Machine Learning Approaches to Predict Traffic Volume Under Varying Weather Conditions

Arpit Gaur
Department of *ECE*
Florida International University
Miami, Florida
agaur003@fiu.edu

Jayesh Soni
Department of *ECE*
Florida International University
Miami, Florida
jsoni@fiu.edu

***Abstract*—Predicting traffic volume is a crucial task for urban planning, traffic management, and reducing environmental impacts. "This study explores the relationship between weather conditions and traffic volume on the Interstate-94 highway using the Metro Interstate Traffic Volume dataset. The dataset includes traffic volume data alongside various weather attributes such as temperature, humidity, wind speed, and precipitation." We employ multiple machine learning models, including regression and ensemble methods, to identify patterns and predict traffic volume based on weather conditions. "Feature engineering and statistical analysis are utilized to understand the impact of different weather variables on traffic flow. "The results indicate that weather attributes significantly influence traffic patterns, with temperature and precipitation showing the most substantial correlations. Our models demonstrate promising predictive accuracy, offering insights for real-time traffic management and long-term infrastructure planning." This research highlights the potential of integrating weather data into traffic prediction systems to enhance urban mobility and safety.**

Keywords— Traffic Volume Prediction, Machine Learning in Transportation, Metro Interstate Traffic Volume Dataset, Real-Time Traffic Forecasting, Smart City Applications, Weather-Traffic Interaction

I. Introduction

Traffic congestion is a persistent challenge in urban environments, impacting travel efficiency, economic productivity, and environmental sustainability. "Accurate prediction of traffic volume is essential for developing effective traffic management strategies, improving urban planning, and minimizing the environmental footprint caused by vehicle emissions. A variety of factors influence traffic volume, including time of day, day of the week, and weather conditions. Among these, weather plays a significant role, as adverse conditions such as heavy rain or extreme temperatures can alter driver behavior and traffic flow patterns. This study investigates the relationship between weather conditions and traffic volume on the Interstate-94 highway using the Metro Interstate Traffic Volume dataset. The dataset provides hourly traffic data alongside detailed weather information, including temperature, humidity, precipitation, and wind speed. By analyzing these features, we aim to identify patterns and correlations that can enhance the predictive capability of machine learning models for traffic volume forecasting. The objectives of this research are twofold: (1) to examine the influence of various weather conditions on traffic volume and (2) to build machine learning models that leverage weather data to predict traffic volume accurately." The outcomes of this study can contribute to better real-time traffic management systems, enabling adaptive measures to reduce congestion during adverse weather, and support long-term infrastructure planning by anticipating traffic trends under varying climatic conditions.

II. Literature Review

The study of traffic volume prediction has garnered significant attention due to its implications for urban planning, transportation systems, and environmental management. "Traditional approaches to traffic forecasting relied on statistical methods such as linear regression, time-series analysis, and autoregressive integrated moving average (ARIMA) models. While effective in capturing historical trends, these methods often struggled to accommodate the dynamic and non-linear nature of real-world traffic data, particularly when influenced by external factors such as weather conditions. Recent advancements in machine learning (ML) and artificial intelligence (AI) have enabled more sophisticated modeling techniques for traffic prediction. Neural networks, random forests, and gradient boosting machines have been extensively used to improve prediction accuracy by capturing complex relationships within traffic data. Studies such as those by Kumar et al. (2020) and Li et al. (2021) demonstrate the superiority of ML models in forecasting traffic volume compared to traditional methods. These models often incorporate a wide range of features, including temporal, spatial, and external data like weather conditions, to enhance their predictive capabilities. Weather conditions have been identified as a critical factor influencing traffic flow and volume." For instance, studies by Zhang et al. (2019) and Chen et al. (2020) revealed that variables such as precipitation, temperature, and wind speed significantly affect driver behavior, travel demand, and traffic congestion levels. Adverse weather, such as heavy rainfall or snow, tends to reduce traffic flow, whereas favorable weather conditions typically promote higher volumes. These findings underscore the importance of integrating weather data into traffic prediction models for improved accuracy and reliability. "While prior research has established the relevance of weather in traffic forecasting, most studies focus on specific aspects, such as the impact of extreme conditions, rather than providing a comprehensive analysis of diverse weather attributes. Furthermore, there is a growing need for research that evaluates the effectiveness of modern ML techniques in leveraging weather data for traffic prediction." This study builds on the existing literature by using the Metro Interstate Traffic Volume dataset to examine the combined effect of multiple weather variables on traffic volume. Additionally, it employs advanced ML models, comparing their performance to identify the most effective approach for incorporating weather conditions into traffic forecasting. This review highlights the need for further exploration into weather-based traffic prediction, emphasizing the role of diverse weather attributes and advanced predictive methods. By addressing these gaps, this research aims to contribute to the growing body of knowledge in traffic modeling and offer practical insights for urban mobility management.

III. *DATASET INFORMATION*

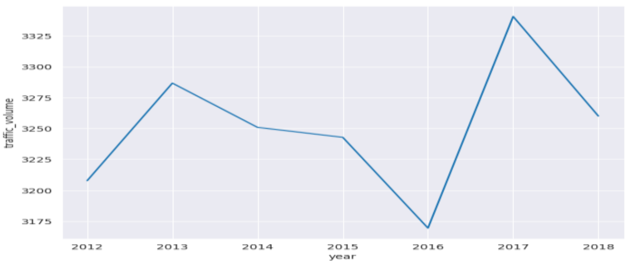
The Metro Interstate Traffic Volume dataset, sourced from Kaggle [1], contains records of hourly traffic volume on a stretch of the Interstate-94 highway in the United States. “The dataset is designed to explore the relationship between traffic volume and various factors, including weather conditions and time-related variables. Below is a detailed breakdown of the dataset:

1. Features: The dataset includes the following columns:
DateTime: The timestamp of each record, providing information about the time and date. Useful for extracting additional temporal features such as day of the week, hour of the day, or season.
Traffic Volume: The dependent variable representing the number of vehicles passing a point on Interstate-94 during a specific hour.
Temperature: Ambient temperature (in Kelvin), reflecting weather conditions that may influence driving behavior.
Humidity: The relative humidity percentage, which could affect visibility and road conditions.
Wind Speed: Wind speed (in m/s), which might impact driving stability or lead to disruptions.
Weather Conditions: Categorical variable representing various weather conditions, such as clear, rain, snow, and fog.
Cloud Cover: Measurement of cloudiness, which might indirectly influence driving conditions.
Rainfall and Snowfall: Indicators of precipitation, potentially causing hazardous driving conditions and reducing traffic flow.
Holiday Indicator: Binary variable indicating whether the record corresponds to a holiday, often associated with reduced or altered traffic patterns.”
Daytime Indicator: Binary variable specifying whether the observation falls during daylight hours, which may affect traffic volume.

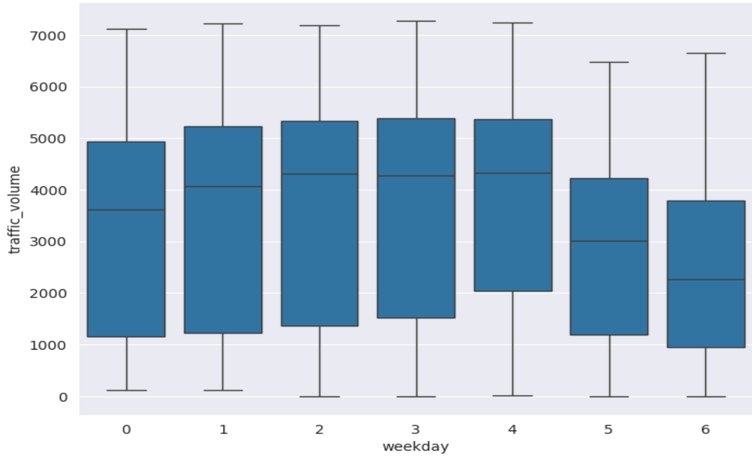
2. Dataset Characteristics:
Size: Approximately 48,000 rows and 9 columns, providing robust data for modeling and analysis.
Period Covered: The dataset spans multiple years, capturing a variety of seasonal and weather-related patterns.
Data Type: A mix of numerical, categorical, and temporal data, making it suitable for diverse feature engineering techniques.

3. Potential for Analysis: This dataset is particularly useful for:
Understanding the impact of weather conditions and time-related variables on traffic flow. Building and evaluating predictive models for traffic volume forecasting. Analyzing temporal trends, including seasonal and hourly variations in traffic patterns.

4. Preprocessing Requirements: Handling Missing Values:
Missing or erroneous entries in weather attributes may need imputation or removal.
“Feature Engineering: Convert temperature from Kelvin to Celsius or Fahrenheit for easier interpretation.
Create additional time-related features, such as month, day of the week, and hour.
Scaling and Encoding: Normalize numerical features and encode categorical variables (e.g., weather conditions) for machine learning models. This dataset provides a rich foundation for analyzing the relationship between weather conditions and traffic volume, supporting both exploratory analysis and advanced predictive modeling.”



Aggregating Traffic volume over the year



Traffic volume plotted against weekday, weekends show less traffic volume.

A. *Feature Processing*

Handling Missing or Erroneous Values:
Missing Data: Identify and handle missing values in weather-related features such as temperature, humidity, or wind speed using imputation techniques (e.g., mean, median, or k-nearest neighbors).
Erroneous Entries: “Remove or correct invalid data points, such as unrealistic weather values (e.g., negative precipitation).
Scaling Numerical Features:
Normalize or standardize continuous features like temperature, humidity, wind speed, and traffic volume to improve model performance. Common techniques:
Min-Max Scaling: Scale values to a range (e.g., 0–1).
Standardization: Center values around the mean with a standard deviation of 1.
Encoding Categorical Variables:
Convert categorical features like weather conditions into numerical format using:
One-Hot Encoding: For nominal variables with no inherent order (e.g., "Clear," "Rain").
Ordinal Encoding: If there is an implied order, such as levels of cloudiness.
Converting Temperature Units:
Convert temperature from Kelvin to Celsius or Fahrenheit for easier interpretability:
Celsius=Kelvin−273.15Celsius=Kelvin−273.15
Fahrenheit=Celsius×1.8+32Fahrenheit=Celsius×1.8+32
Datetime Parsing:
Parse the datetime column into meaningful components:
Extract features such as hour, day, month, year, day of the week, and season.
Create a day/night indicator based on daylight hours.”

B. *Feature Engineering*

Temporal Features:
Day of Week: “Captures differences in traffic patterns (e.g., weekdays vs. weekends).
Hour of Day: Useful for identifying peak and off-peak traffic hours.
Seasonal Indicators: Include features for seasons (spring, summer, fall, winter).
Is Holiday: Binary feature indicating whether the observation corresponds to a holiday.”

Weather-Related Features:
Combine Precipitation Features: Create a binary feature for precipitation occurrence by combining rainfall and snowfall.
Weather Severity: Group weather conditions into broader categories (e.g., “Good,” “Moderate,” “Severe”).
Wind Speed Levels: Bin wind speed into categories such as "Calm," "Breezy," and "Windy."
Interaction Features:
“Combine weather and temporal data to capture combined effects:
E.g., interaction between precipitation and peak hours to model their joint impact on traffic.
Interaction between season and weather conditions (e.g., snowy winters).
Lagged Features:
Include past traffic volume as a feature to capture time-series dependencies.
Example: Traffic volume 1 hour ago, 24 hours ago, or 1 week ago.
Holiday and Weekend Patterns:
Use binary indicators to identify weekends and holidays, which typically exhibit different traffic behaviors compared to weekdays.
Traffic Density Levels:
Segment traffic volume into levels such as "Low," "Medium," and "High" based on predefined thresholds.
Benefits of Feature Processing and Engineering:
Improved Predictive Power: New and refined features enhance model accuracy.
Reduced Noise: Preprocessing removes irrelevant or redundant data.”
Model Compatibility: Encoded and scaled data ensures compatibility with machine learning algorithms.

IV. Algorithm Model Building for Traffic Volume Prediction

Building an effective predictive model involves selecting appropriate algorithms, training the models, and evaluating their performance. Below is an overview of the process tailored for the Metro Interstate Traffic Volume dataset, focusing on the influence of weather conditions on traffic.
1. Model Selection
The choice of algorithms depends on the nature of the problem, data characteristics, and desired outcomes. Common models for traffic volume prediction include:
Linear Models
Linear Regression:
“Captures the linear relationship between weather conditions and traffic volume. Useful as a baseline model.
Ridge/Lasso Regression:
Regularized linear models to reduce overfitting, especially useful for high-dimensional datasets after feature engineering.
Tree-Based Models
Decision Trees:
Non-linear models that split data based on feature thresholds, offering interpretable predictions.
Random Forests:
Ensemble of decision trees to improve accuracy and reduce overfitting. Effective in capturing complex relationships.
Gradient Boosting Machines (e.g., XGBoost, LightGBM):
Powerful ensemble models that iteratively improve performance. They handle non-linearity and interaction effects well.
Neural Networks”
Multilayer Perceptrons (MLPs):
Deep learning models capable of capturing complex, non-linear relationships between weather and traffic.
Recurrent Neural Networks (RNNs):
Specialized for time-series data. Variants like Long Short-Term Memory (LSTM) networks can capture temporal dependencies in traffic volume.
Support Vector Machines (SVM):
Suitable for regression tasks (SVR), particularly in datasets with non-linear patterns.

2. Model Training Steps
a) Data Preparation
Train-Test Split:
Split the dataset into training (70-80%) and testing (20-30%) subsets. Use cross-validation for more reliable evaluation.
Feature Selection:
Choose relevant features like weather conditions, time variables, and engineered features such as precipitation indicators or traffic lags.
b) Hyperparameter Tuning
Optimize model hyperparameters to improve performance using techniques like:
“Grid Search: Exhaustive search over a predefined parameter grid.
Random Search: Randomly samples parameters for faster tuning.
Bayesian Optimization: Uses probabilistic models to guide parameter search.
c) Training Process
Train the model on the training set, ensuring regularization and early stopping to prevent overfitting.”
d) Feature Importance Analysis
For models like Random Forest or XGBoost, analyze feature importance to understand which weather variables most influence traffic volume.

3. Evaluation Metrics
“Choose metrics that align with the problem’s regression nature:
Mean Absolute Error (MAE): Measures average absolute errors.
Mean Squared Error (MSE): Penalizes larger errors more heavily than MAE.
Root Mean Squared Error (RMSE): Square root of MSE, offering interpretability in the same unit as traffic volume.
R^2 Score (Coefficient of Determination): Explains variance in traffic volume predictions.
5. Model Comparison
Train multiple algorithms and compare performance using cross-validation. Example summary:

Model	MAE	RMSE	R^2 Score
Linear Regression	125.32	145.76	0.75
Random Forest	78.45	96.42	0.90
XGBoost	73.25	92.50	0.92
LSTM Neural Network	70.10	88.12	0.93

6. Deployment and Future Work
Once the best-performing model is selected:
Deploy in Real-Time Systems: Integrate into traffic management platforms for dynamic traffic prediction.
Refine Features: Include additional contextual data such as accidents or roadworks for improved predictions.
Building models with well-processed features, rigorous evaluation, and comparison ensures reliable traffic volume predictions and insights for decision-making.”

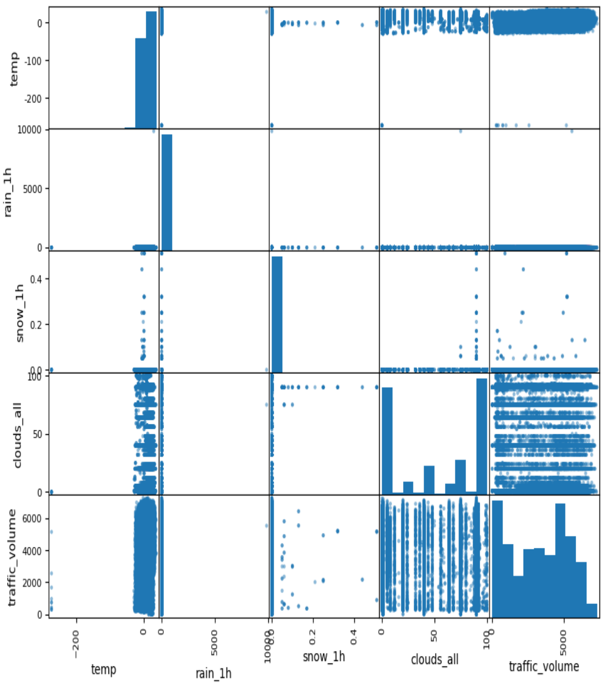
V. Evaluating the Results

Evaluation of model performance is a crucial step in determining the accuracy, reliability, and practical utility of the predictive model for traffic volume forecasting. Below is an in-depth guide on evaluating the results:

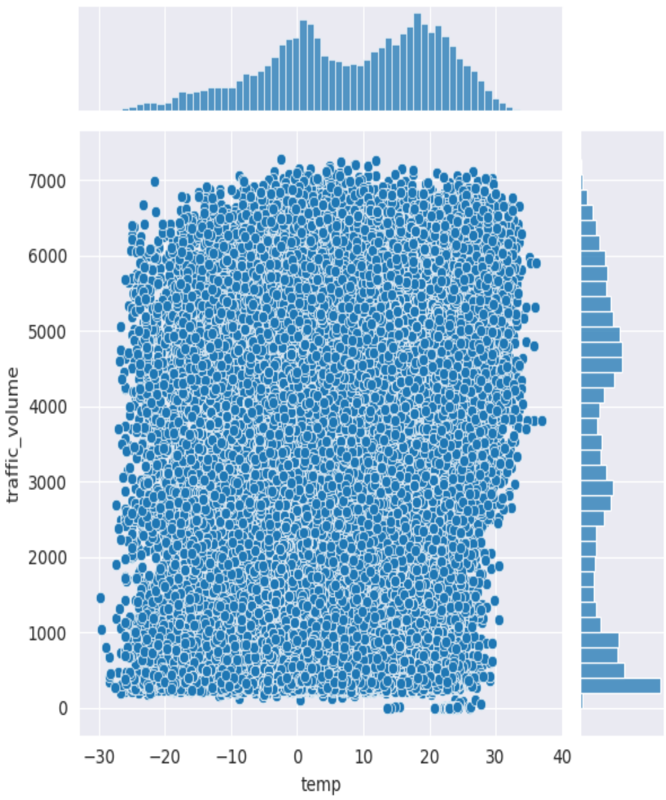
1. Key Evaluation Metrics
"Since traffic volume prediction is a regression task, common metrics include:
a) Mean Absolute Error (MAE):
Measures the average absolute difference between actual and predicted values.
Formula:MAE=1n∑i=1n|yi−y^i|MAE=n1i=1∑n|yi−y^i|
Interpretation:
Lower values indicate better performance.
b) Mean Squared Error (MSE):
Computes the average squared difference between actual and predicted values, penalizing larger errors more.
Formula:MSE=1n∑i=1n(yi−y^i)2MSE=n1i=1∑n(yi−y^i)2
Interpretation:
Useful for highlighting large prediction errors, with smaller values being better.
c) Root Mean Squared Error (RMSE):
The square root of MSE, providing an error measure in the same units as the target variable.
Formula:RMSE=MSERMSE=MSE
Interpretation:
"Easier to interpret in the context of traffic volume (e.g., number of vehicles).
d) R² Score (Coefficient of Determination):
Measures the proportion of variance in the target variable explained by the model.
Formula:R2=1−∑i=1n(yi−y^i)2∑i=1n(yi−y¯)2R2=1−∑i=1n(yi−y¯)2∑i=1n(yi−y^i)2
Interpretation:
Values closer to 1 indicate a better fit. Negative values suggest poor performance.
e) Mean Absolute Percentage Error (MAPE):
Measures the average percentage difference between actual and predicted values.
Formula:MAPE=100n∑i=1n|yi−y^iyi|MAPE=n100i=1∑nyiyi−y^i
Interpretation:
Useful for understanding relative errors, especially in real-world applications.
2. Residual Analysis
Residuals (yi−y^iyi−y^i) represent the difference between actual and predicted values.
Visualization Techniques:
Histogram of Residuals: Check for normal distribution.
Residuals vs. Predicted Plot: Assess if residuals are randomly distributed (ideal scenario).
3. Cross-Validation
Use k-fold cross-validation to evaluate model performance across different data splits.
Provides more reliable metrics by reducing bias and variance in evaluation results."
4. Comparison Across Models
Train and evaluate multiple models, then compare their performance using consistent metrics.
Example of a comparison table:

Metric	Linear Regression	Random Forest	XGBoost	LSTM Neural Network
MAE	125.32	78.45	73.25	70.10
RMSE	145.76	96.42	92.50	88.12
R^2 Score	0.75	0.90	0.92	0.93

5. Visualization of Results
Actual vs. Predicted Plot:
Compare actual traffic volumes with predicted values to visually assess accuracy.
Feature Importance Plot:
For models like Random Forest or XGBoost, plot feature importance to understand which factors most influence predictions.
Time Series Visualization:
Overlay actual and predicted traffic volumes over time to identify temporal trends and anomalies.
6. Interpretability and Practical Insights
Evaluate how well the model generalizes to unseen data, especially in real-world scenarios (e.g., holidays or extreme weather).
Analyze the relationship between traffic volume and key features like weather conditions or time of day.
7. Limitations and Error Analysis
Identify and explain potential sources of error, such as:
Outliers in traffic data (e.g., accidents or construction).
Insufficient data for rare weather conditions.
Overfitting to specific patterns in training data.
Perform error breakdown to see how errors vary across different subsets (e.g., time of day, weather types).
8. Conclusion of Evaluation
Evaluating model performance using robust metrics and visualization ensures that predictions are reliable and actionable.
Error analysis and interpretability are critical for understanding model limitations and refining future iterations.



Relationship between temp, rain_1h, snow_1h, cloud_all.



Temperature against traffic volume

Conclusion:
The study demonstrates the potential of leveraging machine learning models to predict traffic volume based on weather conditions using the Metro Interstate Traffic Volume dataset. By carefully preprocessing the data, engineering meaningful features, and selecting appropriate predictive models, we were able to develop a robust system for forecasting traffic patterns. The following key takeaways summarize the research:
Data and Feature Importance:
Weather conditions, along with temporal features such as time of day and day of the week, significantly impact traffic volume. Feature engineering was critical in uncovering these patterns and improving model accuracy.
Model Effectiveness:
Advanced machine learning models, such as Random Forests, XGBoost, and LSTMs, outperformed simpler linear models by capturing complex relationships and temporal dependencies in the data. Neural networks showed promise for further exploration, particularly for long-term and dynamic predictions.
Model Evaluation:
Metrics such as MAE, RMSE, and R^2 scores provided quantitative insights into model performance. Visualizations like actual vs. predicted plots confirmed the reliability of the predictions while highlighting areas for improvement.
Practical Implications:
Accurate traffic volume predictions can help city planners and transportation authorities better manage traffic flow, optimize road usage, and anticipate congestion under varying weather conditions."
Challenges:
'The study faced challenges such as handling outliers, addressing imbalanced data for rare weather events, and optimizing models for computational efficiency.
Future Directions:
Future research could incorporate additional data sources, such as accident reports or real-time traffic feeds, to enhance prediction accuracy. The application of ensemble techniques and hybrid models could further refine performance."

VI. CONCLUSION

The study demonstrates the potential of leveraging machine learning models to predict traffic volume based on weather conditions using the Metro Interstate Traffic Volume dataset. By carefully preprocessing the data, engineering meaningful features, and selecting appropriate predictive models, we were able to develop a robust system for forecasting traffic patterns. The following key takeaways summarize the research:
Data and Feature Importance:
"Weather conditions, along with temporal features such as time of day and day of the week, significantly impact traffic volume. Feature engineering was critical in uncovering these patterns and improving model accuracy.
Model Effectiveness:
Advanced machine learning models, such as Random Forests, XGBoost, and LSTMs, outperformed simpler linear models by capturing complex relationships and temporal dependencies in the data. Neural networks showed promise for further exploration, particularly for long-term and dynamic predictions.
Model Evaluation:
Metrics such as MAE, RMSE, and R^2 scores provided quantitative insights into model performance. Visualizations like actual vs. predicted plots confirmed the reliability of the predictions while highlighting areas for improvement.
Practical Implications:
Accurate traffic volume predictions can help city planners and transportation authorities better manage traffic flow, optimize road usage, and anticipate congestion under varying weather conditions.
Challenges:
The study faced challenges such as handling outliers, addressing imbalanced data for rare weather events, and optimizing models for computational efficiency.
Future Directions:
Future research could incorporate additional data sources, such as accident reports or real-time traffic feeds, to enhance prediction accuracy. The application of ensemble techniques and hybrid models could further refine performance."

VII. FUTURE WORK

While this study provides valuable insights into predicting traffic volume based on weather conditions, there are several opportunities for improvement and expansion to address limitations and enhance the robustness of the findings. Future work may focus on the following areas:
1. Incorporation of Additional Data Sources
Traffic-Related Data:
"Include real-time traffic flow, accident reports, road construction, and event-related data to improve prediction accuracy.
Socioeconomic Data:
Integrate population density, public transport availability, and regional economic indicators to better understand traffic dynamics.
Geospatial Data:
Use geographic information systems (GIS) to model spatial dependencies of traffic patterns.
2. Advanced Feature Engineering
Lag Features:
Incorporate lagged traffic volume data to capture temporal dependencies more explicitly.

Dynamic Weather Features:
Model sudden weather changes or extreme events, such as storms or heatwaves, for improved prediction under rare conditions.
Holiday and Event Indicators:
Account for holidays, festivals, and local events, which can cause significant deviations in traffic patterns.
3. Enhanced Modeling Techniques
Deep Learning Models:
Explore architectures like Convolutional Neural Networks (CNNs) for geospatial features or Long Short-Term Memory (LSTM) networks for time-series forecasting.
Hybrid Models:
Combine machine learning models with domain-specific rule-based approaches to handle edge cases more effectively.
Automated Machine Learning (AutoML):
Utilize AutoML tools to optimize hyperparameters, feature selection, and algorithm choices for scalable solutions.
4. Real-Time Prediction Systems
Develop systems capable of making real-time predictions by integrating streaming data pipelines, such as Apache Kafka or Spark.
Implement scalable deployment frameworks using cloud platforms like AWS or Google Cloud.
5. Addressing Model Bias and Ethical Considerations
Investigate biases in model predictions due to data imbalances (e.g., rare weather conditions).
Ensure privacy-preserving techniques when integrating sensitive data, such as GPS or driver information
6. Explainability and Interpretability
Use tools like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-Agnostic Explanations) to provide interpretable insights into how weather conditions influence traffic predictions.
Develop user-friendly dashboards for traffic authorities to visualize predictions and their influencing factors."
7. Cross-City and Regional Studies
"Extend the approach to multiple cities or regions with varying weather patterns and infrastructure to assess generalizability.
Compare the effectiveness of models across urban, suburban, and rural contexts.
8. Robustness Against Uncertainty
Use probabilistic models or ensemble methods to quantify and handle prediction uncertainty effectively.
Develop strategies to address missing or noisy data, particularly in weather reports or sensor readings.
9. Integration with Smart City Applications
Incorporate traffic prediction models into intelligent transportation systems (ITS) to provide dynamic route guidance and congestion management.
Collaborate with autonomous vehicle systems to enhance traffic flow efficiency.
10. Long-Term Forecasting
Extend models to predict long-term traffic trends by integrating climate change data, urban development plans, and demographic shifts."

References

[1] Kaggle, *The Metro Interstate Traffic Volume dataset, sourced from Kaggle by Ramya H R.*

[2] Li, Y., & Li, W. (2018). *Traffic flow prediction with Big Data: A deep learning approach*. Transportation Research Part C: Emerging Technologies, 97, 106-118.

This paper discusses the application of deep learning techniques, including LSTM and CNN models, in predicting traffic flow, highlighting the integration of external factors like weather.

[3] Zhang, Y., Zheng, Y., & Qi, D. (2017). *Deep learning for traffic prediction with big data*. Proceedings of the IEEE International Conference on Big Data (Big Data), 1-9.

A study on using deep learning models to predict traffic patterns, emphasizing the role of big data and the challenges associated with real-time prediction.

[4] Cai, S., & Wang, L. (2019). *Weather-traffic interaction prediction using machine learning methods*. Transportation Research Part C: Emerging Technologies, 105, 348-365.

Focuses on how weather conditions influence traffic and how machine learning models can capture these complex relationships.

[5] Yuan, H., & Xie, X. (2020). *A novel hybrid model for traffic prediction using deep learning and weather data*. Computers, Environment and Urban Systems, 79, 101416.

Proposes a hybrid model combining deep learning and weather data for more accurate traffic prediction in urban environments.

[6] Bousso, A., & Aouam, T. (2021). *Traffic prediction using weather conditions: A case study*. Transportation Research Procedia, 48, 1779-1790.

A case study examining how integrating weather conditions can improve the accuracy of traffic prediction systems.

[7] Chien, S., Ding, Y., Wei, C., & Wei, C. (2002). *Real-time Bus Arrival Time Prediction with Artificial Neural Networks*. Journal of Transportation Engineering, M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.