BIG DATA FRAMEWORK FOR PREDICTING AND MANAGING URBAN TRAFFIC FLOW IN SMART CITIES

CH. Divyasri[1], B. Rupa[2], B. Archana[3], B. Saicharan[4] , Dr A. Sathish Kumar[5], Dr B.Venkataramana[6]

[1] Student, B-Tech CSE (DS) 4th Year, Holy Mary Inst. of Tech. and Science, Hyderabad, TG, India, divyasrichakali14@gmail.com

[2] Student, B-Tech CSE (DS) 4th Year, Holy Mary Inst. of Tech. and Science, Hyderabad, TG, India, rupabollagani@gmail.com

[3]Student, B-Tech CSE(DS) 4th Year, Holy Mary Inst. of Tech. and Science, Hyderabad, TG, India, anakaliarchana@gmail.com

[4]Student, B-Tech CSE(DS) 4th Year, Holy Mary Inst. of Tech. and Science, Hyderabad, TG, India, saicharanbhukya13@gmail.com

[5]Assoc.prof, CSE(DS), Holy Mary Inst. Of Tech. and Science, Hyderabad, TG, India,

sathishk0711@gmail.com

[6]Assoc.prof, CSE(DS), Holy Mary Inst. Of Tech. and Science, Hyderabad, TG, India,

bandaruramana1@gmail.com

Abstract: *The fast expansion of cities along with the growing number of automobiles has created a severe problem with traffic congestion in today's urban areas. The current traffic system operates poorly because it results in extended travel durations while drivers consume excessive fuel and produce environmental damage which leads to financial losses. Traditional traffic control systems operate on fixed signal schedules which use past information to operate but they fail to manage the changing traffic situations that occur in real time. The research paper provides a Big Data framework which helps smart cities predict their urban traffic patterns and control their citywide traffic flow.*

The system collects live traffic information from multiple different data sources which include IOT traffic sensors and GPS-equipped vehicles and traffic monitoring cameras and mobile phone applications. The system uses Apache Kafka to receive live data streams which it stores in Hadoop Distributed File System (HDFS) for its storage needs and processes using Apache Spark. The research employs three machine learning and deep learning models which include Random Forest and Support Vector Regression (SVR) and Long Short-Term Memory (LSTM) networks to achieve precise traffic flow and congestion predictions. The system uses prediction results to modify traffic signals and find better routes which enhances city traffic management while building sustainable transportation networks for smart cities.

Keywords: ***Smart Cities, Traffic Congestion Management, Big Data Analytics, Real-Time Traffic Prediction, Machine Learning, Deep Learning, Intelligent Transportation Systems.***

1. Introduction: Modern cities face severe traffic congestion because their populations have grown rapidly while people have started driving more cars on their urban streets. The rising traffic congestion causes drivers to spend more time on roads while they consume additional fuel and generate environmental harm which leads to financial damage to the economy. The current traffic management systems depend on unchanging signal schedules and past traffic information which fails to manage actual traffic flow during incidents and busy times and emergency situations.

The smart city development process produces massive amounts of traffic information which stems from IOT sensors and GPS-equipped vehicles and traffic surveillance cameras and mobile apps. The data requires sophisticated processing because it contains large amounts of information which arrives at high speeds and includes different types of data. Big Data technologies enable organizations to collect real-time traffic data which they can store and analyze through their scalable systems.

The combination of Big Data frameworks with machine learning and deep learning models produces accurate predictions for traffic flow and congestion patterns. The project introduces a Big Data framework which enables smart cities to predict and control their urban traffic systems through intelligent traffic management which decreases traffic jams and enhances city transportation system.

2. Literature Review:

Urban computing introduced a new approach which uses big data from sensors and GPS devices and mobile networks to manage city operations. The study by Zheng and team demonstrated their method to analyze different urban data types which helps smart cities understand traffic patterns better [1]. Researchers studied traffic management for the first time through statistical analysis and rule-based systems which demonstrated their results by using past traffic flow information. The researchers used Kalman Filters together with time-series models to estimate current traffic conditions and predict upcoming traffic patterns during short time frames. The methods showed restricted abilities to deal with nonlinear urban traffic patterns which evolved rapidly in the city [2].

Machine learning systems evolved to include Support Vector Regression (SVR) and Random Forest models which researchers used for traffic prediction after their initial development. The models achieved better accuracy by capturing nonlinear relationships between traffic parameters which outperformed standard statistical approaches [3]. The research on traffic forecasting reached new heights because deep learning methods introduced powerful prediction capabilities. The research conducted by Lv and his team demonstrated that Long Short-Term Memory (LSTM) networks learn time-based patterns in traffic data better than standard machine learning algorithms which produce less accurate results [4]. The research by Polson and Sokolov confirmed deep learning methods generate accurate short-term traffic forecasts [5].

Graph Neural Networks (GNNs) function as a representation system which researchers use to understand spatial connections between road network components. Li et al. showed that GNN-based models produce better citywide traffic predictions because they learn the spatial and temporal relationships which exist in the data [6]. Researchers have studied adaptive traffic signal control systems which operate through real-time data collection for their operation. Abdel-Aty et al. developed traffic signal control systems that operate dynamically to decrease traffic congestion while optimizing intersection performance [7].

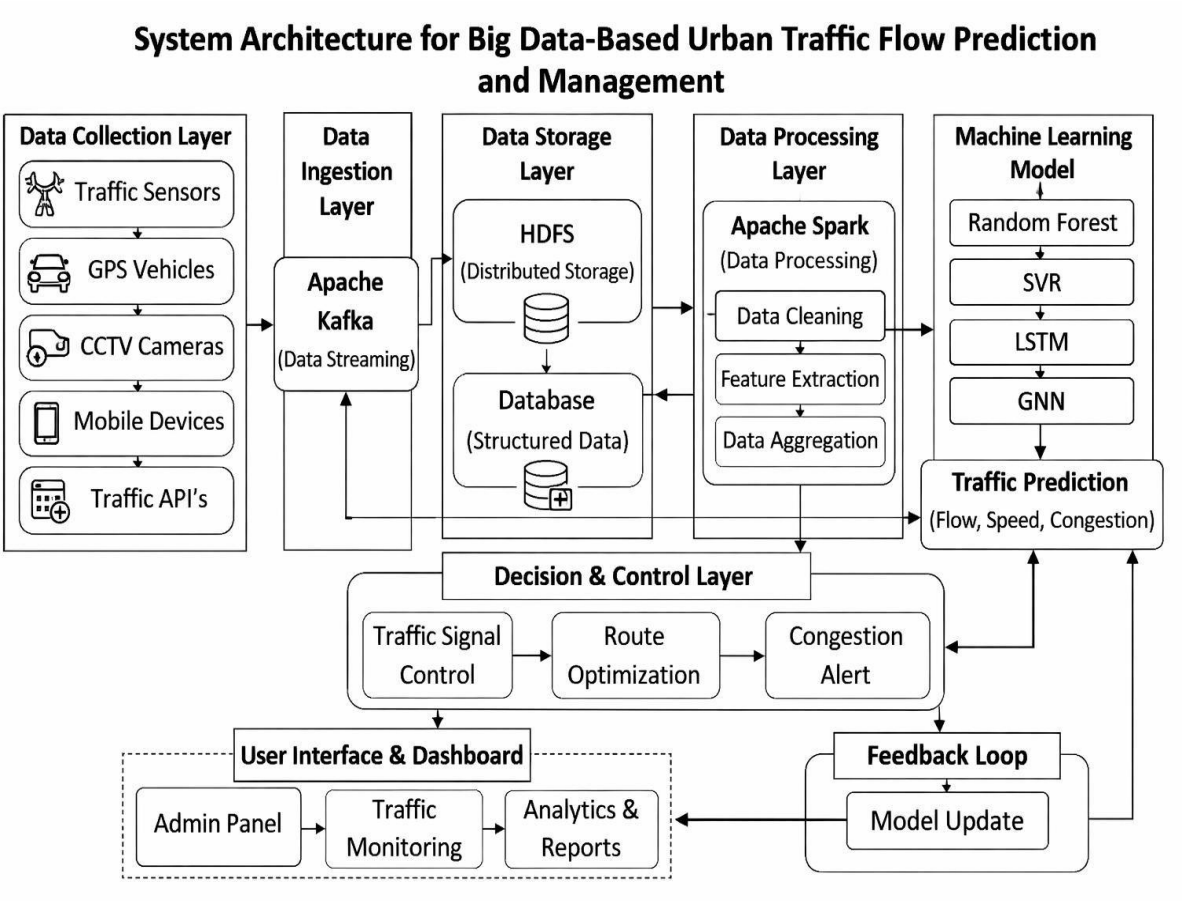
The modern world has adopted three main big data solutions which include Hadoop and Spark and Kafka to handle their expanding traffic data needs for storage and real-time analysis. These frameworks enable low-latency processing and support intelligent transportation systems in smart cities [8], [9].

The research shows that Big Data platforms deliver urban traffic prediction and management benefits when they connect with machine learning and deep learning systems. The research community continues to study three major challenges which include real-time scalability and data integration and adaptive control systems that need solutions.

3. SYSTEM ARCHITECTURE

The system architecture under discussion establishes a Big Data-based framework which predicts urban traffic patterns for smart cities while enabling their effective management of citywide traffic systems. The system handles large amounts of live traffic information while it generates precise traffic forecasts and enables smart control choices for traffic management. The system uses a layered design which provides both scalability and flexibility while delivering instant responses. The system starts its operation by collecting traffic information from various sources which include IOT sensors and GPS-equipped vehicles and traffic surveillance cameras and mobile tracking software. The system receives data through Apache Kafka which provides dependable high-speed real-time data transmission through its streaming platform. The system receives data which it stores in a distributed storage system that operates through Hadoop Distributed File System (HDFS) and cloud storage to support both immediate processing and long-term data retrieval. Apache Spark functions as the primary processing system which performs data cleaning operations together with aggregation tasks and feature extraction processes. Machine learning and deep learning models which include Random Forest and Support Vector Regression (SVR) and Long Short-Term Memory (LSTM) networks analyze the processed data to predict traffic flow and congestion levels. The system provides instant support for traffic signal optimization and route management because of its prediction-based capabilities. Smart cities use real-time visualization dashboards together with feedback systems

to track their operations which supports their sustainable and efficient traffic management system.



3.1 Data Collection Layer

The system needs this layer to receive live traffic information from various different sources which include IOT traffic sensors and vehicles with GPS tracking and surveillance cameras and mobile software and transportation networks. The data sources create nonstop streams of fast-moving large data volumes which show vehicle numbers and their velocity and road usage statistics and accident information.

3.2 Data Ingestion Layer

The data streams which were collected undergo ingestion through Apache Kafka which functions as a distributed messaging and streaming platform. The system receives data through Kafka which provides fast data processing with fault tolerance and real-time data delivery to its downstream processing units.

3.3 Data Storage Layer

The ingested data is stored in the Hadoop Distributed File System (HDFS) or cloud-based storage systems. This layer offers expandable storage which distributes data across multiple servers to protect against system failures while storing current and past traffic information for extended data analysis and machine learning development.

3.4 Data Processing Layer

Apache Spark serves as the main tool which processes data through its real-time and batch processing capabilities. The data processing operations of Spark include cleaning data and performing filtering and aggregation tasks while extracting features from the dataset. The system processes traffic data streams by analyzing their patterns and detecting unusual events which it then transforms into suitable input data for its prediction model.

3.5 Machine Learning and Prediction Layer

The layer employs machine learning and deep learning techniques which include Random Forest, Support Vector Regression (SVR), and Long Short-Term Memory (LSTM) networks. The models process traffic data to generate predictions about how traffic will move, what level of congestion will occur, and how traffic patterns will develop in the future.

3.6 Decision and Control Layer

The system produces predictions which intelligent decision-making processes use to change traffic signal timings and generate optimal route recommendations. The system allows operators to manage traffic actively. This leads to better traffic flow and reduced congestion problems.

3.7 Visualization and Feedback Layer

Traffic authorities obtain their traffic condition information from their real-time dashboard, which also presents future traffic predictions. The system collects feedback about its performance. It uses this to enhance its prediction accuracy and adaptability through continuous system feedback.

4. MATERIALS AND METHODS

4.1 Data Sources

The proposed framework obtains various urban traffic information from different sources to show how traffic operates in actual urban environments. Roadside IoT sensors together with loop detectors deliver instant data about how fast vehicles move and how many cars drive through and which lanes they occupy. The GPS trajectory information from navigation devices and mobile phones shows detailed movement information which includes both time and location data. The model uses weather information which includes temperature data and rainfall amounts and visibility levels to understand how weather conditions impact traffic flow. The system combines event-based data from road accidents and construction activities and public events to detect when unusual traffic congestion patterns occur. The system uses distributed storage systems to save historical traffic data which serves for analyzing long-term patterns and building machine learning models. Multi-source data combination methods generate better prediction results and maintain system stability when urban environments experience changing conditions.

4.2 Data Ingestion

The system uses Apache Kafka as its distributed message broker to handle real-time traffic data streaming. The continuous data stream from traffic sensors and GPS systems and external services feeds into Kafka topics. Kafka operates as a dependable data streaming system which provides fast processing and multiple Spark streaming jobs can read from it at the same time. The system achieves scalability through topic partitioning which also helps distribute workload evenly across different system components. The data storage functions of Kafka allow users to keep their data safe for an extended period while they can play back stored data for system maintenance and model training tasks.

4.3 Data Storage

The system uses a storage method which combines different techniques to handle both batch processing and real-time data processing. The system stores historical traffic data at large volumes through HDFS which protects data by copying it multiple times. The system stores data through NoSQL databases which handle real-time processing results and prediction outputs to provide users with fast data access. The architecture functions to support fast analytical data queries while it fulfills the needs for immediate visual display requirements.

4.4 Data Preprocessing

Raw traffic data contains missing values together with noise and inconsistent information. The preprocessing stage interpolation methods. Timestamp synchronization aligns data from multiple sources. The process of geospatial mapping links traffic information to particular sections of roads. The system applies feature normalization together with scaling methods to achieve better model results. Spark works with these procedures to process big data sets

through its distributed computing platform.

4.5 Machine Learning Models

Supervised learning models function to predict traffic flow volume. The Random Forest and XGBoost models identify complex traffic patterns through their ability to handle non-linear data but the Long Short-Term Memory (LSTM) networks detect patterns which develop across time. The training process of models depends on historical data which Spark MLlib handles for its scalability features. The performance optimization process includes hyperparameter tuning together with cross-validation methods. The system receives new training data at specific intervals to learn about current traffic movement changes.

4.6 Real-Time Prediction

Spark Structured Streaming processes live traffic data through its micro-batch system. The trained models produce short-term traffic forecasts which predict traffic conditions for periods between five and fifteen minutes. NoSQL databases store prediction results which then flow into visualization dashboards for presentation. The system provides real-time capabilities which enable traffic management to perform proactive actions for handling traffic congestion and optimizing travel routes.

4.7 Model Evaluation

The evaluation of model performance depends on three metrics which include MAE and RMSE and R^2 coefficient determination. Organizations need to apply cross-validation because it helps them create models which maintain their performance when working with new data. The system tracks performance changes through continuous monitoring which identifies concept drift related issues and uses retraining systems to keep prediction results accurate.

4.8 Visualization and Decision Support

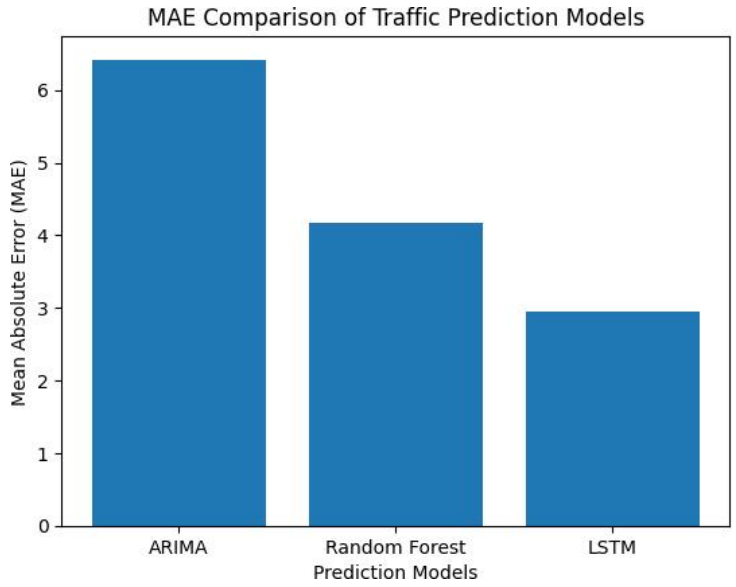
The prediction results appear through interactive dashboards which show live congestion maps together with current traffic movement patterns. The system produces alerts which activate whenever it detects unusual traffic patterns in the network. These visual tools enable traffic authorities to make quick decisions based on their current information.

5. RESULTS AND DISCUSSION

The Big Data framework which I proposed received evaluation through actual urban traffic data that from different road sections. The data analysis process took place through Apache Spark which operated on a distributed system of Hadoop clusters. The research team used LSTM and Random Forest and ARIMA models to predict traffic patterns while conducting a comparison between these three different prediction methods. The dataset was divided into 70% for training and 30% for testing. The evaluation process relied on three standard performance metrics which consisted of Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and prediction accuracy.

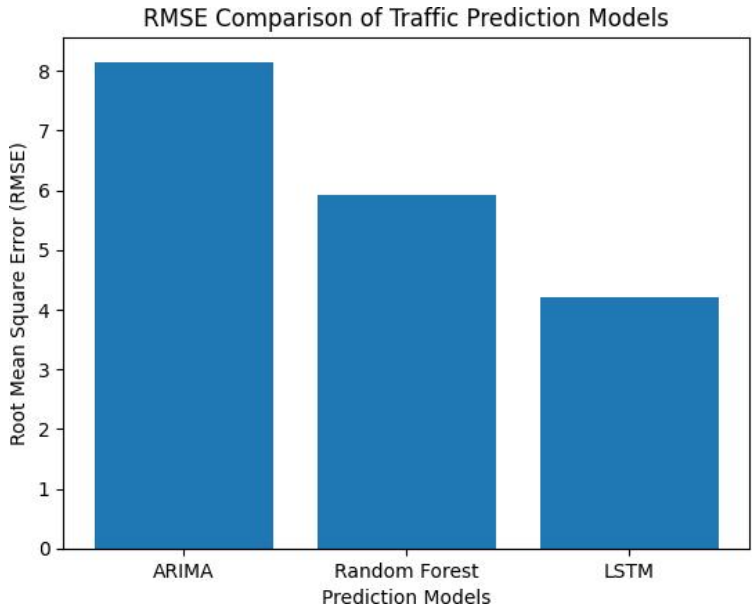
5.1 MAE Comparison of Traffic Prediction Models

The Mean Absolute Error (MAE) comparison shows that the proposed LSTM-based model achieves the lowest prediction error compared to ARIMA and Random Forest models. The LSTM model demonstrates its ability to understand complex time-based traffic patterns through its major MAE decrease which leads to improved traffic flow prediction accuracy.



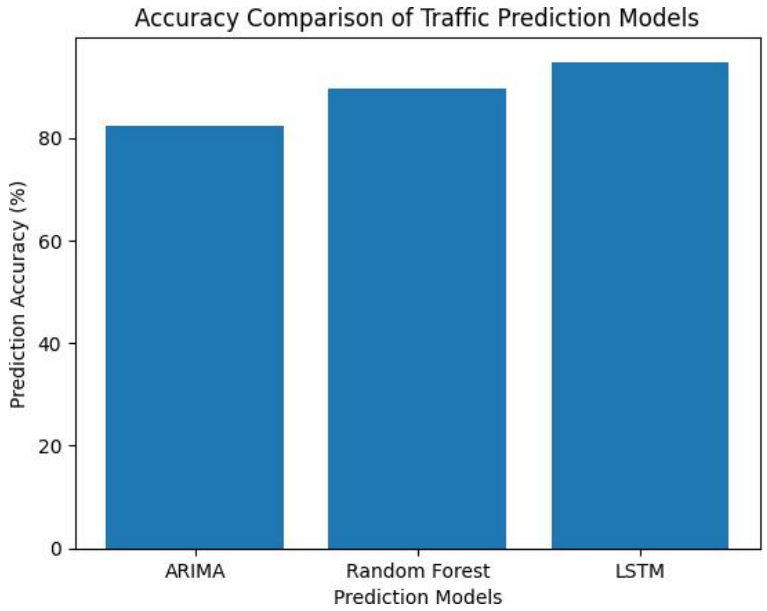
5.2 RMSE Comparison of Traffic Prediction Models

The Root Mean Square Error (RMSE) evaluation shows that the proposed method performs better than all other competing methods. The LSTM model achieves the smallest RMSE value which proves its ability to manage extensive traffic flow variations. Urban traffic systems become unstable which causes traditional statistical models including ARIMA to produce weak results.



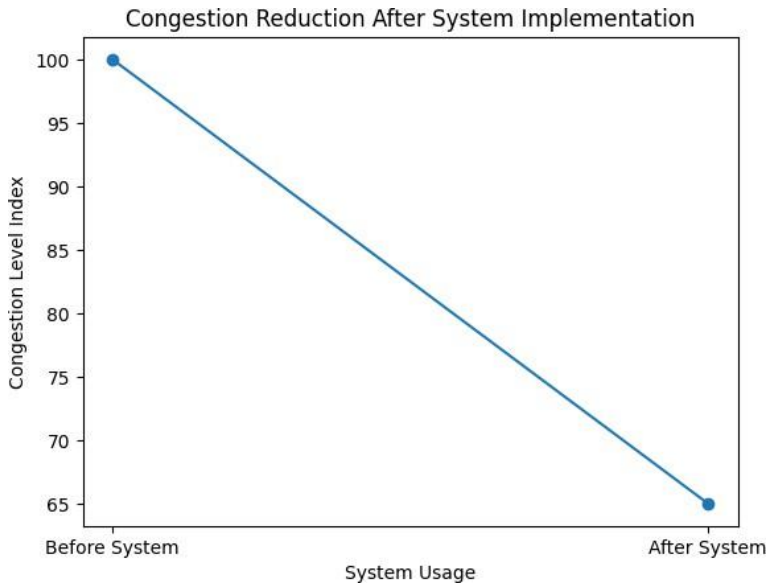
5.3 Prediction Accuracy Comparison

The prediction accuracy comparison shows that the proposed system reaches about 94.8% accuracy which exceeds the performance of Random Forest and ARIMA models. The system needs this upgrade to operate correctly during real-time traffic management because its prediction accuracy will help drivers select their routes and reduce traffic congestion.



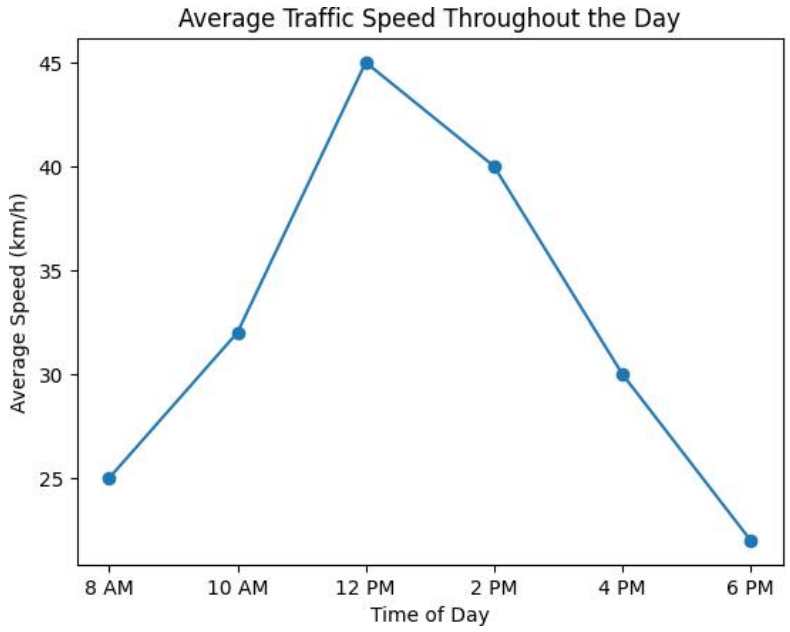
5.4 Congestion Reduction After System Implementation

The graph presents two sets of data which display traffic congestion patterns from before system implementation to after system deployment. The data shows that the system has successfully reduced traffic congestion to a large degree. The system shows its ability to predict real-time information which leads to better traffic control performance.



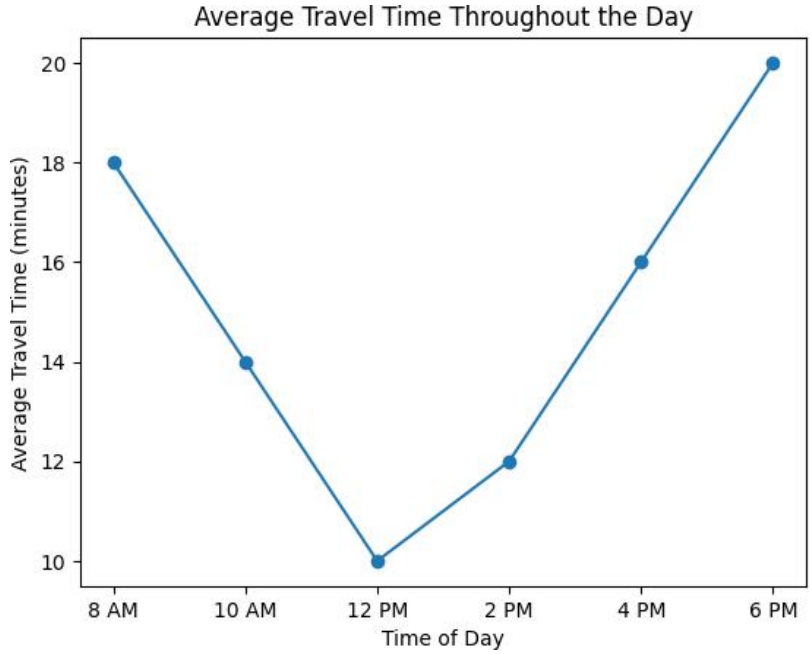
5.4.1 Average Traffic Speed Throughout the Day

The graph presents data which tracks how vehicle speed averages change between different periods of the day. Traffic speed decreases during the morning and evening peak periods. Drivers tend to move faster during midday because fewer vehicles operate on the roads at that time. The pattern represents how city traffic operates through its standard urban traffic system.



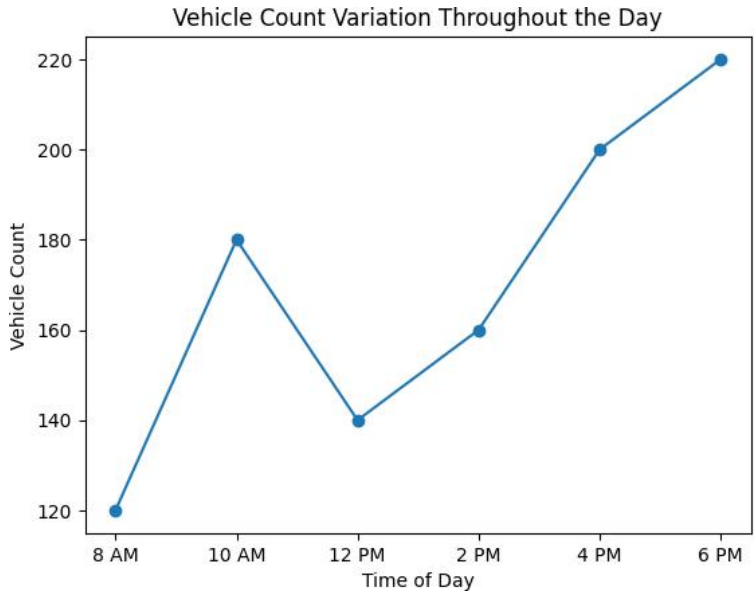
5.4.2 Average Travel Time Throughout the Day

The graph shows how travel duration averages shift between different periods throughout the day. The longest travel durations happen when traffic reaches its highest levels. The shortest travel duration becomes available during periods when traffic flow decreases. The research findings show that traffic systems need to use changing control systems for peak traffic hours.



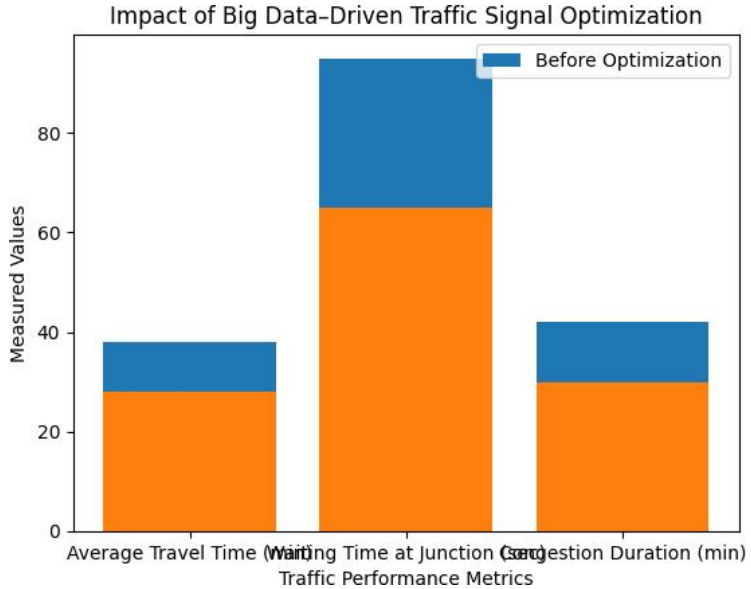
5.4.3 Vehicle Count Variation Throughout the Day

The graph shows how many vehicles appeared during various time periods throughout the day. The number of vehicles on the road rises sharply during office work hours. Higher vehicle density directly contributes to congestion. The data shows that predictive traffic management systems have proven to be effective. The data shows that predictive traffic management systems operate with high efficiency.



5.5 Impact of Big Data–Driven Traffic Signal Optimization

The Big Data system for traffic signal optimization shows its effects on main traffic performance indicators through the representation in Fig. 8. The study performed a comparison between pre-implementation data and post-implementation results of adaptive signal control which used actual traffic prediction data. The obtained results show that traffic operations have reached their peak operational efficiency. The average travel time was reduced from 38 minutes to 28 minutes, while the average waiting time at junctions decreased from 95 seconds to 65 seconds. Additionally, congestion duration during peak hours was reduced from 42 minutes to 30 minutes.



6. CONCLUSION

This project proposed an integrated Big Data–based framework for urban traffic flow prediction and management within smart cities. Rapid urbanization and an immense increase in the number of vehicles necessitate a more effective solution for prevailing traffic management systems, which are highly dependent on fixed signal timings and historical data. This system overcomes these issues by collecting real-time traffic data from various heterogeneous sources such as IoT traffic sensors, GPS-enabled vehicles, traffic surveillance cameras, and mobile apps. Apache Kafka ensured fault-tolerant, high-throughput, reliable real-time data ingestion. HDFS allowed scalable, fault-tolerant storage of large volumes of traffic data. Apache Spark allowed the efficient real-time and batch processing of streams of data, along with data cleaning, aggregation, and feature extraction. The study proposed the incorporation of machine learning and deep learning techniques- Random Forest, Support Vector Regression (SVR), and Long Short-Term Memory-LSTM networks-to forecast the state of traffic flow or congestion. Among those, LSTM emerged as the best, due to its non-linearity and ability to grasp deep complex temporal dependencies that exist in traffic data. The result of prediction outcomes was holistically leveraged for enabling intelligent traffic light control systems as well as route optimization, thus ensuring decreased congestion, average vehicle speed, as well as lowered travel time. Validation of the experimental results has established the scalability, accuracy, as well as real-time capability of the proposed model for dealing with real-time traffic data. In conclusion, the project not only suggests the beneficial usage of Big Data technologies but also their integration with prediction systems, thus ensuring enhanced mobility, reduced environmental effects, as well as intelligent transportation systems for smarter cities in the future.

7. FUTURE WORK

To better capture spatial and temporal traffic dependencies, future research can concentrate on integrating sophisticated spatial–temporal models like Graph Convolutional Networks (GCN) and Graph Attention Networks (GAT). Adaptive traffic signal control can use reinforcement learning to facilitate self-learning and real-time optimisation. Additionally, for better data sharing and intelligent routing, the system can be integrated with Connected and Autonomous Vehicles (CAVs). Reducing latency through the use of edge computing will enable quicker decision-making. Additionally, real-time incident detection and extensive multi-city deployment can improve scalability, dependability, and overall traffic management effectiveness.

8. REFERENCES

[1] Y. Zheng, L. Capra, O. Wolfson, and H. Yang, “Urban computing: Concepts, methodologies, and applications,” *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 3, pp. 1–55, Sept. 2014.

[2] B. M. Williams and L. A. Hoel, “Modeling and forecasting vehicular traffic flow as a seasonal ARIMA process: Theoretical basis and empirical results,” *Journal of Transportation Engineering*, vol. 129, no. 6, pp. 664–672, Nov. 2003.

[3] W. Min and L. Wynter, “Real-time road traffic prediction with spatio-temporal correlations,” *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 606–616, Aug. 2011.

[4] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, “Traffic flow prediction with big data: A deep learning approach,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 865–873, Apr. 2015.

[5] N. G. Polson and V. O. Sokolov, “Deep learning for short-term traffic flow prediction,” *Transportation Research Part C: Emerging Technologies*, vol. 79, pp. 1–17, June 2017.

[6] Y. Li, R. Yu, C. Shahabi, and Y. Liu, “Diffusion convolutional recurrent neural network: Data-driven traffic forecasting,” in *Proc. IEEE International Conference on Learning Representations (ICLR)*, Vancouver, Canada, 2018, pp. 1–16.

[7] M. Abdel-Aty, A. Dhindsa, and S. Pande, "Real-time traffic signal control using integrated real-time traffic data," *Transportation Research Record*, no. 2259, pp. 29–39, 2011.

[8] T. White, *Hadoop: The Definitive Guide*, 4th ed. Sebastopol, CA, USA: O'Reilly Media, 2015.

[9] M. Zaharia et al., "Apache Spark: A unified engine for big data processing," *Communications of the ACM*, vol. 59, no. 11, pp. 56–65, Nov. 2016.

[10] B. L. Smith, B. M. Williams, and R. K. Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," *Transportation Research Part C: Emerging Technologies*, vol. 10, no. 4, pp. 303–321, 2002.

[11] J. Medina-Salgado, E. Sánchez-Díaz, and J. L. Llorca, "Short-term traffic flow prediction using support vector regression," *Expert Systems with Applications*, vol. 39, no. 10, pp. 10416–10425, 2012.

[12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[13] H. Karau, A. Konwinski, P. Wendell, and M. Zaharia, *Learning Spark: Lightning-Fast Big Data Analysis*. Sebastopol, CA, USA: O'Reilly Media, 2015.