ISSN (O): 2320-5407
ISSN (P): 3107-4928

Journal Homepage: - www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH (IJAR)

Article DOI: 10.21474/IJAR01/22559
DOI URL: http://dx.doi.org/10.21474/IJAR01/22559

RESEARCH ARTICLE

SUPERVISED MODELS FOR ESTIMATING LINK-LEVEL TRAFFIC DENSITY USING TRAJECTORY DATA

Amadou Diabagate[1], Abdellah Azmani[2] and Adama Coulibaly[1]

1. Faculty of Mathematics and Computer Science, University Felix Houphouet-Boigny, Abidjan, Cote d'Ivoire.
2. Faculty of Sciences and Technologies, University Abdelmalek Essaadi, Tangier, Morocco.

Manuscript Info

Manuscript History
Received: 04 November 2025
Final Accepted: 06 December 2025
Published: January 2026

Key words:-
Urban congestion prediction, Supervised learning, Urban mobility, Trajectory-based mobility data, Abidjan (African cities).

Abstract

Traffic congestion is a growing concern in rapidly expanding cities, particularly in contexts where conventional traffic monitoring systems provide limited spatial and temporal coverage. This challenge is especially visible in many cities of the Global South, where the scarcity of fine-grained data restricts detailed analysis of urban mobility at the road-segment level. This study examines the prediction of link-level traffic density in Abidjan using trajectory data collected from an app-based mobility platform during September 2023, comprising more than 11,000 observations aggregated over approximately 814 road segments, and supervised machine learning methods.Road segments are described through a combination of geometric, regulatory, and trajectory-based features, and several regression models are evaluated within a common experimental framework.The results indicate that reliable traffic density estimates can be obtained even in the absence of dense sensing infrastructure. Random Forest provides consistently accurate and stable predictions across heterogeneous traffic conditions. The analysis also suggests that regulatory characteristics, such as speed limits and road hierarchy, exert a stronger influence on traffic density than detailed geometric descriptors.These findings highlight the practical relevance of trajectory-based supervised learning as a flexible and affordable solution for traffic analysis and mobility planning in data constrained urban environments.

Introduction:-

Traffic congestion has become a structural challenge in large metropolitan areas, particularly in cities undergoing rapid urban growth and increasing motorization [1]. Beyond longer travel times and higher fuel consumption, congestion has been widely shown to undermine economic productivity and degrade environmental sustainability in urban systems [2]. These effects are especially acute in cities of the Global South, where transport infrastructure development and traffic monitoring capacitiesoften remain limited relative torising mobility demand [3].In sub-Saharan African cities, traffic dynamics reflect a combination of strong demographic pressure, spatial expansion, and a highly heterogeneous transport supply [4]. In Abidjan, formal public transport services operate alongside

Corresponding Author:- Amadou Diabagate
Address:-Faculty of Mathematics and Computer Science, University Felix Houphouet-Boigny, Abidjan, Cote dIvoire.

informal modes, private vehicles, and app-based mobility platforms, resulting in pronounced spatial and temporal variations in congestion across the road network [1]. Although recent investments in major road infrastructure have improved connectivity on selected corridors, congestion remains a persistent daily constraint, largely due to the lack of continuous and fine-grained information on traffic conditions at the level of individual road segments [2,3]. Conventional traffic monitoring systems are primarily based on fixed sensing infrastructure, such as loop detectors, cameras, and dedicated counting stations [5]. While these technologies provide accurate measurements where they are deployed, their high installation and maintenance costs often limit spatial coverage, particularly in rapidly expanding urban environments [6]. As a consequence, large portions of road networks in cities like Abidjan remain insufficiently observed, restricting comprehensive congestion assessment and evidence-based traffic management [3].

In recent years, the rapid diffusion of digital mobility platforms has created new opportunities for traffic observation. Ride-hailing services continuously generate high-resolution trajectory data that capture vehicle movements across extensive parts of the urban network [7]. When properly anonymized and aggregated, these trajectory data have been shown to provide a reliable proxy for traffic conditions, enabling link-level analysis of congestion dynamics in complex and heterogeneous urban settings [8].The growing availability of trajectory-based data has coincided with significant advances in supervised machine learning for traffic analysis. Previous studies have demonstrated that nonlinear and ensemble-based models outperform classical linear approaches when modeling complex relationships between traffic density, road characteristics, and temporal demand variations [9]. In particular, machine learning techniques such as tree-based ensembles, kernel-based models, and Artificial Neural Network (ANN) are well suited to capturing the nonstationary and heterogeneous nature of urban traffic dynamics [10].

Building on these developments, this study investigates link-level traffic density prediction in Abidjan using trajectory data derived from an e-hailing platform. The objective is to evaluate the ability of supervised learning models to estimate traffic density at the scale of individual road segments in a data-constrained urban environment. Five regression approaches are examined: a Dummy Regressor used as a baseline, Linear Regression enhanced with Polynomial Ridge Regularization, Random Forest, Support Vector Regression (SVR), and Artificial Neural Network. All models are assessed within a unified experimental framework to ensure a consistent comparison of predictive performance and robustness.The remainder of this paper is organized as follows. Section 2 presents the urban mobility context of Abidjan and motivates the use of trajectory-based data. Section 3 reviews related work on traffic density estimation, trajectory-based traffic analysis, and supervised machine learning approaches. Section 4 describes the methodological framework, and the experimental protocol, including descriptive statistics, correlation analysis, and hyperparameter tuning. Section 5 reports the experimental results, with particular emphasis on error analysis and predicted–actual relationships. Section 6 discusses the implications and limitations of the findings. Finally, Section 7 concludes the paper and outlines perspectives for future research.

Urban Mobility Context in Abidjan:-

Abidjan is the economic capital of Cote d'Ivoire and one of the major metropolitan areas in West Africa. Over recent decades, the city has experienced sustained population growth and rapid spatial expansion, which have been accompanied by a steady increase in daily travel demand. This evolution has progressively intensified pressure on the urban road network. These trends are documented in recent empirical studies focusing on traffic data collection and network characterization in Abidjan, which describe the growing challenges associated with monitoring and managing mobility in a rapidly expanding urban environment [3,11].Urban mobility in Abidjan exhibits a high degree of modal diversity. Formal public transport systems coexist with a wide range of informal services, including shared minibuses and taxis, alongside private vehicles and, more recently, app-based ride-hailing platforms. This heterogeneous transport supply is reflected in traffic patterns that vary substantially across space and time, depending on land-use configuration, peak-hour demand, and network structure. Previous analyses of mobility transformations in Abidjan report that such diversity increases the complexity of traffic observation and modeling tasks, particularly at the level of individual road segments [11].

Traffic monitoring in Abidjan is characterized by a limited deployment of fixed sensing infrastructure. Conventional monitoring technologies such as loop detectors, cameras, and permanent counting stations are typically installed on selected parts of the network, leading to partial spatial coverage and discontinuous temporal information. Similar situations have been reported in many cities of the Global South. Empirical studies on traffic state estimation and monitoring indicate that these constraints limit comprehensive congestion assessment and restrict the operational use of data-driven traffic management strategies in large urban networks [12].In parallel with these limitations, the growing adoption of digital mobility platforms has generated new sources of traffic-related data. Ride-hailing services produce large volumes of high-resolution trajectory data that record vehicle movements across substantial portions of the urban network. Several recent studies show that trajectory-based data can provide a reliable proxy for traffic conditions, enabling the analysis of speed variations and congestion patterns at the link level, particularly in contexts where fixed sensors are sparse or unevenly distributed [13,14].Taken together, these empirical characteristics—rapid urban growth, heterogeneous mobility patterns, limited fixed sensing infrastructure, and increasing availability of trajectory data—make Abidjan a relevant case study for investigating alternative approaches to traffic density estimation. From an analytical perspective, data-constrained urban environments such as Abidjan offer a suitable setting for exploring trajectory-based methods combined with supervised machine learning, with the aim of supporting more comprehensive, scalable, and cost-effective link-level traffic analysis.

Related Work:-

Traffic density and congestion estimation constitute a long-standing research topic within intelligent transportation systems. Early studies primarily relied on data collected from fixed sensing infrastructure, including loop detectors, cameras, and permanent counting stations, to estimate traffic states and congestion levels. Such infrastructure has supported numerous operational traffic models and control strategies, particularly in cities equipped with dense monitoring networks [12]. At the same time, several studies point out that the deployment and maintenance of fixed sensors remain costly and often result in uneven spatial coverage, especially in rapidly expanding urban environments. As a consequence, these systems may fail to capture fine-grained congestion patterns at the level of individual road segments, limiting their effectiveness for detailed traffic analysis [15–17]. These limitations are further accentuated in complex urban networks characterized by heterogeneous demand and highly variable traffic conditions [18].

In response to these constraints, a growing body of literature has explored vehicle trajectory data as an alternative or complementary source of traffic information. The widespread availability of GPS-enabled devices, probe vehicles, and digital mobility platforms has substantially increased access to trajectory data for large-scale traffic analysis. Empirical studies show that trajectory-derived indicators, such as speed profiles and travel time distributions, can provide meaningful representations of congestion dynamics and support link-level traffic state inference in urban road networks [19,20]. More recent contributions further demonstrate that traffic density and congestion levels can be inferred from GPS-based trajectories even in contexts where direct measurements are unavailable or unreliable [13,18]. In particular, ride-hailing trajectory data have attracted growing attention due to their high temporal resolution and extensive spatial coverage, making them especially suitable for traffic analysis in cities with sparse or unevenly distributed sensing infrastructure [14]. Additional empirical investigations confirm that trajectory-based approaches are capable of capturing spatial heterogeneity and localized congestion patterns across large urban networks [21,22].

Beyond data source considerations, recent research has increasingly focused on modeling frameworks that explicitly account for the structure of road networks. Graph-based spatiotemporal models, including attention-driven temporal graph convolutional architectures, have been widely adopted to capture spatial dependencies and temporal dynamics in traffic forecasting tasks [23]. Subsequent developments emphasize the importance of jointly modeling local and global spatial interactions, with local–global spatiotemporal graph convolutional formulations proposed to better represent traffic flow dynamics across urban networks [24]. Complementary to these advances, several studies highlight the benefits of data fusion strategies, in which heterogeneous information sources are combined with machine learning models to improve robustness and generalization, particularly under sparse or noisy sensing conditions [25].

Alongside these methodological developments, supervised machine learning techniques have become central to traffic prediction and congestion analysis. Linear regression models are commonly employed as baseline approaches due to their simplicity and interpretability, often serving as reference points for more complex models [26]. However, numerous empirical studies report that linear models struggle to capture the nonlinear relationships inherent in traffic dynamics, especially under variable demand and complex network interactions [6,7]. Kernel-based methods such as Support Vector Regression have been shown to better represent nonlinear patterns, while ensemble approaches, including Random Forest, provide increased robustness to noise and heterogeneous feature distributions [27–29]. Artificial Neural Network have also been extensively applied to traffic forecasting tasks, with several studies demonstrating their ability to model complex temporal and spatial dependencies when sufficient data and appropriate regularization strategies are available [30].Recent survey studies underline the rapid expansion of deep learning and hybrid learning paradigms in traffic prediction, while emphasizing the critical role of data characteristics and evaluation protocols in determining comparative performance outcomes [31,32]. Comparative analyses consistently indicate that no single learning paradigm systematically outperforms others across all traffic prediction scenarios. Instead, model effectiveness is strongly dependent on data availability, feature design, and experimental settings, underscoring the importance of systematic and controlled model comparison within a unified evaluation framework [8,33]. Despite these advances, relatively few studies provide comprehensive comparisons of multiple supervised learning models for link-level traffic density estimation using real-world trajectory data in sub-Saharan African cities. Existing work in African urban contexts tends to focus on broader mobility challenges and data scarcity, with limited quantitative assessment of fine-grained traffic density models [3,11].

Against this background, the present study contributes to the literature by addressing these gaps through a systematic evaluation of multiple supervised regression models for link-level traffic density prediction in Abidjan, using trajectory data derived exclusively from an e-hailing platform. By comparing a baseline model with linear, ensemble-based, kernel-based, and Artificial Neural Networkapproaches within a consistent experimental framework, the study provides empirical insights into the relative suitability and robustness of different modeling strategies in a data-constrained urban environment. The proposed approach emphasizes practicality and scalability, offering a data-driven framework that can support traffic analysis and decision-making in cities where conventional monitoring infrastructure remains limited.

Methodology:-

This study relies on a supervised machine learning framework to estimate traffic density at the level of individual road segments in Abidjan.

The methodological pipeline integrates trajectory-based data processing, feature construction, model training, and performance evaluation, with a focus on robustness and reproducibility under realistic urban mobility conditions. An overview of the workflow is provided in Figure 1.

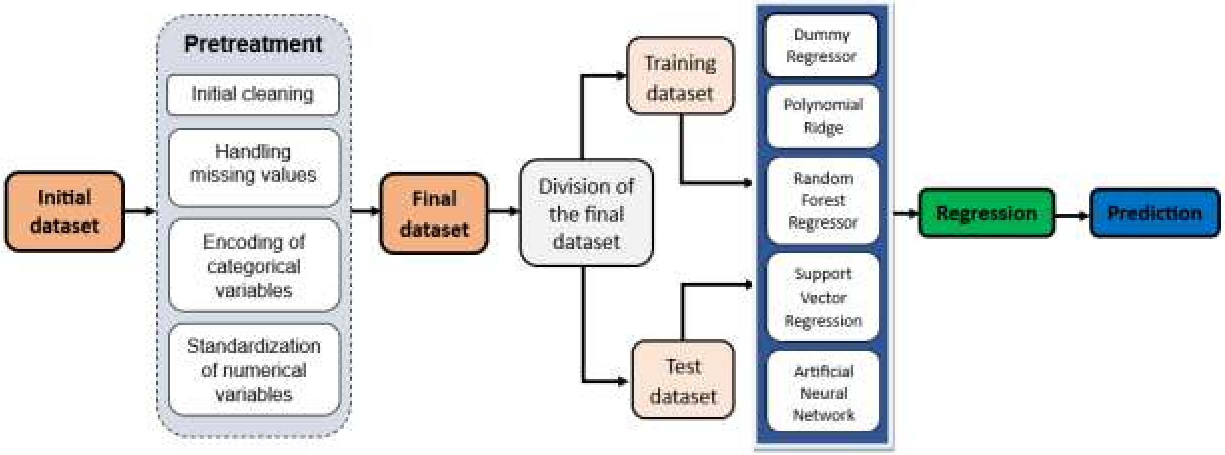

Figure 1: - Overview of the machine learning pipeline used

Dataset and Study Area:-

The empirical analysis is based on trajectory data collected from an e-hailing platform operating in Abidjan, Cote d’Ivoire. The dataset consists of anonymized vehicle trajectories describing completed trips within the metropolitan area. Each record contains spatiotemporal information that allows trajectories to be associated with individual road segments and aggregated at the link level. In total, the dataset comprises more than 11,000 individual trip records collected during September 2023, aggregated over 814 distinct road segments across the Abidjan metropolitan area. Trajectories were filtered to retain only completed trips with valid timestamps. Traffic density was estimated at the link level by aggregating trajectory-derived indicators over predefined temporal intervals, using the number of observed vehicle passages per segment as a proxy for density. This aggregation procedure ensures consistency between spatial units and supports reproducibility of the target variable construction.During preprocessing, only trips with consistent timestamps, valid GPS traces, and origins and destinations located within the study area were retained. Cancelled trips, incomplete trajectories, and corrupted records were systematically removed. All data were handled in aggregated and anonymized form to ensure compliance with privacy and ethical requirements. Although the present analysis focuses on Abidjan, the overall methodological framework is designed to remain applicable to other urban environments facing similar data constraints.

Feature Construction and Descriptive Statistics:

A set of explanatory variables was constructed to characterize traffic conditions at the road-segment level. These variables reflect complementary aspects of the urban network, including geometric properties of links, regulatory attributes such as posted speed limits, and trajectory-derived indicators capturing vehicle movement patterns.The selection of geometric variables was guided by their ability to represent structural characteristics of road segments that directly influence traffic accumulation and flow capacity, such as segment length, curvature, connectivity, and local network complexity. Regulatory attributes were included to account for differences in road hierarchy and operational constraints, while trajectory-derived indicators provide empirical information on observed vehicle behavior.

The target variable corresponds to traffic density estimated for each segment over predefined time intervals.This aggregation-based density estimation was adopted to ensure consistency across road segments and to produce a comparable link-level target variable compatible with the spatial resolution imposed by the available trajectory data. By estimating density at the segment level rather than at individual trajectory points, the approach aims to capture persistent traffic conditions while reducing the impact of short-term fluctuations and measurement noise.Descriptive statistics were computed to summarize the distributions of the explanatory variables and the target variable. Table 1 reports key summary measures, including indicators of central tendency, dispersion, and range. The results reveal pronounced heterogeneity across road segments. Several geometry-related variables exhibit strongly skewed distributions, with median values substantially lower than means, indicating the presence of a limited number of large or structurally complex segments alongside a majority of shorter and simpler links.

Table 1: - Summary of descriptive statistics for key variables

Variable	**Mean**	**Variance**	**Standard Deviation**	**Median**	**Mode**	**Range**	**Min**	**Max**
BBox Area (m²)	6085	330856000	18189,40	303,28	0	141191	0	141191
BBox Height (m)	71,26	8202,66	90,57	35,9	10,56	487,2	0,33	487,54
BBox Width (m)	43,6	4869,94	69,78	16,96	0	466,39	0	466,39
Mean Bearing (°)	164,04	13478,6	116,1	179,72	0	357,61	0	357,61
Chord (m)	93,05	11388,2	106,72	50,79	10,57	564,79	2,35	567,13
End Bearing (°)	164,07	13437,5	115,92	179,72	0	359,82	0	359,82
Length (m)	93,64	11734,2	108,32	51,06	10,57	587,54	2,35	589,89
Vertices	3,27	5,94	2,44	2	2	14	2	16
Seg. Max (m)	53,3	1884,03	43,41	36,51	10,57	179,32	2,35	181,66
Seg. Mean (m)	44,62	1360,6	36,89	29,82	10,57	174,01	2,35	176,36
Seg. SD (m)	5,58	119,16	10,92	0	0	45,22	0	45,22
Sinuosity	1	0	0,01	1	1	0,05	1	1,05
Start Bearing (°)	164,34	13609,4	116,66	180	0	359,37	0	359,37

Straightness	1	0	0,01	1	1	0,04	0,96	1
Max Turn (°)	4,08	5,78	2,4	4,08	4,08	14,96	0	14,96
Mean Turn (°)	2,83	3,24	1,8	2,83	2,83	13,12	0	13,12
Turn p90 (°)	3,7	4,53	2,13	3,7	3,7	14,22	0	14,22
Vertex Density (m^{-1})	0,04	0	0,05	0,03	0,03	0,42	0,01	0,43
Length	93,33	11653,9	107,95	50,7	21,2	585,2	2,4	587,6
Segments	2,27	5,94	2,44	1	1	14	1	15
Speed Limit	66,11	624,79	25	60	50	70	50	120

Regulatory attributes also display considerable variability across the network, reflecting differences in road function and hierarchy. Taken together, these descriptive patterns underline the structural diversity of Abidjan's road network and motivate the use of flexible regression models capable of capturing nonlinear relationships.

Correlation Analysis:

To explore relationships among explanatory variables and assess potential redundancy, a correlation matrix was computed using Pearson correlation coefficients. The resulting heatmap is presented in Figure 2. The analysis highlights distinct correlation structures associated with different feature groups.Regulatory attributes form a coherent cluster, while geometric descriptors related to segment size and extent are strongly correlated with one another. In contrast, indicators of local structural complexity and orientation display weaker associations with size-related variables. Importantly, correlations between regulatory and geometric feature groups remain moderate, suggesting limited redundancy across these dimensions.Overall, the correlation patterns indicate that the selected features provide complementary information rather than duplicating the same signal. On this basis, all constructed variables were retained for the supervised learning experiments.

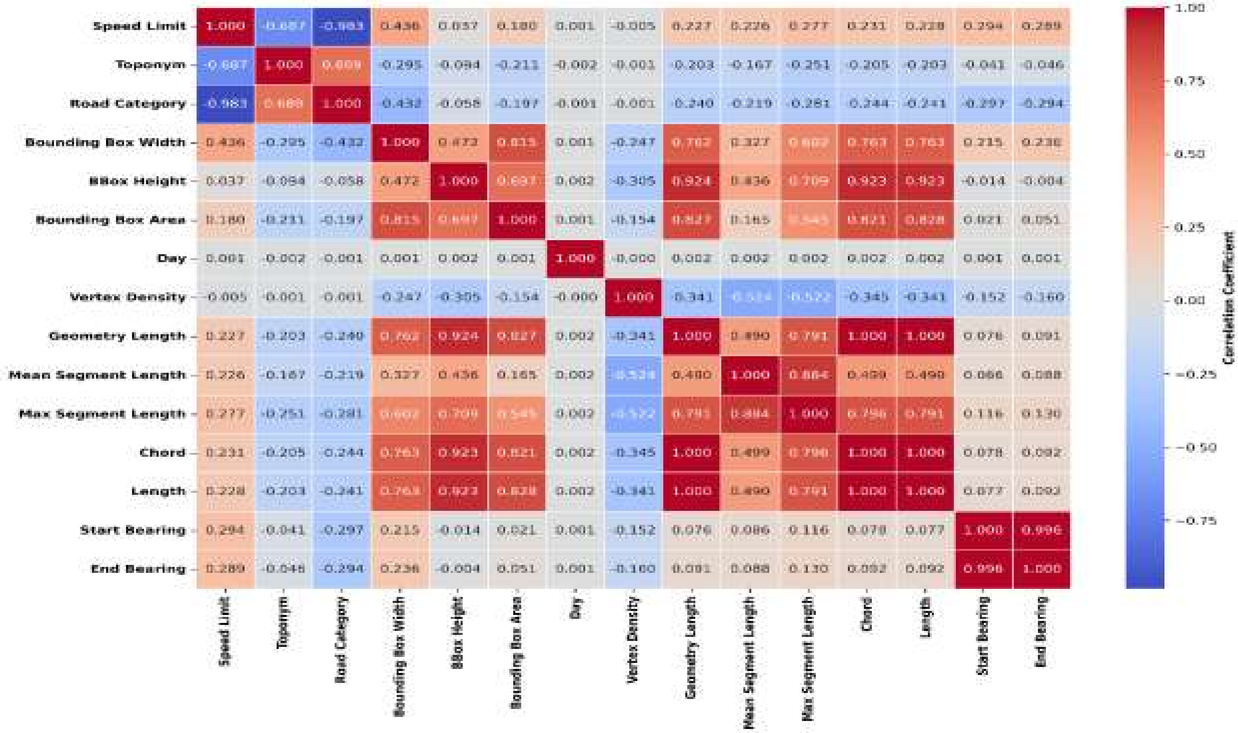

Figure 2: - Heatmap of the Correlation Between the Top 15 Predictive Features

Supervised Learning Models:

Rather than relying on a single predictive approach, this study compares several regression models with distinct assumptions and levels of flexibility, selected to cover the main families of supervised learning approaches commonly applied in traffic prediction. These range from simple baseline and linear models to ensemble-based, kernel-based, and neural methods, thereby ensuring a balanced and methodologically sound comparison [34, 35].

To capture nonlinear relationships while maintaining model stability, Linear Regression with Polynomial Ridge regularization was retained. Polynomial feature expansion allows interaction effects to be modeled, while L2 regularization helps control estimation variance in the presence of correlated predictors [26].An ensemble-based approach is represented by the Random Forest Regressor, which aggregates multiple decision trees trained on randomized subsets of the data. This method is well suited to heterogeneous feature spaces and complex nonlinear dependencies [29].Support Vector Regression (SVR) was also considered, as it uses kernel-based transformations to approximate nonlinear relationships through margin-based optimization, often yielding strong generalization performance on structured datasets [36, 37].Finally, an Artificial Neural Network (ANN) was employed to learn nonlinear interactions across multiple explanatory variables. The network relies on layered representations optimized through gradient-based learning and is capable of capturing complex feature interactions [30].A Dummy Regressor serves as a baseline, providing a reference level of performance against which more advanced models can be evaluated [38].

Hyperparameter Tuning and Experimental Protocol:

To ensure a fair comparison across models, hyperparameter tuning was conducted using a randomized search strategy. This approach enables efficient exploration of the hyperparameter space while limiting computational cost. The Dummy Regressor was excluded from this tuning procedure and evaluated using baseline strategies.

Model performance was assessed through a K-fold cross-validation scheme in order to obtain stable estimates of predictive accuracy and generalization. Evaluation relied on standard regression metrics, including the coefficient of determination (R^2) and error-based measures such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The optimized hyperparameter configurations retained for each model are summarized in Table 2.

Table 2: - Summary of Machine Learning Models and their Optimized Hyperparameter Settings

Model	Best Hyperparameters
Dummy Regressor	{'strategy': 'mean'}
Polynomial+Ridge	{'poly__degree': 2, 'poly__include_bias': False, 'poly__interaction_only': False, 'ridge__alpha': 0.1, 'ridge__fit_intercept': True}
Random Forest	{'max_depth': None, 'min_samples_split': 2, 'n_estimators': 200}
Support Vector Regression	{'C': 10, 'gamma': 'scale', 'kernel': 'rbf'}
Artificial Neural Network	{'activation': 'relu', 'alpha': 0.01, 'hidden_layer_sizes': (50, 30)}

All remaining model parameters were kept at their default values as defined in the scikit-learn implementation.

Experimental Results and Analysis: -

This section reports the experimental results obtained from the supervised learning models used to estimate link-level traffic density in Abidjan. The analysis builds exclusively on results already produced in the complete study and focuses on global performance, error behavior, and calibration quality. No explainable AI techniques are considered at this stage, and the discussion is deliberately limited to empirical observations.

Global Model Performance:

The first level of analysis compares the overall predictive performance of the models using standard regression metrics. Cross-validated values of R^2, RMSE, MAE, MSE, and MAPE are summarized in Table 3, providing a consistent basis for comparison across models.As expected, the Dummy Regressor performs poorly across all metrics, yielding a coefficient of determination close to zero and very large errors, with an RMSE exceeding 13. It therefore serves only as a baseline reference. In contrast, Polynomial Ridge Regression represents a clear improvement, reaching an R² of about 0.96 and reducing the RMSE to roughly 2.7. This gain suggests that the inclusion of nonlinear terms captures a significant part of the structure underlying traffic density variation across road segments.

Table 3: - Cross-validation performance of the regression models for traffic density prediction

Models	R2	RMSE	MAE	MSE	MAPE
Dummy Regressor	-0.000045	13.116091	10.470514	172.031836	9.184379
Polynomial + Ridge	0.961704	2.700657	2.023422	7.293547	1.842884
Random Forest	0.990780	1.472332	1.058527	2.167762	0.969000
Support Vector Regression	0.907712	4.273489	3.179472	18.262707	2.965131
Artificial Neural Network	0.953348	3.396052	2.569213	11.533171	2.312873

Among the remaining models, Random Forest stands out as the best-performing approach. It achieves the highest explanatory power, with an R² close to 0.99, while maintaining low prediction errors (RMSE ≈ 1.47 and MAE ≈ 1.06). Support Vector Regression and the Artificial Neural Network also outperform the linear baseline, with coefficients of determination above 0.90, but they are associated with larger residual errors and greater variability across cross-validation folds. These differences in predictive behavior are further illustrated in Figure 3, which highlights the progressive improvement obtained when moving from simpler to more flexible learning models.

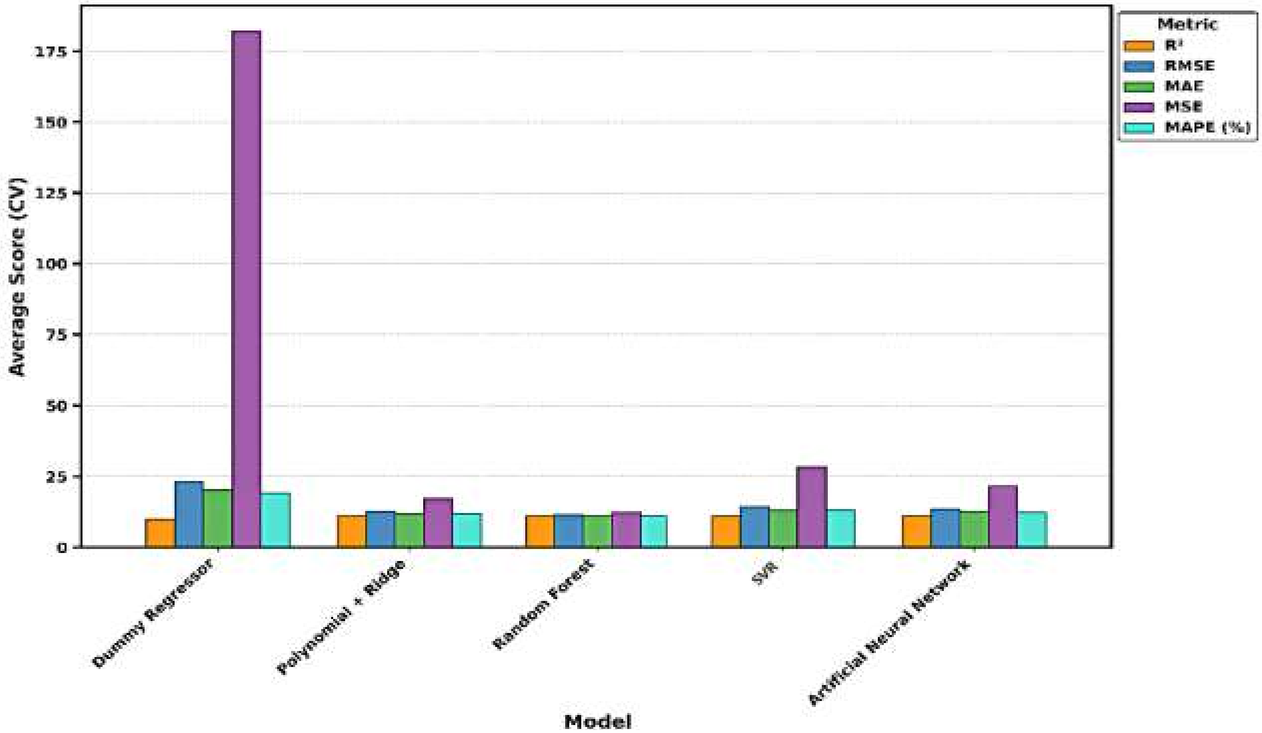

Figure 3: - Comparative Barplot of Model Performance by Metric Using Cross-Validation

Random Forest emerges as the most effective model across the evaluated performance metrics. Nevertheless, these results are best interpreted within the context of the adopted experimental framework. K-fold cross-validation provides evidence of performance stability across data splits, while the reported outcomes remain specific to the dataset characteristics and feature set used in this analysis.Taken together, the global metrics reveal a clear hierarchy among the evaluated approaches, with ensemble-based methods providing the most accurate and reliable estimates of traffic density at the link level.

Error Analysis and Diagnostic Plots:

While aggregate metrics provide a first indication of model performance, residual analysis offers deeper insight into stability and robustness. The distributions of prediction errors obtained under cross-validation are shown in Figure 4.

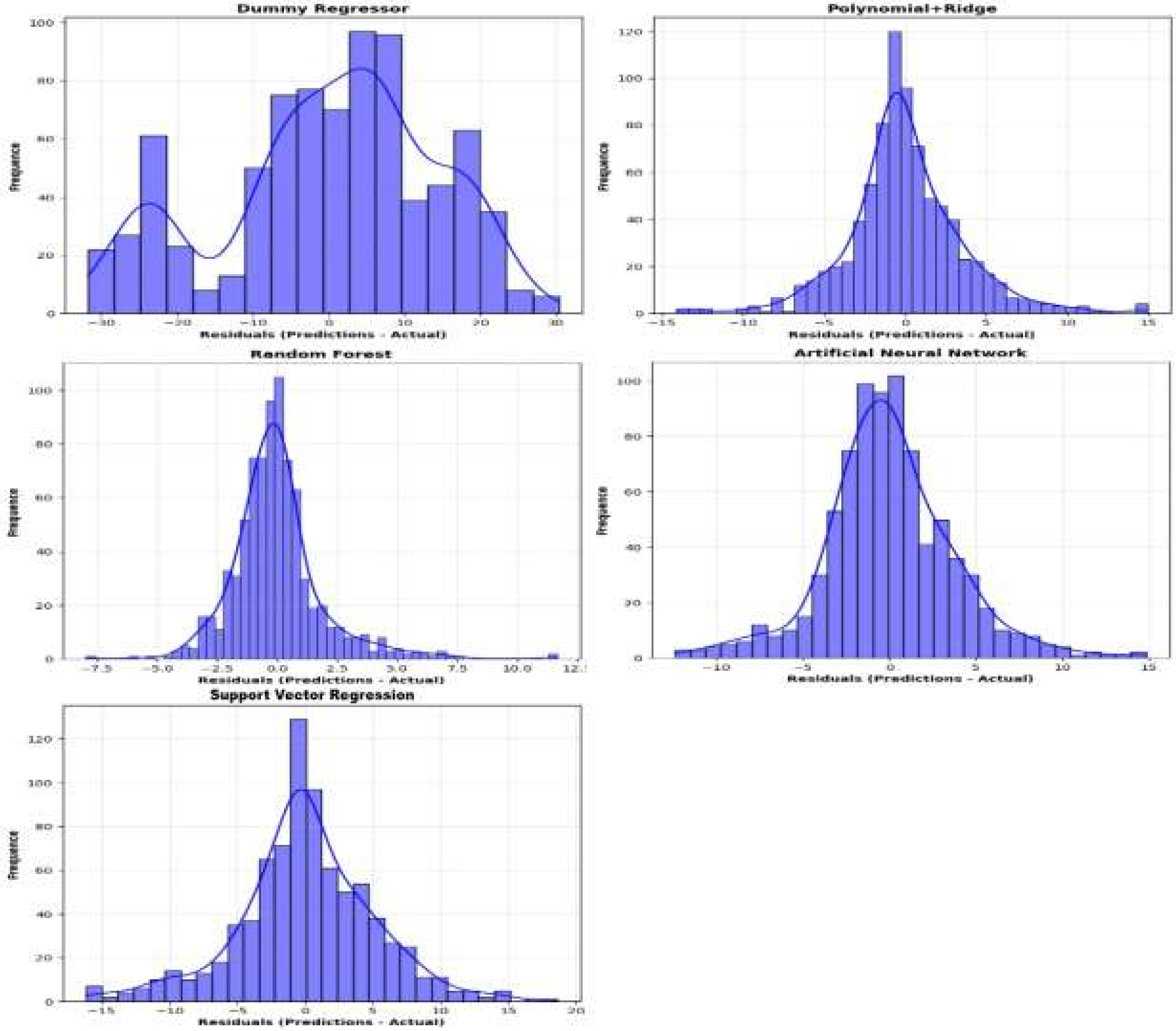

Figure 4: - Residual distributions (Predictions – Actual) for all regression models under cross-validation

The Dummy Regressor produces wide and unstructured residual distributions, confirming its inability to capture meaningful variation in traffic density. Polynomial Ridge Regression yields residuals that remain centered around zero but exhibit heavier tails, suggesting reduced accuracy for extreme density values.Random Forest displays the most balanced residual behavior. Its error distribution is narrow, approximately symmetric, and closely centered on zero, indicating both low variance and limited systematic bias across different traffic regimes. Support Vector Regression and the Artificial Neural Network also generate centered residuals, although with broader dispersion, reflecting higher sensitivity to local fluctuations and model configuration.Overall, the diagnostic plots confirm that ensemble-based models not only achieve higher accuracy but also provide more stable and consistent error behavior, a desirable property for link-level traffic density estimation in heterogeneous urban networks.

Predicted–Actual Relationship Analysis:

Model calibration was further examined by comparing predicted and observed traffic density values. Scatter plots of predicted versus actual densities are presented in Figure 5, with the identity line included as a reference.

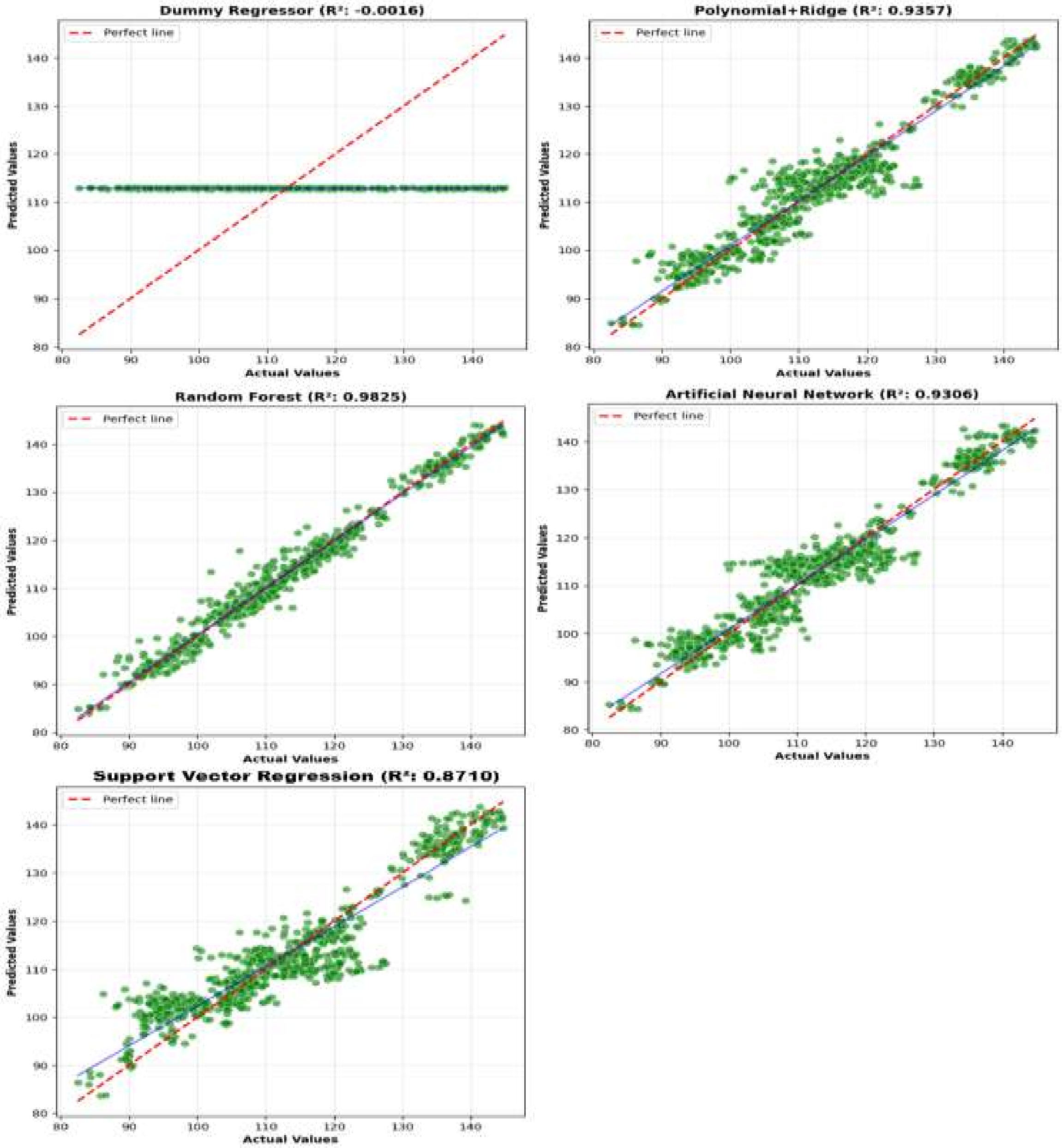

Figure 5: -Predicted vs. Actual traffic density plots for the regression models under cross-validation

Polynomial Ridge Regression shows a reasonable alignment with the diagonal but tends to smooth high-density observations, resulting in mild underestimation at the upper end of the range. Random Forest exhibits the strongest agreement with observed values, with predictions tightly clustered around the identity line across both low- and high-density conditions.Support Vector Regression and the Artificial Neural Network capture the overall trend but display greater dispersion around the diagonal, indicating increased variability in predictions. As expected, the Dummy Regressor shows no meaningful alignment with observed densities.These visual patterns are consistent with the numerical results and residual diagnostics. Together, they indicate that Random Forest provides the most accurate and well-calibrated representation of link-level traffic density among the models considered.

Global Feature Importance:

To complement the performance analysis with a global view of variable influence, feature importance was examined using the Random Forest model. The relative importance of the most influential predictors is shown in Figure 6.

Regulatory and contextual variables dominate the ranking. Speed Limit emerges as the most influential feature, followed by indicators related to road category and spatial context. Geometric descriptors contribute more moderately, while fine-grained orientation and segmentation variables appear at the lower end of the ranking.This global importance analysis confirms that traffic density patterns in Abidjan are driven primarily by regulatory context and network hierarchy, with geometric characteristics providing secondary refinement. The analysis remains strictly global and does not rely on local explainability techniques.

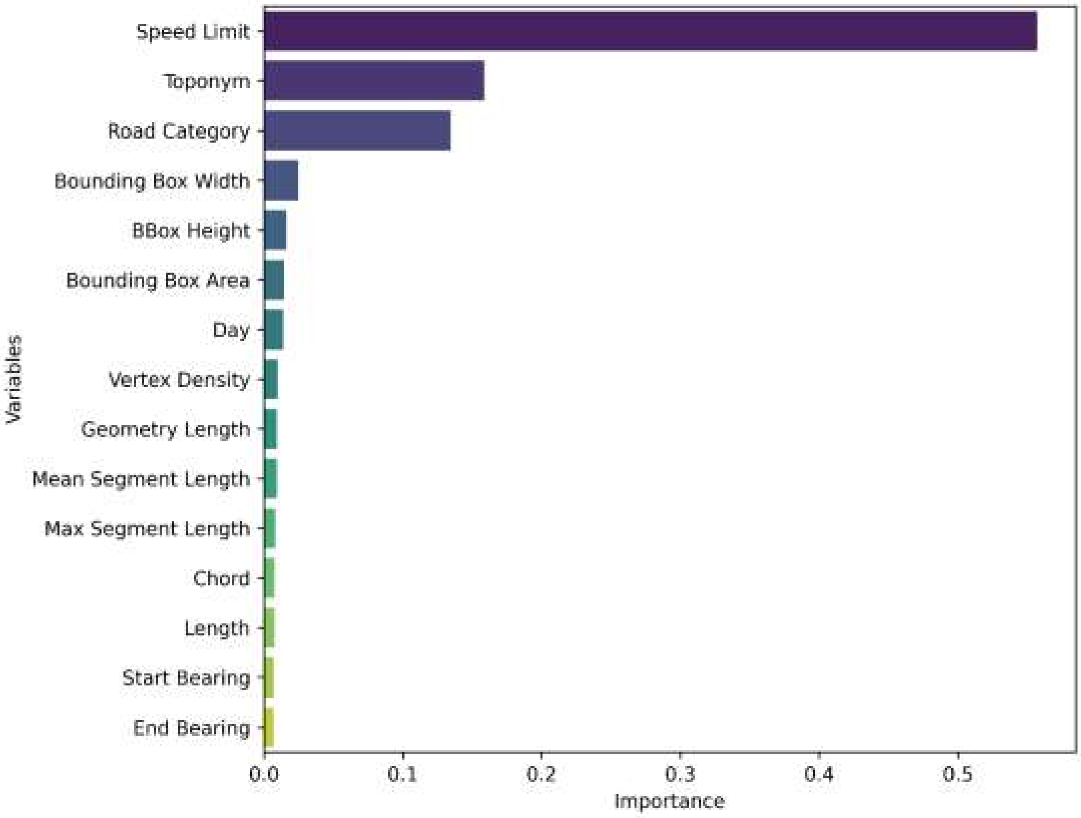

Figure 6: -Top 15 most important features according to the Random Forest model

Discussion:-

This study shows that supervised machine learning can provide a reliable and effective framework for estimating link-level traffic density in Abidjan in situations where conventional traffic sensing infrastructure remains sparse or unevenly deployed. By combining trajectory data from an e-hailing platform with geometric and regulatory descriptors, the proposed approach captures structural congestion patterns that are difficult to observe through traditional monitoring systems alone.

Why Some Models Perform Better Than Others:

The results indicate that Random Forest tends to outperform linear, kernel-based, and Artificial Neural Network approaches in the considered setting. This advantage can largely be attributed to the ability of tree-based ensemble methods to model complex and nonlinear interactions between heterogeneous predictors, including road geometry, regulatory constraints, and spatial context. Such interactions are particularly relevant at the link level, where traffic density is strongly shaped by structural characteristics of the road network rather than by purely temporal dynamics. Similar observations have been reported in recent comparative studies on traffic flow and congestion prediction, in which ensemble-based methods consistently demonstrate strong robustness in heterogeneous urban environments [4,5,28].Polynomial regression achieves reasonable performance but remains limited in its capacity to capture higher-order interactions across diverse road segments. Kernel-based methods and Artificial Neural Networks also provide acceptable levels of accuracy; however, their performance appears more sensitive to feature scaling, hyperparameter configuration, and data distribution. This sensitivity may reduce their stability in operational contexts where calibration data are limited or unevenly distributed [39,40].

Consistency with the Existing Literature:

The observed dominance of ensemble-based models is well aligned with recent findings in the traffic prediction literature. Several reviews of machine learning applications in intelligent transportation systems emphasize that tree-based ensembles offer a favorable balance between predictive accuracy, robustness to noise, and computational efficiency, particularly in contexts characterized by uneven data quality and coverage [6,30,41]. Empirical studies conducted in African and North African cities report similar trends, highlighting the suitability of these models for congestion estimation using trajectory-based data [4,5].In addition, the strong influence of regulatory variables, such as speed limits and road hierarchy, is consistent with prior work showing that contextual and functional attributes often play a more decisive role than fine-grained geometric descriptors when explaining congestion patterns at the scale of urban road networks [9,42]. This finding reinforces the importance of integrating regulatory information when modeling traffic density in rapidly urbanizing cities.

Relevance for Data-Constrained Cities:

From an applied perspective, these results underline the practical value of trajectory-driven learning frameworks for cities with limited fixed sensing infrastructure. In Abidjan, as in many cities of the Global South, the uneven deployment of traffic sensors restricts the ability to monitor congestion comprehensively across the network. Trajectory data generated by e-hailing services therefore represent a valuable alternative source of high-resolution information that can support network-wide traffic analysis at relatively low cost [13,14].By focusing on supervised learning models rather than complex spatiotemporal architectures, the proposed approach remains computationally tractable and adaptable to other urban contexts facing similar data constraints. This makes it particularly relevant for transport authorities seeking scalable tools to support congestion diagnosis and mobility planning in environments where data availability remains heterogeneous [15,43].While the empirical results reported in this study are grounded in the specific urban context of Abidjan, the proposed methodological framework is not inherently city-specific. The combination of trajectory-derived data, link-level feature construction, and supervised learning models can be applied to other urban environments facing similar constraints in terms of limited fixed sensing infrastructure and heterogeneous traffic conditions. In such contexts, local calibration of features and aggregation schemes would be required to reflect network characteristics and mobility patterns, while preserving the overall modeling approach.

Limitations:-

Some limitations nevertheless deserve to be acknowledged. First, the analysis relies on trajectory data from a single mobility platform, which may introduce spatial sampling bias toward high-demand corridors and central areas. As a result, peripheral neighborhoods with lower e-hailing activity may be underrepresented, potentially affecting prediction accuracy in these zones [44,12]. Second, the temporal scope of the dataset is limited, which restricts the ability to capture long-term seasonal effects or atypical congestion patterns associated with special events or disruptions. Finally, although the selected features capture key structural and regulatory drivers of traffic density, other potentially relevant factors, such as land-use intensity or weather conditions, were not explicitly modeled and may explain part of the residual variability observed in the predictions [16,45].Despite these limitations, the consistency of the results with prior studies and the stability of the best-performing models suggest that the proposed framework constitutes a robust and relevant foundation for link-level traffic density estimation in data-constrained urban environments.

Conclusion:-

This paper investigated the problem of link-level traffic density prediction in Abidjan using trajectory data derived from an e-hailing platform and supervised machine learning models. The study was motivated by the persistent lack of fine-grained traffic monitoring infrastructure in many rapidly growing cities of the Global South, where conventional sensing systems provide only partial and uneven coverage of urban road networks.By comparing a baseline model with linear, ensemble-based, kernel-based, and Artificial Neural Network within a unified experimental framework, the results demonstrate that supervised learning can effectively capture traffic density patterns at the scale of individual road segments. Among the evaluated approaches, ensemble-based methods, and in particular Random Forest, consistently provide the most accurate and stable predictions across global performance metrics, residual diagnostics, and calibration analyses. These findings highlight the importance of modeling nonlinear interactions between regulatory context, road hierarchy, and trajectory-derived indicators when addressing heterogeneous urban traffic conditions.Beyond predictive accuracy, the analysis shows that regulatory and contextual variables play a dominant role in shaping traffic density patterns in Abidjan, while detailed geometric descriptors contribute more moderately once higher-level structural information is taken into account. This observation reinforces the relevance of incorporating regulatory and functional attributes in data-driven traffic models, especially in urban environments characterized by mixed transport systems and uneven infrastructure development.From an applied perspective, the proposed framework illustrates the practical value of trajectory-based data for traffic analysis in data-constrained contexts. By relying on widely available mobility data and supervised learning models that remain computationally tractable, the approach offers a scalable alternative to sensor-dependent monitoring systems. It can support network-wide congestion assessment and provide quantitative insights that are difficult to obtain through traditional data sources alone.

From an operational perspective, the proposed framework offers practical insights for transport authorities and urban mobility planners operating in data-constrained environments. By leveraging trajectory data from existing digital mobility services, the approach enables the estimation of link-level traffic density without relying on extensive fixed sensing infrastructure. Such information can support congestion diagnosis, network performance assessment, and the prioritization of targeted interventions, while remaining compatible with the data availability and resource constraints commonly faced by cities in the Global South.Several limitations nonetheless remain. The reliance on a single mobility data provider may introduce spatial and behavioral biases, and the indirect estimation of traffic density from trajectories cannot fully replace ground-truth measurements. In addition, the analysis focuses on global predictive behavior and does not explicitly address temporal dynamics or localized congestion phenomena.Despite these constraints, the study provides a solid empirical foundation for the use of supervised learning and trajectory data in link-level traffic density estimation in rapidly urbanizing cities. Future work may extend this framework by integrating additional data sources, exploring temporal modeling strategies, or applying the approach to other urban contexts facing similar monitoring challenges. Taken together, the results contribute to ongoing efforts to develop data-driven, scalable, and context-aware tools for urban traffic analysis and mobility planning.

References:-

[1] J. Doherty, “Mobilizing social reproduction: Gendered mobility and everyday infrastructure in Abidjan,” Mobilities, vol. 16, no. 5, pp. 758–774, 2021, doi: 10.1080/17450101.2021.1944288.

[2] G. Falchetta, M. Noussan, and A. T. Hammad, “Comparing paratransit in seven major African cities: An accessibility and network analysis,” Journal of Transport Geography, vol. 94, p. 103131, 2021, doi: 10.1016/j.jtrangeo.2021.103131.

[3] G. Sylla, P. Apparicio, and A. N. (coauthors), “Mapping road traffic noise descriptors in a sub-Saharan city: An extensive mobile data collection in Abidjan (Ivory Coast),” African Transport Studies, vol. 3, p. 100067, 2025, doi: 10.1016/j.aftran.2025.100067.

[4] U. U. Imoh and M. Movahedi Rad, “Analysis and prediction of traffic conditions using machine learning models on Ikorodu Road in Lagos State, Nigeria,” Infrastructures, vol. 10, no. 5, p. 122, 2025, doi:10.3390/infrastructures10050122.

[5] L. Hammoumi et al., “Leveraging machine learning to predict traffic jams: Case study of Casablanca, Morocco,” J. Urban Manag., 2025, doi: 10.1016/j.jum.2025.02.004.

[6] P. Qi, C. Pan, X. Xu, J. Wang, J. Liang, and W. Zhou, “A review of dynamic traffic flow prediction methods for global energy-efficient route planning,” Sensors, vol. 25, no. 17, p. 5560, 2025, doi:10.3390/s25175560.

[7] K. N. Lam, “Traffic prediction using LSTM, RF and XGBoost,” in Proc. 2nd Int. Conf. Data Analysis and Machine Learning (DAML), 2024, vol. 1, pp. 267–274, doi:10.5220/0013515600004619.

[8] N. A. M. Razali, N. Shamsaimon, K. K. Ishak et al., “Gap, techniques and evaluation: Traffic flow prediction using machine learning and deep learning,” J. Big Data, vol. 8, p. 152, 2021, doi:10.1186/s40537-021-00542-7.

[9] K. Hamad, E. Alotaibi, W. Zeiada, G. Al-Khateeb, S. Abu Dabous, M. Omar, B. R. K. Mantha, M. G. Arab, and T. Merabtene, “Explainable artificial intelligence visions on incident duration using eXtreme Gradient Boosting and SHapley Additive exPlanations,” Multimodal Transportation, vol. 4, no. 2, p. 100209, 2025, doi: 10.1016/j.multra.2025.100209.

[10] B. Lv, H. Gong, B. Dong; Z. Wang, H. Guo, J. Wang, and J. Wu, "An Explainable XGBoost Model for International Roughness Index Prediction and Key Factor Identification," Applied Sciences, vol. 15, no. 4, p. 1893, 2025. doi: 10.3390/app15041893.

[11] G. Spire, A. Steck, and S. M. Koffi, “La modernisation urbaine depuis la portière d’un minibus à Yopougon (Abidjan). Les effets du nouvel ordre infrastructurel sur les vies citadines,” Flux, no. 135, pp. 103–114, 2024, doi: 10.3917/flux1.135.0103.

[12] W. Deng, H. Lei, and X. Zhou, “Traffic state estimation and uncertainty quantification based on heterogeneous data sources: A three detector approach,” Transp. Res. Part B, vol. 57, pp. 132–157, 2013, doi: 10.1016/j.trb.2013.08.015.

[13] S. Xu, L. Zhao, C. Wang & Z. He,“Traffic congestion estimation on urban road segments considering dynamic critical bottleneck based on GPS trajectory data,”Transportation Letters, pp. 1–20, 2025,doi: 10.1080/19427867.2025.2546422.

[14] Y. Liu et al., “How machine learning informs ride-hailing services: A survey,” Mach. Learn. Appl., Vol. 2, p. 100075, 2022, doi:10.1016/j.commtr.2022.100075.

[15] A. Y. Asuah, R. A. Acheampong, “Transport accessibility research in African cities: Systematic evidence review, knowledge gaps and directions for future research,” Urban Transitions, vol. 3, p. 100013, 2025, doi: 10.1016/j.ubtr.2025.100013.

[16] Y. Hou, Z. Deng, and H. Cui, “Short-term traffic flow prediction with weather conditions: Based on Deep Learning Algorithms and Data Fusion,” Complexity, Vol. 2021, no 1, p. 6662959, 2021, doi: 10.1155/2021/6662959.

[17] S. Yu, J. Peng, Y. Ge, X. Yu, F. Ding, S. Li, C. Ma, “A traffic state prediction method based on spatial-temporal data mining of floating car data by using autoformer architecture,” Computer-Aided Civil and Infrastructure Engineering, Vol. 39, no. 18, pp. 2774 – 2787, 2024, doi: 10.1111/mice.13179.

[18] S. Sun, J. Chen, and J. Sun, “Traffic congestion prediction based on GPS trajectory data,” Int. J. Distrib. Sensor Netw., vol. 15, no. 5, 2019, doi: 10.1177/1550147719847440.

[19] I. Benfaress, B. Afaf and Z. Ahmed, "Enhancing Traffic Accident Severity Prediction Using ResNet and SHAP for Interpretability," AI, vol. 5, no. 4, pp. 2568-2585, doi: 10.3390/ai5040124.

[20] A. Grigorev et al., “Traffic incident duration prediction: A systematic review of techniques,” Adv. Transportation Rev., Vol. 2024, no.1, p. 3748345, 2024, doi:10.1155/atr/3748345.

[21] Y. Zhang et al., “Incorporating multimodal context information into traffic speed forecasting through graph deep learning,” International Journal of Geographical Information Science, Vol. 37, no. 9, pp. 1909 – 1935, 2023, doi: 10.1080/13658816.2023.2234959.

[22] S. Guo et al., “Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting,” in Proc. AAAI, vol. 33, no. 1, pp. 922–929, 2019, doi: 10.1609/aaai.v33i01.3301922.

[23] Bai et al., “A3T-GCN: Attention Temporal Graph Convolutional Network for Traffic Forecasting,” ISPRS Int. J. Geo-Inf., vol. 10, no. 7, p. 485, 2021, doi: 10.3390/ijgi10070485.

[24] X. Zong, Z. Chen, F. Yu & S. Wei, “Local-global spatial-temporal graph convolutional network for traffic flow forecasting,” Electronics, vol. 13, no. 3, p. 636, 2024, doi: 10.3390/electronics13030636.

[25] Qiu et al., “Traffic prediction with data fusion and machine learning,” Digital, vol. 4, no. 2, p. 12, 2025, doi:10.3390/analytics4020012.

[26] A. E. Hoerl and R. W. Kennard, “Ridge Regression: Biased Estimation for Nonorthogonal Problems,” Technometrics, vol. 12, no. 1, pp. 55–67, 2012, doi: 10.1080/00401706.1970.10488634.

[27] C. Wang, Y. Hou, and M. Barth, “Data-driven multi-step demand prediction for ride-hailing services using convolutional neural network,” in Advances in Computer Vision, vol. 944, Springer, 2020, pp. 11–22, doi:10.1007/978-3-030-17798-0_2.

[28] R. Liu and S. Shin, “A review of traffic flow prediction methods in intelligent transportation system construction,” Appl. Sci., vol. 15, no. 7, p. 3866, 2025, doi:10.3390/app15073866.

[29] L. Breiman, “Random forests,” Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[30] M. Attioui et al., “Congestion forecasting using machine learning: A systematic review,” Smart Cities, vol. 5, no. 3, p. 76, 2025, doi:10.3390/futuretransp5030076.

[31] M. Veres & M. Moussa, “Deep learning for intelligent transportation systems: A survey of emerging trends,” IEEE Transactions on Intelligent transportation systems, vol. 21, no. 8, p. 3152-3168, 2019, doi: 10.1109/TITS.2019.2929020.

[32] X. Liu, L. Qin, M. Xu et al., “A comprehensive review of traffic flow forecasting based on deep learning,” Neurocomputing, p. 132269, 2025, doi: 10.1016/j.neucom.2025.132269.

[33] R. A. Acheampong, E. Agyemang, and A. Y. Asuah, “Is ride-hailing a step closer to personal car use? Exploring associations between car-based ride-hailing and car ownership and use aspirations among young adults,” Travel Behaviour and Society, vol. 33, p. 100614, 2023, doi: 10.1016/j.tbs.2023.100614.

[34] Yizhe Wang, Yangdong Liu & Xiaoguang Yang, “An Empirical Comparison of Urban Road Travel Time Prediction Methods — Deep Learning, Ensemble Strategies and Performance Evaluation”, Applied Sciences, 15(14): 8075, 2025. https://doi.org/10.3390/app15148075 MDPI

[35] Ali, R., Ali, A., Naeem, H. M. Y., Asad, M., Alsarhan, T., & Heyat, M. B. B., “A Comprehensive Survey of Deep Learning-Based Traffic Flow Prediction Models for Intelligent Transportation Systems”,ICCK Transactions on Advanced Computing and Systems, 1(3): 117–137, 2024. https://doi.org/10.62762/TACS.2025.795448

[36] A. J. Smola and B. Schölkopf, “A tutorial on support vector regression,” Stat. Comput., vol. 14, no. 3, pp. 199–222, 2004, doi: 10.1023/B:STCO.0000035301.49549.88.

[37] K.-L. Du, B. Jiang, J. Lu, J. Hua, M. N. S. Swamy, “Exploring Kernel Machines and Support Vector Machines: Principles, Techniques, and Future Directions,” Mathematics, Vol. 12, no. 24, p. 3935, 2024, doi: 10.3390/math12243935.

[38] S. Yang et al., “Ensemble learning for short-term traffic prediction,” J. Sensors, Vol. 2017, no. 1, 2017, doi: 10.1155/2017/7074143.

[39] Y. Ning et al., “A review of research on traffic flow prediction methods based on deep learning,” ACM Comput. Surv., 2024, pp. 166–170, doi:10.1145/3677892.3677922.

[40] S. Afandizadeh, S. Abdolahi, and H. Mirzahossein, “Deep learning algorithms for traffic forecasting: A comprehensive review and comparison with classical ones,” J. Adv. Transportation, Vol. 2024, no. 1, p. 9981657, 2024, doi:10.1155/2024/9981657.

[41] B. Gomes, J. Coelho, and H. Aidos, “A survey on traffic flow prediction and classification,” Intell. Syst. Appl., vol. 20, p. 200268, 2023, doi:10.1016/j.iswa.2023.200268.

[42] J. Dong et al., “TCEVIS: Visual analytics of traffic congestion influencing factors based on explainable machine learning,” Visual Informatics, vol. 8, no. 1, pp. 56–66, 2024, doi: 10.1016/j.visinf.2023.11.003.

[43] H. Zhang and Z. Jing, “Machine learning in intelligent transportation: A systematic review,” Adv. Eng. Technol. Res., vol. 14, no. 1, p. 945, 2025, doi: 10.56028/aetr.14.1.945.2025.

[44] W. Xu and Y. Huang, “Mining urban congestion evolution characteristics,” Am. J. Traffic Transp. Eng., vol. 5, no. 1, pp. 1–7, 2020, doi: 10.11648/j.ajtte.20200501.11.

[45] Y. Deng, “A hybrid network congestion prediction method integrating association rules and LSTM for enhanced spatiotemporal forecasting,” Transactions on Computational and Scientific Methods, vol. 5, no. 2, 2025, doi: 10.5281/zenodo.14912727.