scheme reduced the error rate

by 30% compared to previous studies that utilized boosting for traffic flow prediction in smart cities. This improvement is attributed to the presence of outliers in the dataset, which caused the boosting ensemble models to overfit. The experiments further suggest that the proposed bagging ensemble scheme mitigated the impact of overfitting and resulted in a more accurate error rate. In fact, it decreased the error by 12% compared to the KNN and stacking ensemble models employed in the study. However, there are some concerns associated with bagging. Firstly, it requires additional storage space and computation time. Secondly, due to the utilization of a large number of base classifiers, bagging can potentially compromise the interpretability of the model and introduce bias if the standard procedure of applying bagging is not followed. This can lead to underfitting of the data. Moreover, since the dataset used in the study pertains specifically to different areas of Aarhus, Germany, the model's performance may be slightly compromised in other regions of the world where seasonal and traffic patterns differ.

In the future, there will be a comprehensive examination of the impacts that different seasons, including summer and winter, have on various aspects of society. Specifically, there will be a focus on investigating how traffic patterns change as individuals in European countries embark on vacations and depart from urban areas during the month of August. Furthermore, the forthcoming study will aim to identify the extent to which road traffic, household activities, and various types of equipment contribute to air pollution. It is important to note that the correlation between traffic and pollution data may differ among different cities. To gain a more comprehensive understanding of traffic flow, it may be beneficial to combine satellite measurements with ground sensor values to determine if they can be used in tandem to predict traffic patterns. Additionally, to ensure the accuracy and reliability of the findings, data from additional cities around the world will be incorporated into the research, thereby assessing the robustness and other performance parameters of the study.