Chapter 8

Traffic Forecasting and Route Optimization Using Classified Vehicle Data and GPS Augmentation

8.1 Preface

This chapter presents a comprehensive evaluation of a machine learning-based framework for urban traffic forecasting and route optimization. Building upon the classification of vehicle types via vibration signals (Chapter 7) and leveraging the data acquisition protocols and GPS overlay mechanism described in Section 4.8, the study employs time-series forecasting models, namely Autoregressive Integrated Moving Average (ARIMA) and Support Vector Machine (SVM), to predict short-term traffic volumes. The forecasts are based on classified vehicle count data collected across 24 representative locations in two Indian cities.

To translate traffic forecasts into actionable routing decisions, the framework further integrates XGBoost classifiers for evaluating route efficiency based on predicted traffic densities. A series of controlled experiments examines the influence of model selection, spatial

sampling density, preprocessing strategies, classification accuracy and training–testing data partitions on forecasting performance. In addition, a real-world case study involving 40 volunteers compares the proposed GPS-augmented navigation approach against conventional GPS routing, validating its effectiveness in reducing travel time and enhancing route selection.

The methodological developments and experimental findings underscore the feasibility of integrating machine learning with vehicle classification datasets to improve spatio-temporal traffic prediction and vibration-aware navigation in urban environments.

8.2 Experimental evaluation and results

This section describes the experimental evaluations and results of the proposed technique to verify its effectiveness on the collected dataset using ARIMA and SVM-based forecasting models. First, the section discusses implementation details and performance metrics. Further, the section presents the evaluation results for different settings and parameters using the XGBoost model and the dataset collected from pre-specified sites of Indian Roads (Gorakhpur and Varanasi). Finally, a real-world study was conducted to verify the effectiveness of the proposed approach using data collected from 20 volunteers.

8.2.1 Dataset

This work uses the dataset collected from the selected sites in two cities in India. The details of the dataset collection are covered in Section 4.8, where the dataset comprises the instances of the traffic volume from 24 different sites in two major Indian cities (Gorakhpur and Varanasi).

8.2.2 Implementation details

This work uses Python language for implementing the forecasting models elaborated in Section 4.8 and training the XGBoost model. Specifically, the Keras library in Tensorflow was used to implement the system. All the experiments were performed on the Dell PC having *i*7 processor and 24 GB RAM. In the experiments, the collected dataset is divided into training and testing datasets with 70% and 30% instances, respectively. Mean Square Error (MSE) was used to measure the forecasting accuracy. These results demonstrate that the proposed approach achieves state-of-the-art results on the collected dataset. During the experiment, the batch size is set to 32 and the optimizer to SGD. The training epoch was set to 50 and utilized 70% of the dataset instances. All the presented results are repeated for 50 times to depict the average results.

In the proposed work, a 15-minute interval was chosen for recording traffic volume videos based on the following considerations. A 15-minute interval provides a sufficient sample size to capture variations in traffic volume throughout different times of the day. It allows for a more accurate representation of traffic patterns and helps in obtaining statistically significant data for analysis. It offers a balance between capturing short-term fluctuations and providing a broad overview of traffic trends over time. This resolution is often suitable for understanding daily traffic patterns, peak hours and potential congestion points. Many traffic management decisions, such as adjusting signal timings or deploying resources, are made on a relatively short timescale. Next, traffic control systems often operate on shorter time intervals and a 15-minute data collection interval aligns with the frequency at which adjustments to traffic signal timings or other control strategies are made. Furthermore, the reviewer's perspective on the ARIMA model is acknowledged, emphasizing the importance of using a short interval duration to optimize forecasting accuracy. However, this demand for a short interval can be addressed by the significantly higher number of data instances, a characteristic applicable to our specific scenario as

well. Consequently, utilizing a 15-minute interval proves effective for us, yielding superior results. Finally, the data collection spanned an extended period, justifying the selection of a 15-minute sampling rate based on the recommendations of experts from both our institute and our collaborating partner.

8.2.3 Experimental results

This section presents various experimental results on collected datasets to study the impact of different parameters and forecasting techniques on the proposed work.

8.2.3.1 Impact of forecasting model

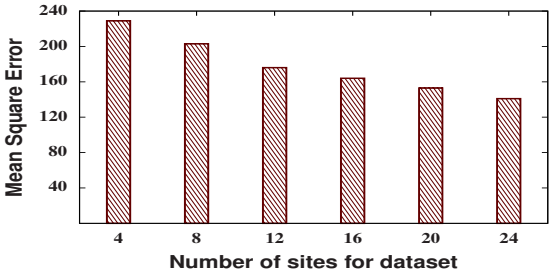
This experiment studies the impact of different forecasting models used in this work, *i.e.,* ARIMA and SVM. The results in Table 8.1 illustrate that different forecasting models have varying impacts on the accuracy and reliability of forecasts. The choice of a forecasting model largely depends on the nature of the data and the purpose of the forecast. Some of the commonly used forecasting models include ARIMA, SVM, exponential smoothing, neural networks and regression analysis. ARIMA and SVM models are widely used for time-series forecasting and are particularly useful when the data has a strong trend or seasonality. Exponential smoothing models, on the other hand, are suitable for data with little or no trend or seasonality and are commonly used in short-term forecasting. Neural network models, which are inspired by the structure and function of the human brain, are known for their ability to handle large datasets and complex patterns. They are particularly useful in situations where traditional statistical models may not be effective. Regression analysis models are used to predict the relationship between two or more variables and are often used in business forecasting. These models can help identify the key factors that drive performance and predict future trends.

Table 8.1 Illustration of the impact of different forecasting models on Mean Square Error (MSE). 2-w = Two Wheeler, 3-w = Three Wheeler and LCV = Light Commercial Vehicle.

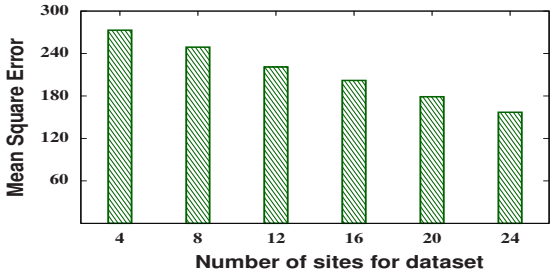
Vehicle	*2-w*	*3-w*	*Car*	*Jeep*	*LCV*	*Tractor*	*Bus*	*Truck*	*eRickshaw*	*Cycle*
ARIMA	141	157	**93**	105	119	124	135	123	129	132
SVM	153	165	**104**	113	127	133	142	132	144	152

The impact of different forecasting models also depends on the data quality used to develop the model. High-quality data is essential for accurate and reliable forecasts and the choice of model can only improve accuracy up to a certain point. Inaccurate or incomplete data can limit the effectiveness of even the most sophisticated forecasting models. By understanding the strengths and limitations of different models and using high-quality data, organizations can improve the accuracy and reliability of their forecasts and make more informed decisions; thus, in this work, ARIMA and SVM models were used.

The impact of various forecasting models on Mean Square Error (MSE) using ARIMA and SVM models is presented in Table 8.1. The results reveal that the vehicle category has the lowest MSE, while the "other vehicle" category has the highest MSE. This can be attributed to the fact that the collected dataset contains a significant number of "car" instances, which results in the predictor learning more identifiable features for "cars" compared to other vehicles. Additionally, the ARIMA models outperform SVM due to their ability to detect low-level features from the input data by performing auto-regressive operations.



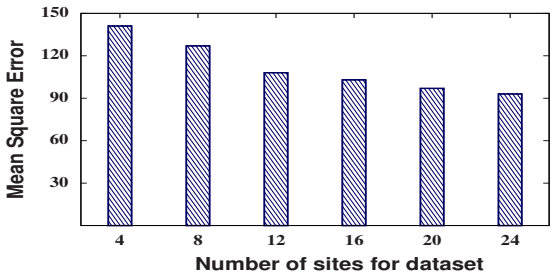
(a) Two Wheeler



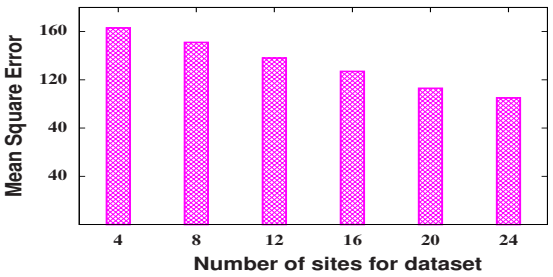
(b) Three Wheeler

Continued on next page

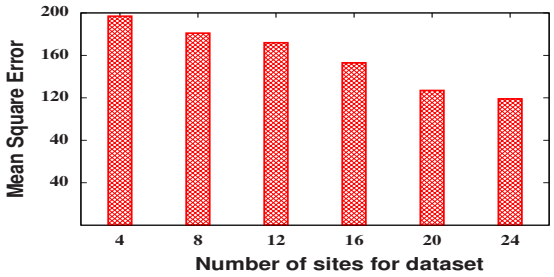
Figure 8.1 continued from previous page



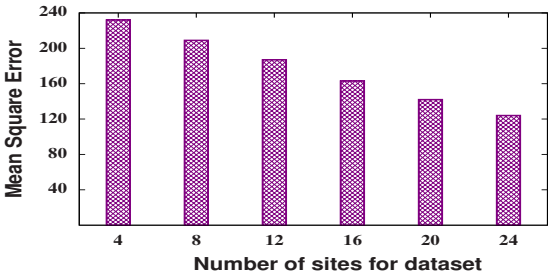
(c) Car



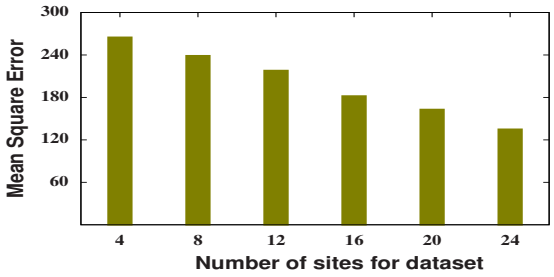
(d) Jeep



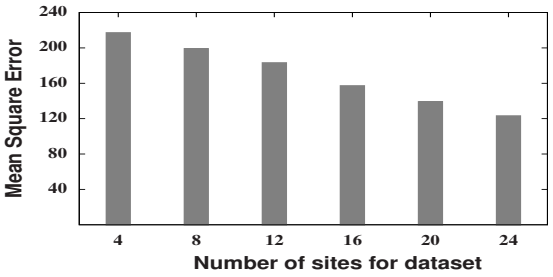
(e) Light Commercial Vehicle



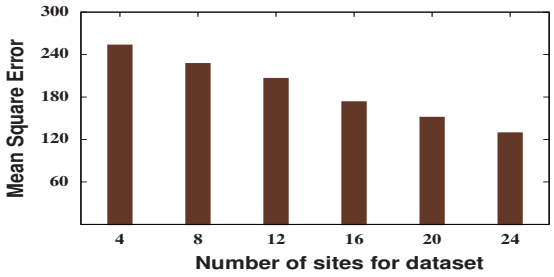
(f) Tractor



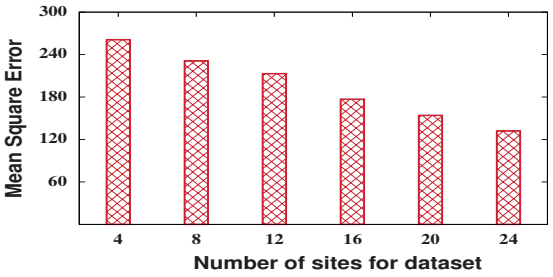
(g) Bus



(h) Truck



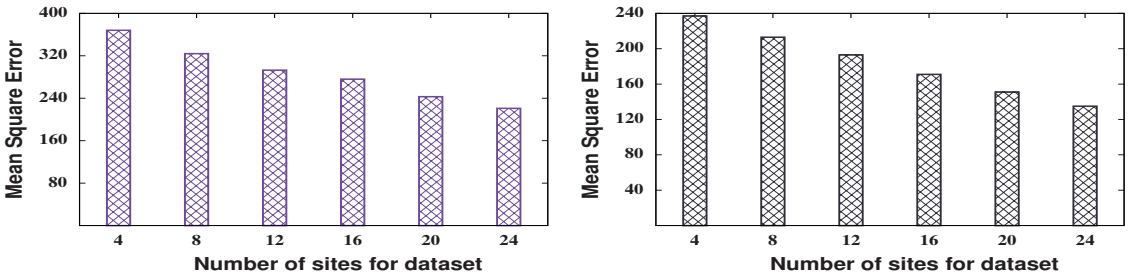
(i) e-Rickshaw



(j) Cycle

Continued on next page

Figure 8.1 continued from previous page



(k) Others (l) Average

Fig. 8.1 Impact of number of sites for data collection (assuming 4, 8, 12, 16, 20 and 24) on MSE using ARIMA and SVM-based forecasting models.

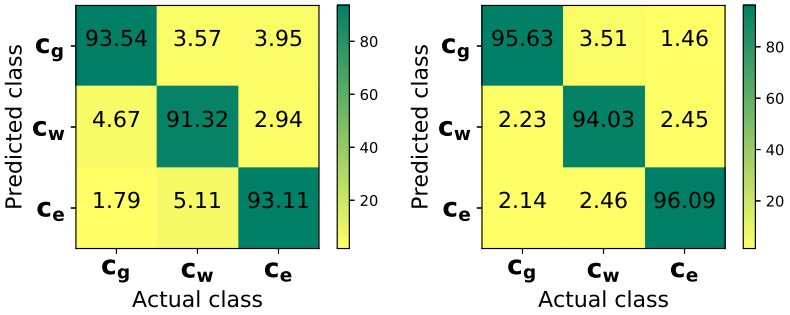
8.2.3.2 Impact of number of sites

The impact of the number of sites (24 in this work) for the data collection on performance depends on the specific application and the nature of the data being collected. In general, collecting data from multiple sites can increase the amount of data available for analysis and improve the accuracy of the resulting models or predictions. This is particularly true for complex systems or processes exhibiting significant variability across different locations. However, collecting data from multiple sites can also increase the complexity and cost of data collection and the potential for data quality issues. In addition, the number and location of sites may need to be carefully chosen to ensure that the data collected is representative of the overall system or process being studied. When analyzing data collected from multiple sites, it is essential to consider the potential sources of variability between sites and to develop appropriate statistical methods to account for the differences. This may include techniques such as hierarchical modelling or mixed-effects models, which allow for variation between sites while accounting for the overall trends and patterns in the data.

In this experiment, ARIMA and Support Vector Machine (SVM) models were used for forecasting traffic volume. Fig. 8.1 illustrates the MSE of forecasting averaged over ARIMA and SVM. Overall, the impact of the number of sites for the data collection on performance is complex and depends on various factors. While collecting data from multiple sites can improve the accuracy and reliability of the resulting models or predictions, it is important to carefully consider the potential costs and challenges of data collection and to use appropriate statistical methods to account for variation between sites. In Fig. 8.1, the impact of the number of sites on the individual vehicle MSE was presented. To study site variation, this study considered a set of 4, 8, 12, 16, 20 and 24 sites and also depicted the average MSE in Fig. 8.1. The results show that increasing the number of sites for dataset collection leads to a rapid decrease in MSE. This is due to the increased number of data instances available for training the predictor, reducing the MSE. As previously observed, the MSE for any number of sites is minimal for the "car" class label due to the larger number of available instances. However, the average MSE is higher for all combinations of the sites set (4, 8, 12, 16, 20 and 24) due to the scarcity of instances. Notably, the number of sites, more precisely the number of data instances, plays a crucial role in minimizing error or maximizing the predictor's performance.

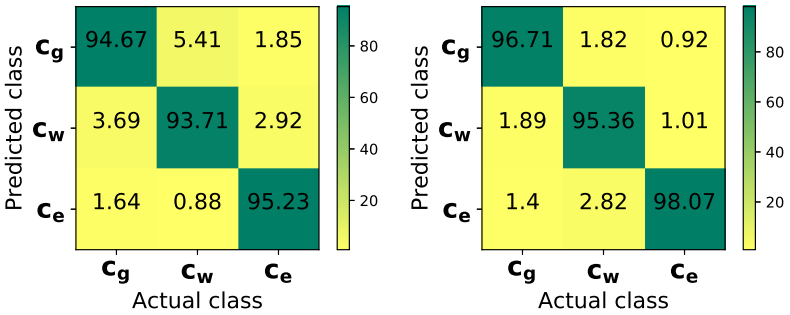
8.2.3.3 Impact of pre-processing

To study the impact of preprocessing, this study considered the accuracy of XGBoost + SVM and XGBoost + ARIMA with and without preprocessing in predicting the mode of transportation or vehicles. The considered preprocessing involves transforming and cleaning the data before training the prediction model. The impact of preprocessing on performance is significant because it can affect the model's accuracy, efficiency and generalization capabilities. An important preprocessing step is data normalization, which involves scaling the data so that it falls within a specific range or distribution, which is



(a1) Without preprocessing. (a2) With preprocessing.

(a) SVM+ XGBoost.



(b1) Without preprocessing. (b2) With preprocessing.

(b) ARIMA + XGBoost.

Fig. 8.2 Impact of preprocessing on the performance of XGBoost-based classifier using SVM and ARIMA model on a collected dataset.

adopted in this work. This improved the model's efficiency by reducing the impact of outliers and ensuring that all features are treated equally.

The impact of preprocessing on performance depends on the considered applications and the nature of the collected dataset. However, in general, effective preprocessing can significantly improve the accuracy, efficiency and generalization capabilities of machine learning models. It can also help prevent overfitting, which occurs when the model is too closely tailored to the training data and performs poorly on new or unseen data.

The experiment discusses the impact of preprocessing on the performance of machine learning models, specifically SVM and ARIMA for forecasting and XGBoost for prediction. Fig. 8.2 illustrates the impact of preprocessing on the performance of individual classes, namely c_g, c_w and c_e. The results demonstrate that preprocessing can signifi-

cantly improve the performance of both SVM and ARIMA-based forecasting models and the XGBoost model for prediction. The SVM-based forecasting model and XGBoost combination underperform compared to the ARIMA and XGBoost combination due to ARIMA's higher capability to achieve high-order performance in forecasting. Therefore, the experiment implies that preprocessing the dataset is necessary to achieve high-order performance in machine learning models. Moreover, the results indicate that the performance achieved by the class c_w is the lowest, while the performance for c_e is the highest due to the availability of correspondingly least and highest data instances. Interestingly, the experiment emphasizes the importance of preprocessing the dataset for achieving high performance in machine learning models, especially for forecasting and prediction tasks. Additionally, it highlights the impact of the number of data instances on the performance of the models. Furthermore, the marginal improvement in ARIMA model is due to its ability to capture and model linear and non-linear temporal patterns in the time series data more efficiently than SVM model. Moreover, SVM is a powerful machine learning algorithm that works well with high-dimensional data and complex relationships. However, our forecasting task involves a relatively simpler structure; specifically, the data has a more straightforward temporal pattern and the simplicity of ARIMA might contribute to its slightly better performance.

8.2.3.4 Impact of classification models

The primary objective of the study is to compare the accuracy and effectiveness of each classification model in predicting or classifying the given data. By comparing the performance of different models, the experiment identifies the strengths and limitations of each model and determines the most effective one. This experiment considered different classification models; applied over ARIMA and SVM based-forecasting models. Table 8.3 presents an illustration of the impact of different classification models (Gaussian Naive

Table 8.3 An illustration of the impact of different classification models (Gaussian Naive Bayes, k-Nearest Neighbor, Logistic regression, SVM and XGBoost) using forecasted data from the ARIMA and SVM using the collected dataset.

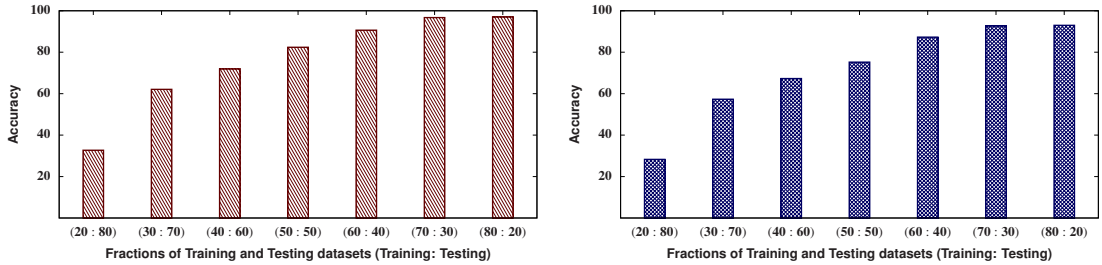
Forecasting model	**Classification model**	**Accuracy**			
		Class (c_g)	Class (c_w)	Class (c_e)	**Average**
ARIMA	**Gaussian Naive Bayes**	94.83%	92.41%	95.21%	94.15%
	k-Nearest Neighbor	95.07%	92.76%	95.53%	94.46%
	Logistic Regression	95.21%	93.81%	95.87%	94.97%
	SVM	95.81%	94.19%	96.84%	95.62%
	XGBoost	96.71%	95.36%	98.07%	96.71%
SVM	**Gaussian Naive Bayes**	91.89%	88.91%	90.43%	90.42%
	k-Nearest Neighbor	92.17%	89.43%	91.89%	91.17%
	Logistic Regression	92.72%	90.21%	92.43%	91.78%
	SVM	93.07%	90.82%	92.76%	92.22%
	XGBoost	93.54%	91.32%	93.11%	92.65%

Bayes, k-Nearest Neighbor, Logistic regression, SVM and XGBoost) using forecasted data from the ARIMA and SVM using the collected dataset. It highlights a research experiment aimed at evaluating the effectiveness of different classification models on a given dataset. The five models chosen for the study are Gaussian Naive Bayes, k-Nearest Neighbor, Logistic Regression, SVM and XGBoost. These models are widely used in machine learning and have proven to be effective in the considered application.

Table 8.3 illustrates the performance of different classification models using the forecasted dataset from both SVM and ARIMA models. Similar to previous experiments, this experiment also employs the data from 24 different sites. The ARIMA and SVM-based forecasting model is a time-series forecasting method that predicts future values based on past data patterns. At the same time, the collected dataset contains historical data that can be used to train and test the models. The goal was to find the best combination of forecasting and classification models for the given dataset. The experiment results showed that combining ARIMA-based forecasting with the XGBoost classification model performed better than any other combination of SVM-based forecasting and different classification models. This conclusion was based on the performance metrics used to evaluate the models.

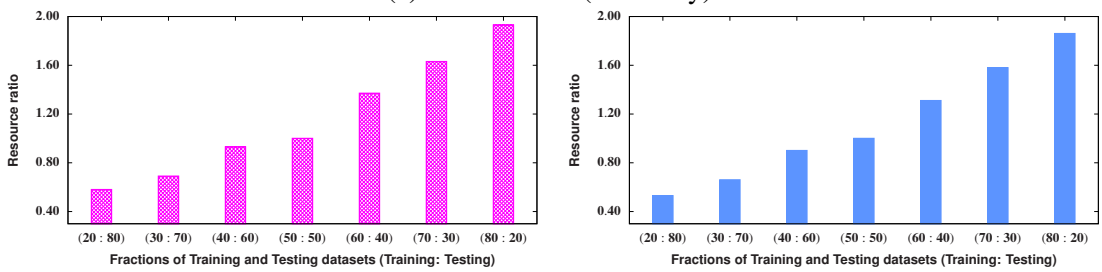
Thus, the combination of ARIMA-based forecasting and the XGBoost classification model can be considered the most suitable for this particular dataset.

This experiment provides valuable insights into the suitability of different classification models for different data types and helps guide the selection of the most appropriate model for a given application. By understanding the strengths and weaknesses of each model, one can make informed decisions when selecting a model for a specific task.



(a1) ARIMA+XGBoost. (a2) SVM +XGBoost.

(a) Performance (Accuracy).



(b1) ARIMA+XGBoost. (b2) SVM +XGBoost.

(b) Resource ratio with respected to $(50 : 50)$.

Fig. 8.3 Impact of training and testing datasets on the performance of XGBoost-based classifier.

Table 8.4 An illustration of the details of the volunteers used in this work for real-world evaluation.

Volunteer No.	Morning Peak	Off Peak	Evening Peak	Auto	Car	Motorcycle
G1	✓	✓	✓	–	✓	✓
G2	✓	✓	✓	✓	–	✓
G3	✓	–	✓	✓	–	–
G4	–	✓	✓	–	✓	✓
G5	✓	–	✓	–	✓	–
G6	–	✓	–	–	✓	–

Continued on next page

Table 8.4 – continued from previous page

Volunteer No.	Morning Peak	Off Peak	Evening Peak	Auto	Car	Motorcycle
G7	✓	✓	✓	✓	–	✓
G8	✓	–	✓	–	–	✓
G9	✓	✓	–	✓	–	–
G10	✓	✓	✓	–	✓	–
G11	✓	–	✓	✓	–	✓
G12	✓	✓	–	✓	✓	–
G13	–	✓	✓	–	✓	–
G14	✓	✓	✓	–	✓	–
G15	✓	✓	–	✓	–	–
G16	✓	–	✓	–	✓	–
G17	✓	✓	✓	✓	–	✓
G18	–	✓	✓	✓	–	–
G19	✓	✓	–	–	✓	–
G20	✓	–	✓	✓	–	✓
M1	✓	✓	✓	–	✓	–
M2	✓	–	✓	✓	–	✓
M3	✓	✓	–	✓	–	✓
M4	–	✓	✓	–	✓	–
M5	✓	✓	–	✓	–	✓
M6	✓	✓	✓	–	✓	–
M7	✓	–	✓	✓	–	–
M8	✓	✓	–	–	–	✓
M9	✓	–	✓	✓	✓	–
M10	✓	–	–	–	–	✓
M11	–	✓	✓	✓	✓	–
M12	✓	✓	–	–	✓	–
M13	✓	✓	–	✓	✓	–
M14	–	✓	✓	–	✓	✓
M15	✓	✓	–	–	✓	–
M16	✓	✓	✓	✓	–	✓
M17	✓	✓	✓	–	–	✓
M18	✓	✓	–	✓	✓	–
M19	✓	–	✓	–	✓	–
M20	✓	✓	✓	✓	–	✓

8.2.3.5 Impact of training and testing datasets

This experiment studies the impact of choosing different fractions of training and testing sub-datasets on the performance of the XGBoost model operating over the forecasted values

from the SVM and ARIMA models using the collected datasets from 24 different sites. The fraction of data used for training and testing can significantly impact the performance of a machine-learning model. Using a small training set and an extensive testing set may end up with an underfitting model. The model may not have learned enough about the data during training and thus may be unable to make accurate predictions on new data. On the other hand, the large testing set may be representative of the overall population, which can give you a reasonable estimate of the model's performance on unseen data.

Contrarily, for the extensive training set and small testing set, the model may overfit the training data. This means the model has learned the training data too well and may not generalize well to new, unseen data. This can result in poor performance on the testing set. However, a small testing set may not be representative of the overall population, which can lead to inaccurate estimates of the model's performance. Furthermore, when there are equal fractions of training and testing sets, the model may be able to learn enough from the training data to make accurate predictions. However, the size of the training and testing sets will determine the overall accuracy. In general, larger training sets can lead to better model performance but may require more computational resources and time.

This experiment investigates the impact of different fractions of training and testing sub-datasets $(training : testing)$ on performance. The fractions examined include $(20 : 80)$, $(30 : 70)$, $(40 : 60)$, $(50 : 50)$, $(60 : 40)$, $(70 : 30)$, $(80 : 20)$. It was found that the performance is significantly affected by the limited size of the training sub-dataset and improves as the fraction of the training dataset increases, as shown in Fig. 8.3. However, this also leads to increased computational requirements during training. The results indicate that the optimal fraction for performance and computational requirements is $70 : 30$. To evaluate the computational requirements, the experiment used the Floating Point Operations (Flops) ratio at $(50 : 50)$ with all other combinations. For example, the **resource ratio** for $(30 : 70)$ is defined as: Flops consumed in $(30 : 70)$/ Flops consumed in $(50 : 50)$.

An interesting observation from the results is as follows: the fraction of training and testing data can have a significant impact on the performance of a machine learning model. It is essential to choose a fraction that balances the trade-off between underfitting and overfitting and that is representative of the overall population.

8.2.4 Real-world evaluation

The section you provided discusses the real-world evaluation of a proposed approach, which involves testing a GPS overlay scheme against traditional GPS navigation. To conduct the evaluation, 40 student volunteers were recruited and divided into two groups of 20 each, as depicted in Table 8.4. One group used only GPS for navigation (referred to as Group G), while the other group used the proposed GPS overlay scheme (referred to as group M). The volunteers were asked to travel along a road corridor from Lanka to Varanasi Cantt railway station and they were exposed to different mediums during their travels. There were three possible paths for the volunteers to take between the two locations. The volunteers all had their own smartphones running on the Android operating system, with sufficient RAM and battery to run the GPS-only and the developed applications. The details of all the volunteers are provided in the Table 8.4, which likely includes their names, ages, phone models and other relevant information.

The purpose of this evaluation is likely to compare the performance of the GPS overlay scheme with traditional GPS navigation. By testing both methods in real-world conditions and with a diverse group of volunteers, the researchers can gather valuable data on the effectiveness and usability of the proposed approach. According to the findings of this case study, the average travel time of the volunteers who used the GPS overlay scheme (M) was lower than that of the volunteers who relied solely on GPS (G). The augmentation of GPS with additional data helped to accurately predict travel time, resulting in identifying the most efficient routes during peak hours. As a result, the GPS overlay scheme was able

to save both time and money for the volunteers who used it. The quantitative analysis reveals a 15% reduction in travel time and an almost 17% decrease in costs. These findings are based on the movements of volunteers within a smaller radius of 10 kilometers. It is anticipated that the improvements will be even more substantial with larger trip radii and an increased number of volunteers.

8.3 Field implications from this work

The field implications from this work are listed as follows:

1. **Enhanced Traffic Management:** Implementation of the machine learning-based overlay technique could lead to more accurate and reliable road traffic predictions, enabling authorities to better manage traffic flow and congestion in urban areas.

2. **Improved Navigation Systems:** Integrating the proposed technique into navigation systems utilizing GPS data could result in more efficient route planning and real-time traffic updates for drivers, reducing travel time and fuel consumption.

3. **Urban Planning and Development:** Accurate traffic prediction facilitated by the overlay technique can provide valuable insights for urban planners and policymakers, aiding in the design of more efficient road networks and transportation infrastructure to accommodate growing urban populations.

4. **Environmental Impact Reduction:** By optimizing traffic flow and reducing congestion, the proposed technique can lower vehicle emissions and mitigate the environmental impact of urban transportation, contributing to sustainability goals and air quality improvement efforts.

5. **Economic Benefits:** Improved traffic prediction can lead to cost savings for businesses involved in transportation and logistics by optimizing delivery routes and schedules, minimizing delays and enhancing overall operational efficiency.

6. **Public Safety:** More accurate traffic predictions enable better emergency response planning and management, ensuring timely access to emergency services and enhancing public safety in urban areas.

7. **Data-driven Decision Making:** The machine learning-based approach employed in the overlay technique highlights the importance of utilizing data-driven methodologies in addressing complex transportation challenges, setting a precedent for future research and development in traffic prediction and management.

8. **Technological Advancement:** The development and implementation of innovative techniques like the proposed overlay method demonstrate the potential of technology, particularly machine learning, to revolutionize traditional approaches to traffic prediction and transportation management.

9. **Societal Impact:** By reducing traffic congestion and improving overall transportation efficiency, the proposed technique can enhance the quality of life for residents in urban areas, offering them a smoother and more convenient commuting experience.

10. **Research and Development Opportunities:** The successful application of the overlay technique opens up avenues for further research and development at the intersection of machine learning, GPS technology and transportation engineering, fostering innovation and advancement in the field.

8.4 Conclusion

This study presents a novel approach to enhancing GPS-based road traffic prediction by integrating classified vehicle counts and vehicle categories, resulting in improved accuracy in predicting traffic congestion and delays. While popular navigation tools like Google Maps provide valuable traffic information, they lack detailed insights into vehicle types and counts, limiting their predictive capabilities. Our research addresses this limitation by leveraging video recording cameras to count and classify vehicles in two mid-sized Indian cities manually. Machine learning-based forecasting models were developed to predict traffic volume by integrating the vibration data with GPS coordinates. The findings underscore the effectiveness of our approach, emphasizing the significance of incorporating classified vehicle counts and categories into GPS-based traffic prediction systems. A comparative case study conducted between two locations demonstrates the tangible benefits of our technique, showcasing a significant reduction in travel time. These results underscore the practical utility of our approach in facilitating informed route selection and efficient traffic management. In conclusion, our study not only contributes to advancing the field of traffic prediction but also underscores the importance of integrating detailed vehicle information with GPS-based systems. Moving forward, such integration holds promise for enhancing transportation efficiency, improving urban mobility and ultimately enhancing the quality of life for commuters in congested urban areas.

One important limitation of our study is the reliance on manual intervention at various stages, introducing subjectivity and potential biases that may affect the overall robustness of the prediction and forecasting models. The complexity of developing and transferring the model to different geographical areas presents a significant challenge. The necessity for data collection beforehand, before actual deployment, hinders the model's adaptability to diverse real-world scenarios. The conducted experiment is controlled and variations

may arise in large-scale implementations. The controlled nature of the experiment may not fully capture the dynamic and complex nature of real-world traffic scenarios.

Future research in this area holds the potential to advance and refine traffic prediction models, leading to greater accuracy and reliability. This advancement could have a substantial impact on reducing mobility time and enhancing overall transportation efficiency. Conducting studies on a larger scale, beyond the controlled environment, will provide valuable insights into the real-world applicability and performance of the developed models.