4. METHODOLOGY

4.1 AIM OF STUDY

4.1.1 Objective:

The primary objective of this study is to explore the use of Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) models in predicting traffic volume. These models are well-suited for handling sequential data and are known for their capabilities in capturing temporal dependencies, which are crucial for understanding traffic flow dynamics.

The study aims to evaluate how these models might be leveraged to improve urban traffic management by providing accurate traffic volume predictions, thereby aiding in the reduction of congestion and enhancing transportation efficiency.

4.1.2 Significance:

Effective traffic management is a fundamental challenge in urban areas, where congestion can lead to significant delays, increased pollution, and reduced quality of life. The potential of advanced machine learning models like LSTM and GRU to provide more accurate traffic predictions lies in their ability to incorporate both temporal patterns and external factors such as weather conditions.

This study seeks to investigate these capabilities, offering preliminary insights into how these models could contribute to a more robust and efficient urban traffic management system.

4.2 OVERVIEW OF THE MODELS

4.2.1 LSTM Model:

he Long Short-Term Memory (LSTM) model is a type of recurrent neural network (RNN) that has been widely utilized in tasks involving sequential data. LSTM networks are equipped with a unique architecture that includes memory cells and gating mechanisms—specifically, input, forget, and output gates—that allow the model to retain or discard information as necessary.

This structure enables the LSTM to maintain context over long sequences, making it particularly effective in scenarios where long-term dependencies are critical. In this study, the LSTM model is explored for its potential to predict traffic volumes by capturing complex temporal patterns within the data.

4.2.2 GRU Model:

The Gated Recurrent Unit (GRU) model is a simplified variant of the LSTM, designed to achieve similar outcomes with fewer computational requirements. GRUs merge the forget and input gates into a single update gate and combine the cell and hidden states, resulting in a more streamlined architecture. This simplification is hypothesized to make GRUs faster to train while still effectively capturing essential temporal dependencies. The study investigates whether this efficiency might offer advantages in specific contexts, such as when computational resources are limited or when faster model training is desirable.

Both models are applied to historical traffic data to examine their capacity to predict future traffic volumes. The LSTM model's more complex architecture may provide fine-tuned control over memory retention, potentially making it effective in capturing detailed traffic patterns. Conversely, the GRU model's simplified structure is evaluated for its potential to balance performance with computational efficiency.

4.3 RESPONSE (DEPENDENT) AND INDEPENDENT VARIABLE(S)

The study's exploratory nature requires an examination of how various factors influence traffic volume, the primary outcome of interest. Traffic volume is defined as the number of vehicles passing a specific point on the road within a given time frame, and accurately predicting this variable is crucial for effective traffic management. To explore the potential impact on traffic volume, the study considers several independent variables:

- **Date and Time:** Temporal features, such as the hour of the day, day of the week, and month, are hypothesized to play a significant role in determining traffic patterns. Traffic volumes are expected to exhibit cyclical behaviour, with peaks during morning and evening rush hours and differences between weekdays and weekends.

- **Weather Conditions:** Weather-related factors, including temperature, precipitation, and cloud cover, are anticipated to affect traffic behaviour. This study explores whether incorporating these variables into the models can enhance prediction accuracy by allowing the models to account for external environmental conditions.

- **Holiday Information:** Holidays are likely to introduce variations in traffic patterns, deviating from regular working days. Including a binary indicator for holidays is expected to help the models adjust their predictions accordingly.

- **Previous Traffic Volume:** Historical traffic data is expected to be a strong predictor of future traffic volumes. The study examines how well the models leverage past patterns to inform their predictions, particularly in a time-series context where the relationship between past and future data points is critical.

4.4 FACTORS, LEVELS, AND PARAMETERS

This section outlines the key factors and parameters considered in the exploratory phase of this study. These elements are crucial for understanding how the models might perform under different conditions:

4.4.1 Number of Layers:

The number of layers in a neural network determines its depth and complexity. A deeper network has the potential to learn more intricate patterns from the data, which could be beneficial for capturing the complexities of urban traffic flows. The study suggests experimenting with different numbers of layers (e.g., 2, 3, or 4 layers) to find an optimal balance between model complexity and performance.

4.4.2 Units per Layer:

The number of neurons in each layer affects the model's capacity to process and store information. More units can increase the model's ability to capture detailed patterns but also make the model more computationally intensive. The study recommends testing different configurations of units per layer (e.g., 128, 64, 32) to evaluate how these adjustments influence the models' predictive power.

4.4.3 Learning Rate:

The learning rate is a critical parameter that determines how quickly the model adjusts its weights during training. A higher learning rate can speed up training but may lead to overshooting the optimal solution, while a lower learning rate allows for more precise adjustments. The study suggests exploring different learning rates (e.g., 0.001, 0.01, 0.0001) and potentially using learning rate schedules to fine-tune the models over time.

4.4.4 Batch Size:

The batch size, which refers to the number of samples processed before the model's internal parameters are updated, can influence the stability and efficiency of the training process. The study proposes testing a range of batch sizes (e.g., 32, 64, 128, 256) to find an optimal balance, potentially starting with smaller batch sizes for more precise updates and increasing them as the models stabilize.

4.4.5 Activation Functions:

The batch size, which refers to the number of samples processed before the model's internal parameters are updated, can influence the stability and efficiency of the training process. The study proposes testing a range of batch sizes (e.g., 32, 64, 128, 256) to find an optimal balance, potentially starting with smaller batch sizes for more precise updates and increasing them as the models stabilize.

4.4.6 Regularization Techniques:

Regularization techniques such as dropout and L2 regularization are considered to prevent overfitting. The study proposes experimenting with different dropout rates (e.g., 0.2, 0.3) and L2 regularization penalties (e.g., 0.001, 0.01) to evaluate how these adjustments impact the models' generalization ability.

4.4.7 Considerations:

Throughout the experimentation phase, the study remains flexible in tuning these parameters. If early experiments suggest overfitting despite using regularization, the study may revisit the model architecture, potentially simplifying it or increasing dropout rates. Conversely, if the models under-fit, adjustments such as increasing the number of layers or units, or modifying the learning rate, may be considered.

4.5 EXPERIMENTAL DESIGN

The experimental design of this study is structured to explore the potential of LSTM and GRU models in predicting traffic volumes. Given the exploratory nature of this work, the design allows for adjustments based on preliminary findings.

4.5.1 Data Preparation:

Data preparation is a foundational step in this exploration, involving the transformation of raw traffic data into a format suitable for model training. The following steps are considered essential for this process:

- **Handling Missing Values:** Missing data can potentially distort predictions, so the study considers techniques like imputation to address any gaps or the removal of incomplete records if necessary.

- **Scaling Features:** Scaling numerical features, such as temperature, is expected to ensure consistency across the dataset, making it easier for the models to process these variables.

- **Encoding Categorical Variables:** Categorical variables, such as weather conditions, are converted into numerical formats using methods like one-hot encoding, enabling the models to incorporate these factors into their predictions.

4.5.2 Randomization (Train/Test Split):

In this study, we aimed to build a robust predictive model for traffic flow using a dataset composed of time-ordered vectors, each representing traffic data for a specific time period. To ensure the accuracy and generalizability of our model, it was crucial to carefully consider how we split the dataset into training and testing subsets.

Given the temporal nature of the data, a simple random split could introduce potential biases, as it might allow the model to "peek" into future trends during training, thereby inflating its performance. To mitigate this risk and more accurately simulate real-world conditions, we employed a sequential, time-based data splitting approach.

Sequential Time-Based Split:

The dataset, spanning several consecutive days of traffic data, was divided based on time. Specifically, the first 70% of the dataset, corresponding to the earlier time periods, was used to train the model. This segment of data represents historical traffic patterns, providing the model with a comprehensive view of how traffic volumes evolved over time.

The rationale behind this approach is to mimic a scenario where the model learns from past data, which is a common practice in time-series forecasting.

By feeding the model data from the first 70% of the time periods, we allow it to recognize patterns, trends, and cyclical behaviours inherent in the traffic data. These insights are crucial for making informed predictions about future traffic volumes.

Testing on Subsequent Data:

Once the model was trained on the first 70% of the dataset, its performance was evaluated using the remaining 30% of the data. This portion of the dataset corresponds to the later time periods that the model had not encountered during training.

By testing the model on this unseen data, we aimed to assess its ability and make accurate predictions in a real-world context, where future traffic conditions are unknown.

This sequential splitting method offers several advantages:

- **Realistic Scenario Simulation:** By training on past data and testing on future data, this closely simulate the conditions under which the model would be used in practice. This approach ensures that the model's performance metrics reflect its true predictive capabilities, rather than being artificially inflated by data leakage.

- A**voidance of Data Leakage:** Randomly splitting the dataset could lead to situations where data points from the same time period are present in both the training and testing sets. This overlap could allow the model to interpolate between closely related points, resulting in predictions that seem highly accurate but do not genuinely reflect the model's ability to predict future trends.

- **Robustness of Predictions:** By focusing the training on historical data and testing on subsequent, unseen data, we ensure that the model's predictions are based on learned patterns and not on the ability to interpolate between known data points. This approach provides a more robust measure of how the model will perform when faced with new, unseen traffic conditions.

4.5.3 Experiment Performance and Revisions:

Throughout the experimental phase, the study conducts multiple rounds of training and testing to refine the models and improve their performance.

Given the exploratory nature of this work, the following steps are considered:

- **Initial Experiment:** The LSTM and GRU models are first trained with default settings over a substantial number of epochs. This initial phase tests various data preprocessing methods to establish a baseline for the models' performance.

- **Model Tuning:** Based on preliminary results, the models were fine-tuned by adjusting key parameters such as the number of layers, units per layer, and learning rate. This tuning process was intended to enhance the models' ability to capture both common patterns and outliers in the traffic data, ensuring that the models could generalize well to new data.

- **Data Imbalance Handling:** Recognizing that traffic data may be imbalanced, with more data points during peak hours compared to off-peak periods, the study explored techniques such as under-sampling and oversampling to address this issue. These methods were applied to create a more balanced dataset, potentially improving the models' ability to predict rare but significant traffic patterns, such as those occurring during late-night hours or extreme weather conditions.

4.5.4 Measuring Classifier Performance:

In this study, evaluating the performance of the LSTM and GRU models is crucial for understanding their effectiveness in predicting traffic volumes. Several performance metrics are employed to provide a comprehensive assessment, with a focus on both accuracy and the ability to handle the nuances of the dataset.

Precision and Recall:

Given the potential imbalance in traffic data, precision and recall were prioritized as key metrics. Precision measures the ratio of correctly predicted positive observations to the total predicted positives, highlighting the model's accuracy in predicting high-traffic volumes.

Recall measures the ratio of correctly predicted positive observations to all observations in the actual class, focusing on the model's ability to identify all instances of high traffic volumes.

Mean Absolute Error (MAE):

MAE is used to measure the average magnitude of errors between predicted and actual values. It provides a straightforward interpretation of prediction accuracy by indicating how much, on average, the predictions deviate from the actual traffic volumes. This metric is particularly useful in scenarios where absolute differences are more critical than relative differences.

Mean Squared Error (MSE) and Root Mean Squared Error (RMSE):

MSE emphasizes larger errors by squaring the differences between predicted and actual values, making it a crucial metric for understanding the model's ability to minimize significant errors. RMSE, the square root of MSE, presents the error in the same units as the original data, offering an intuitive understanding of prediction accuracy.

$$\mathbf{MSE = (1/n) * \Sigma (y_i - \hat{y}_i)^2}$$

where:

- n is the number of data points.
- y_i is the actual value for the ith data point.
- $\hat{y}_i$ is the predicted value for the ith data point.
- $(y_i - \hat{y}_i)^2$ is the squared difference between the actual and predicted values.

Both metrics are essential for assessing how well the models perform in minimizing errors that could have substantial impacts on traffic management decisions.

Application of Metrics:

During the experimental phase, these metrics were applied to both the training and test sets to monitor the models' performance. Precision and recall were closely monitored to ensure that the models could accurately identify high-traffic periods, which are critical for effective traffic management. MAE, MSE, and RMSE were used to gauge the overall accuracy of the models, with particular attention to how well the models handled peak traffic volumes and periods of low traffic.

4.5.5 Algorithm Comparison and Selection:

Both the LSTM and GRU models were implemented to predict traffic volumes. The selection and comparison of these models are essential steps to determine which model provides better predictive performance for urban traffic management applications

Comparison Criteria:

The models were compared based on several performance metrics, including precision, recall, accuracy, MAE, MSE, and RMSE. Each metric offered a different perspective on the models' performance, allowing for a comprehensive evaluation.

The key criteria for comparison were how well each model captured the traffic patterns, particularly during peak hours, and how effectively they minimized prediction errors.