GIS-Based Public Transport Accessibility Assessment and Random Forest Classification for Ward-Level Accessibility Analysis in Chennai, India

J. Nithya[1], *S. Jeyalaksshmi[1]

[1]Department of Computer Science, School of Computing Sciences, Vels Institute of Science Technology and Advanced Studies, Pallavaram, Chennai, Tamil Nadu, India-600117.
profnithyajagadeesan@gmail.com

[2]Department of Computer Applications, School of Computing Sciences, Vels Institute of Science Technology and Advanced Studies, Pallavaram, Chennai, Tamil Nadu, India-600117

*pravija.lakshmi@gmail.com

Abstract—Urbanization and the increasing need for public transport services have raised the need for accessibility assessment in sustainable urban mobility planning. Traditional methods for evaluating accessibility mainly depend on spatial analysis and lack predictive power for pinpointing accessibility patterns within an urban area. This research proposes a novel GIS and machine learning framework for estimating and classifying public transport access at the ward level in Chennai city, India. Spatial datasets such as ward boundaries, population distribution, bus stop distribution, and public transport service areas were integrated in a GIS environment to derive key accessibility indicators. The key indicators derived were population and bus stop density, and service coverage at 400 m and 800 m walking distances. Using these parameters, wards were classified into High, Medium, and Low accessibility classes.

A Random Forest classifier using 6 spatial and demographic features was developed to classify accessibility levels. According to the performance metrics, the overall classification accuracy of the proposed model was 96.67%. Ten-fold cross-validation produced a mean accuracy of 99.00% (standard deviation 3.00%)demonstrating the robustness and stability of the proposed model". The 400 m service coverage was found to be the most important factor influencing accessibility, followed by the 800 m service coverage and bus density. Combining GIS-based spatial analysis with machine learning creates a powerful decision-support framework for identifying underserved areas that should be prioritised for transport improvement. The suggested approach can help city planners and transportation authorities design data-driven plans to enhance public transport accessibility and promote sustainable urban growth in fast-growing metropolitan cities.

Keywords—Public Transport Accessibility; Geographic Information System (GIS); Random Forest; Machine Learning; Spatial Analysis; Accessibility Classification; Sustainable Urban Transportation; Chennai City

How to cite this article: Nithya J, Jeyalaksshmi S. GIS-Based Public Transport Accessibility Assessment and Random Forest Classification for Ward-Level Accessibility Analysis in Chennai, India. Int J Drug Deliv Technol. 2026;16(71s): 522-534. DOI: 10.25258/ijddt.16.71s.64

I. INTRODUCTION

Urban growth and ever-increasing population have resulted in a greater demand for efficient public transport systems all over the world. Public transport plays a vital role in reducing traffic congestion, pollution, and enabling affordable movement for the urban population. The public transport system's effectiveness depends greatly on the accessibility of its transport stations. Poor accessibility causes unequal opportunities in transport, greater dependence on private transport, and lower social and economic mobility, especially in dense cities.

Accessibility evaluation has thus become an integral part of sustainable urban transport planning. A variety of demographic, transportation, and land-use datasets have been used in GIS for the analysis of spatial accessibility. GIS-based analyses help planners identify areas with inadequate public transport, assess service coverage, and support infrastructure development. Common accessibility indicators include: population density, transport network density, distance from transit facilities, and service coverage within standard, acceptable walking distances of 400 m and 800 m. These indicators map the spatial distribution of transportation services across urban areas.

While GIS-based techniques provide effective visual representations of accessibility patterns, they largely result in descriptive spatial analysis, lacking predictive capabilities for categorizing accessibility conditions. Given the accumulated spatial and demographic characteristics of complex urban transportation systems, machine learning (ML) methodologies can recognize the interrelationships among a multitude of accessibility features. Machine learning algorithms can process historical spatial data and capture non-linear interactions among variables to learn patterns, enabling predictive models capable of yielding reliable forecasts for evidence-based transportation planning.

The Random Forest classifier is one of the most widely used machine learning algorithms, offering classification with high accuracy while minimizing the risk of overfitting. It also allows evaluation of input-variable importance. Random Forest works by constructing many decision trees and combining their predictions for classification. It can handle heterogeneous datasets efficiently, while still providing reliable classification. Even when multiple spatial variables are involved, reliable classification can still be achieved. In addition, feature-importance analytics help identify the factors that most influence accessibility. This helps planners target their investments more effectively.

Chennai, a metropolitan city in India, has been rapidly urbanising, growing denser, and facing rising travel demand over the last two decades. Ongoing improvements to public transportation infrastructure have led to improvements in bus services in the study area. However, spatial disparities remain in the level of provision of bus services among wards. Some areas have strong transit coverage, while others have little or none at all for residents. Hence, a spatial analysis and predictive-modelling-based study can provide a comprehensive assessment of the accessibility gaps and support future planning.

This study proposes a framework for assessing and classifying public transport accessibility using GIS and machine learning at the ward level in Chennai City. Within a GIS environment, spatial data on ward boundaries, population distribution, and rates of bus stops and public transport services were used to derive six key accessibility indicators: population, population density, bus count, bus density, 400 m service coverage, and 800 m service coverage. The Random Forest classifier was employed to classify wards into three categories—high, medium, and low—using these indicators. Evaluation of the developed model was conducted using classification accuracy, a confusion matrix, feature-importance ranking, and 10-fold cross-validation for predictive reliability.

The major contributions of this study are summarized as follows:

- ✧ A comprehensive GIS framework developed for the quantification of ward-level public transport accessibility using demographic and spatial accessibility indicators.
- ✧ Accessibility evaluated across two service coverage areas (400 m and 800 m walking distances) to map public transport services in Chennai City.
- ✧ A Random Forest classification model developed to predict accessibility levels using six spatial and demographic variables, achieving strong performance.
- ✧ Identification, through feature-importance analysis, of the factors with the greatest influence on accessibility, useful for future improvements to the public transport system.
- ✧ A scalable, data-driven decision-support framework presented for enhancing accessibility and supporting sustainable urban mobility in metropolitan areas, intended for use by urban planners and transport authorities.

The rest of this paper is organized as follows. Section II reviews the literature on GIS-based accessibility assessment and machine learning applications in transportation planning. Section III discusses the study area, datasets, and data preparation procedures. The proposed methodology is presented in Section IV, covering the GIS processing and classification using random forests. The proposed framework's performance and experimental results are discussed in Section V. Section VI concludes the paper and outlines future research directions.

II. Related Work

The importance of providing accessibility to public transport services has led researchers to investigate it in detail. Accessibility analysis allows planners to assess the spatial distribution patterns for transit services and helps in adding new infrastructure in those regions. As Geographic Information Systems (GIS) have grown, researchers are increasingly utilizing spatial analysis in assessing accessibility based on demographic characteristics, transport network information, and walking-distance service areas.

GIS-based accessibility indicators used in several studies have included population density, distance to the bus stop, road network connectivity, and service coverage. Using standard walking distances of 400 m and 800 m, buffer-based accessibility analysis is increasingly employed to estimate the proportion of urban areas serviced by public transport. These studies provide useful spatial visualizations of service availability; however, they are mostly descriptive and do not lend themselves to predictively classifying accessibility levels across administrative units.

New opportunities for transportation planning have emerged as a result of advances in AI and ML. With machine learning algorithms, a large set of heterogeneous spatial data can be analyzed for nonlinear relationships among accessibility indicators to inform decisions. Algorithms such as Decision Trees, Support Vector Machines, Artificial Neural Networks, Gradient-Boosted Trees, and Random Forest have been successfully applied to travel-demand forecasting, traffic-congestion prediction, route optimization, accident-severity forecasting, and transit-service assessment.

The Random Forest algorithm is one of the better-performing classification algorithms. It operates on the basis of ensemble learning and helps prevent overfitting. It also performs well on complex datasets with several correlated variables and unequal distributions. Random Forest also provides feature-importance analysis, allowing researchers to easily gauge the most important variables for prediction performance. This quality makes it particularly suited to accessibility studies in transport, where various spatial and demographic indicators interact jointly.

Despite these advances, few studies have integrated GIS-based accessibility analysis with machine learning for ward-level public transport accessibility assessment. Most previous studies either focus solely on mapping physical spaces, or on building predictive models without detailed GIS accessibility indicators. In addition, a comprehensive machine-learning framework that integrates several accessibility measures (i.e., population density, bus density, bus stops, and multi-distance service coverage) is lacking in the literature, particularly for rapidly growing Indian metropolitan cities.

To overcome these limitations, this study proposes an integrated GIS and Random Forest framework to assess and classify the public transport accessibility of Chennai City. The proposed method classifies accessibility as high, medium, and low based on six spatial and demographic indicators, namely: total population, population density, bus count, bus density, 400 m service coverage, and 800 m service coverage. Assessment of the model is carried out using classification accuracy, confusion-matrix analysis, feature-importance ranking, and 10-fold cross-validation. By integrating spatial analysis with predictive machine learning, the proposed framework can act as a comprehensive decision-support tool for identifying underserved areas and supporting evidence-based transport planning for sustainable urban development.

III. Study Area and Data

3.1 Study Area

Chennai is the capital city of the Indian state of Tamil Nadu. It is one of the fastest-growing metropolitan areas in the country. Chennai also has a high demand for transport due to the rapid growth in the city's population. The Greater Chennai Corporation (GCC) covers an area of around 426 km² and consists of 200 municipal wards spread across 15 administrative zones.

The city is a major commercial, industrial, educational, and IT hub, attracting a large number of daily commuters from nearby areas. The Metropolitan Transport Corporation (MTC) is the main public bus service provider in Chennai, connecting residential, commercial, industrial, and institutional areas. The city has a good bus network, but bus stop and service availability varies greatly across wards spatially. In general, city-center areas are more accessible, as they are densely serviced by transport networks, compared with several peripheral wards that have access to public transport services but less extensively. Because of these spatial inequalities, Chennai is a suitable case study for assessing public transport accessibility at the ward level using GIS and machine learning. Figure 1 shows the ward boundaries of the study area administratively.

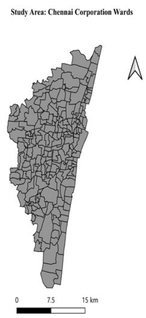

Fig. 1. Study area showing the ward boundaries of Chennai City.

3.2 Data Sources

The accessibility of public transport was studied using various datasets. Administrative ward boundary data for the Greater Chennai Corporation served as the principal spatial layer for the analysis. Ward-level population information was obtained from the 2011 Census of India and was joined with the ward boundary layer to estimate demographic characteristics.

Bus stop data from the Metropolitan Transport Corporation (MTC) network served as the source of public transport information. Bus stop locations were processed in the GIS environment to assess spatial accessibility. To assess the spatial coverage of public transport services within each ward, service areas representing walking distances of 400 m and 800 m were generated around each bus stop.

QGIS was employed for all spatial preprocessing, attribute integration, overlay analysis, and accessibility calculations. The GIS outputs were exported as attribute tables before being used to build the machine learning models in Python.

3.3 Data Preparation

Each ward has its own accessibility attributes, characterized by spatial indicators derived from GIS. Population density was estimated by dividing the ward-level population by the area of the ward in square kilometres. Bus stop counts for each ward were derived using a spatial-overlay technique. Bus density was computed by calculating the number of bus stops per unit of ward area.

To assess service accessibility, 400 m and 800 m walking-distance buffer zones were generated around each bus stop and then dissolved to obtain a continuous service-coverage area. A spatial intersection analysis was then performed to calculate the percentage of each ward covered by the 400 m and 800 m service areas. The coverage percentages reflect the proportion of a ward that falls within walking distance of public transport.

Selection of GIS-Derived Input Features for Machine Learning

Using the derived GIS indicators, six variables were defined as input features for the machine learning model: population, population density, bus count, bus density, 400 m service coverage, and 800 m service coverage. Accessibility levels were classified into three classes—High, Medium, and Low—which served as the target variable for classification. The cleaned dataset comprised 200 wards and six predictive variables after excluding incomplete records. This processed dataset was then used to train and test the Random Forest classification model.

TABLE 1. DATASET DESCRIPTION

Feature	Description	Unit
Population	Total population of each ward	Persons
Population Density	Population per square kilometre	Persons/km²
Bus Count	Number of bus stops in each ward	Count
Bus Density	Bus stops per square kilometre	Stops/km²
Coverage400	Ward area covered within 400 m walking distance	%
Coverage800	Ward area covered within 800 m walking distance	%
Access Class	Accessibility category (High, Medium, Low)	Class

IV. METHODOLOGY

4.1 Overall Framework

The adopted research methodology uses GIS and machine learning to assess the level of ward-wise public transport accessibility within the city. The framework comprises four distinct stages: (i) spatial data collection and preparation, (ii) GIS-based accessibility analysis, (iii) GIS-based feature extraction and dataset preparation, and (iv) Random Forest-based ward classification.

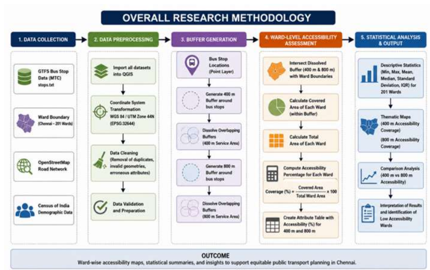

Fig. 2. Overall workflow of the proposed GIS and Random Forest-based public transport accessibility assessment framework.

The overall workflow adopted in this study is shown in Figure 2. First, spatial datasets, such as ward boundaries, ward-level population data, and bus stop locations, were collected and integrated in QGIS. Through spatial-preprocessing operations, ward area and population data were used to calculate bus density and service coverage within 400 m and 800 m walking distances. The indicators derived from GIS were combined to create an ML dataset. Finally, a Random Forest classifier was developed to categorize each ward into High, Medium, or Low accessibility classes. Model performance was evaluated using classification accuracy, confusion-matrix visualization, feature importance, and 10-fold cross-validation.

4.2 GIS-Based Accessibility Assessment

GIS was used to conduct all spatial analyses for accessibility assessment. The analysis used administrative ward boundaries as the main spatial unit. Population data were joined with ward polygons to estimate demographic characteristics. In addition, bus stop locations were spatially joined to assess the availability of public transport facilities within the wards. Accessibility-related indicators were computed using spatial overlay and attribute analysis.

Buffer analysis was performed to derive service areas representing walking distances of 400 m and 800 m from each bus stop. The service areas were dissolved to eliminate overlapping buffers, and then intersected with the respective ward boundaries to calculate the percentage of each ward covered by public transport services.

The GIS environment made it possible to compute ward-level accessibility indicators that form the input variables for machine-learning classification.

4.3 Feature Engineering

Six quantitative variables were selected to represent the accessibility characteristics of each ward:

- ✧ Population: Total population residing within each ward.
- ✧ Population Density: Population per square kilometre of ward area.

Population density represents the number of people residing per square kilometre of ward area and is calculated as:

$$[PD_i = \frac{P_i}{A_i}] \qquad (1)$$

where:

(PD_i) = Population density of ward i (persons/km²)

(P_i) = Population of ward i

(A_i) = Area of ward i (km²)

- ✧ Bus Count: Number of bus stops located within each ward.
- ✧ Bus Density: Number of bus stops per square kilometre.

Bus density represents the number of bus stops available per square kilometre and is calculated as:

$$[BD_i = \frac{B_i}{A_i}] \qquad (2)$$

where:

(BD_i) = Bus density of ward i (stops/km²)

(B_i) = Number of bus stops in ward i

(A_i) = Area of ward i (km²)

- ✧ 400 m Service Coverage: Percentage of ward area accessible within a 400 m walking distance from bus stops.

The percentage of ward area covered within a 400 m walking distance from bus stops is computed as

$$[C_{400} = \frac{A_{400}}{A_w} \times 100] \quad (3)$$

where:

(A_{400}) = Area covered within the 400 m buffer

(A_w) = Total ward area

- 800 m Service Coverage: Percentage of ward area accessible within an 800 m walking distance from bus stops.

Similarly, the percentage of ward area covered within an 800 m walking distance is calculated as

$$[C_{800} = \frac{A_{800}}{A_w} \times 100] \quad (4)$$

where:

(A_{800}) = Area covered within the 800 m buffer

(A_w) = Total ward area

These variables collectively describe the demographic demand and spatial availability of public transportation services.

Following feature generation, incomplete records containing missing values were removed. The final machine learning dataset consisted of 200 ward observations and six predictive variables. Accessibility levels were categorized into three classes: High, Medium, and Low.

The accessibility classes used as the target variable for the Random Forest classifier were generated from the GIS-derived accessibility indicators. For each ward, six indicators were computed: population, population density, bus count, bus density, 400 m service coverage, and 800 m service coverage. These indicators collectively represent both the demand for public transport and the spatial availability of bus services.

An overall accessibility score was first calculated by integrating the normalized values of the six indicators. The resulting scores were then classified into three accessibility levels (High, Medium, and Low) using the Natural Breaks (Jenks) classification method available in QGIS. The Jenks algorithm identifies natural groupings within the data by minimizing the variance within each class while maximizing the variance between classes. This approach provides a more meaningful classification than equal-interval or arbitrary threshold methods because it reflects the inherent distribution of the accessibility scores.

The generated accessibility classes served as the target labels for the Random Forest classification model, while the six GIS-derived indicators were used as predictor variables. This methodology enabled the model to learn the spatial patterns associated with different levels of public transport accessibility across the 200 wards of Chennai City.

4.4 Random Forest Classification

Random Forest was selected as the classification algorithm because it is robust and provides high accuracy when predicting outcomes involving complex, interdependent variables. The algorithm builds an ensemble of decision trees based on bootstrap sampling and random feature selection. The final accessibility class is decided by majority voting across all trees. This approach reduces model variance and helps prevent overfitting.

The prepared dataset was split into training and testing sets using a 70:30 ratio. The predictor variables were the six GIS-derived variables, and the target variable was accessibility class. The Random Forest model, implemented in Python using the scikit-learn library, consists of 100 decision trees with a fixed random state for reproducibility.

The Random Forest classifier predicts the accessibility class by combining the outputs of multiple decision trees through majority voting:

$$[\hat{Y} = mode(T_1(x), T_2(x), \ldots, T_n(x))] \quad (5)$$

where:

$(\hat{Y})$ = Predicted accessibility class

$(T_i(x))$ = Prediction from the (i^{th}) decision tree

(n) = Total number of decision trees

TABLE 2. RANDOM FOREST HYPERPARAMETER SETTINGS

Parameter	Value	Description
Algorithm	Random Forest	Ensemble learning

Parameter	Value	Description
	Classifier	algorithm used for accessibility classification
Number of Trees (n_estimators)	100	Number of decision trees in the forest
Criterion	Gini Impurity	Function used to measure split quality
Maximum Tree Depth (max_depth)	None	Trees grow until all leaves are pure or stopping criteria are met
Minimum Samples Split (min_samples_split)	2	Minimum number of samples required to split an internal node
Minimum Samples Leaf (min_samples_leaf)	1	Minimum number of samples required at each leaf node
Bootstrap Sampling	True	Random sampling with replacement used for building trees
Random State	42	Ensures reproducibility of results
Train-Test Split	70:30	Training and testing data partition
Cross-Validation	10-fold	Used to evaluate model robustness
Software	Python 3.x	Programming environment
Library	Scikit-learn	Machine learning implementation

4.5 Performance Evaluation

Multiple statistical measures were used to evaluate the predictive performance of the Random Forest classifier. Classification accuracy on the testing sample allowed for a comparison of model performance on a single scale. To visualize the classification performance among the three accessibility categories, a confusion matrix was generated. To measure classification quality across the different classes, precision, recall, and F1 score were calculated.

The relative contribution of each predictor variable was assessed using the impurity-based importance measure provided by Random Forest, which is commonly employed for feature-importance evaluation. The analysis identified the variables contributing most to the prediction of accessibility.

Additionally, 10-fold cross-validation was used to assessed model robustness. The dataset was divided into ten subsets, nine of which were used for training and one was used for testing in each iteration. To evaluate the stability and generalization capability of the proposed model, the average cross-validation accuracy and standard deviation were computed.

- Precision

$$[Precision = \frac{TP}{TP+FP}] \qquad (6)$$

- Recall

$$[Recall = \frac{TP}{TP+FN}] \qquad (7)$$

- F1-score

$$[F1 = \frac{2\times Precision\times Recall}{Precision+Rec}] \qquad (8)$$

These mathematical formulations establish the quantitative basis for deriving accessibility indicators and evaluating the performance of the proposed Random Forest classification model.

V. Results and Discussion

5.1 Population Density Analysis

Population density is one of the main factors influencing public transport demand. Wards in Chennai differ in population density, and this variation is shown spatially in Figure 3. Densities differ greatly across wards. Central and more urbanized areas tend to have greater densities than peripheral wards, which tend to be more residential and rely more heavily on public transport.

Population density was calculated by dividing ward-level population by the corresponding ward area. Public transport services in high-density wards should be more frequent and efficient, given the greater passenger demand. Lower-density wards, in contrast, generally experience reduced service demand. Because of this, population density is an important demographic variable for assessing accessibility, and it was therefore included as one of the predictor variables in the machine learning model.

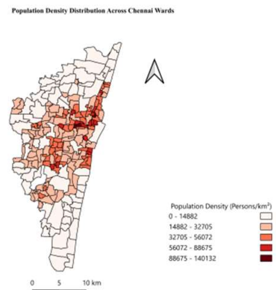

Fig. 3. Ward-level population density distribution in Chennai City.

5.2 Bus Density Analysis

Bus density in Chennai City is shown in Figure 4. Bus density was calculated by dividing the number of bus stops in each ward by its area in square kilometres. The analysis reveals that, spatially, bus stops are not distributed uniformly. Centrally located urban wards tend to have a higher density of bus stops compared to suburban areas.

Higher bus density indicates better spatial accessibility and offers residents more opportunities to reach public transport within shorter distances. In contrast, wards with lower bus density may be less accessible, even when population is high. These findings emphasise the significant role of demographic demand and transport infrastructure in accessibility analysis.

.

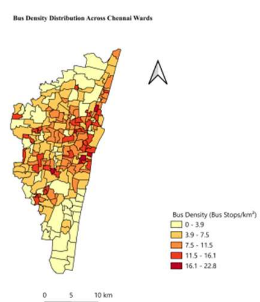

Fig. 4. Spatial distribution of bus density across Chennai wards.

5.3 Analysis of 400 m Service Coverage

The walking-distance service-coverage map shown indicates the proportion of each ward accessible within a walking distance of 400 m—and hence, directly accessible to a bus stop within an approximate five-minute walk. The findings show that accessibility varies greatly across Chennai's wards. Several wards in central locations show substantial 400 m coverage. This is mainly due to the density of public transport infrastructure and closely spaced bus stops.

In contrast, many peripheral wards show considerably lower service coverage, and additional public transport facilities would be required to improve access there. The 400 m threshold is generally considered a comfortable walking distance for users of public transport systems, which makes this indicator an effective measure of immediate access.

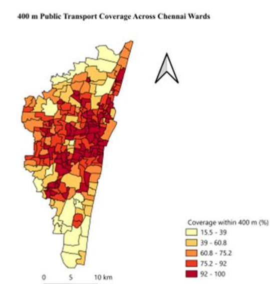

Fig. 5. Public transport service coverage within a 400 m walking distance.

5.4 Analysis of 800 m Service Coverage

Ward-level accessibility to bus stops within an 800 m walking distance is shown in Figure 6. This coverage is much greater than that shown by the 400 m buffer, for most wards.

Several wards that demonstrated a moderate level of accessibility at the 400 m threshold achieved a significant improvement in their accessibility rating at the 800 m service area. This observation suggests that, while residents may not always have immediate access to a bus stop, they can reach a high proportion of the urban area on foot within a reasonable distance. Using both service-coverage indicators allows for a thorough assessment of public transport accessibility.

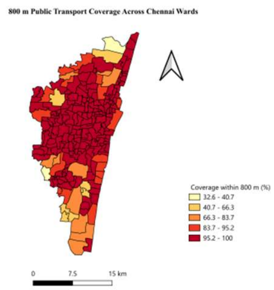

Fig. 6. Public transport service coverage within an 800 m walking distance.

5.5 Accessibility Classification

According to GIS-derived accessibility indicators, wards were classified into three accessibility categories: High, Medium, and Low. The accessibility classification map is shown in Figure 7, which displays the different accessibility patterns across Chennai City. The most accessible wards are located in areas with higher bus density and greater bus service coverage. Wards classified as medium-accessibility exhibit moderate transport availability and partial service coverage.

Low-accessibility wards are generally associated with low bus-stop density and limited coverage. This classification serves as a useful decision-support tool for transportation planners to identify the underserved regions requiring future infrastructure improvement and service expansion.

TABLE 3. ACCESSIBILITY CLASS DISTRIBUTION

Accessibility Class	**Number of Wards**	**Percentage (%)**
High	109	54.50
Medium	46	23.00
Low	45	22.50
Total	200	100

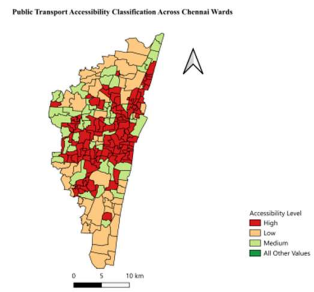

Fig. 7. Ward-level public transport accessibility classification (High, Medium, and Low).

5.6 Random Forest Classification Performance

A Random Forest classifier was trained on six GIS-derived variables: population and population density, bus count and bus density, and 400 m and 800 m service coverage. The dataset was divided into training and testing subsets at a ratio of 70:30.

The developed model achieved an overall classification accuracy of 96.67%. The confusion matrix (Figure 8) shows that most wards were classified accurately with respect to their accessibility levels, with only a few misclassified into adjacent accessibility classes. This high classification accuracy indicates Random Forest's strong ability to model the complex relationships between demographic and spatial accessibility indicators. These findings validate the effective application of machine learning methods in GIS-based transportation planning.

TABLE 4. CLASSIFICATION REPORT

Class	**Precision**	**Recall**	**F1-score**
High	0.97	1.00	0.98
Medium	0.94	0.94	0.94
Low	1.00	0.91	0.95
Overall Accuracy			96.67%

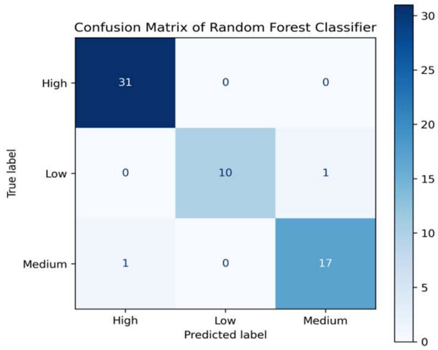

Fig. 8. Confusion matrix of the Random Forest classifier.

5.7 Feature Importance Analysis

An analysis of predictor variable importance was performed to assess each variable's contribution to the Random Forest classification. Figure 9 illustrates the results. Among all predictors, the 400 m service coverage had the greatest importance, contributing about 52.24% of the total model importance. This finding suggests that walkability to a bus stop has the greatest impact on overall accessibility.

The 800 m service coverage was the second most significant variable, followed by bus density and population density. In comparison, population and bus count contributed less to the classification.

These findings indicate that spatial accessibility indicators have stronger predictive capability than demographic variables, highlighting the need for GIS-based service-coverage analysis in assessing transport accessibility.

TABLE 5. FEATURE IMPORTANCE

Rank	Feature	Importance
1	Coverage400	0.522397
2	Coverage800	0.159840
3	Bus Density	0.154230
4	Population Density	0.105738
5	Population	0.030536
6	Bus Count	0.027259

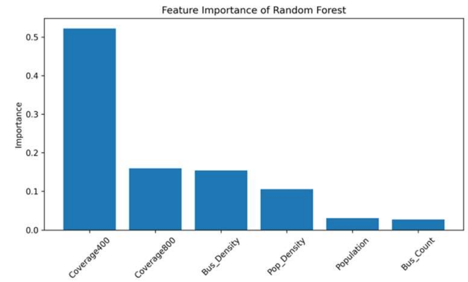

Fig. 9. Feature importance obtained from the Random Forest model.

5.8 Cross-Validation Analysis

To test the robustness and generalization ability of the proposed model, 10-fold cross-validation was performed. The model's cross-validation accuracy averaged 99.00%, with a standard deviation of 3.00%.The consistently high performance across the validation folds indicates that the proposed model generalizes well and provides stable accessibility predictions for unseen data.

The Random Forest model generalizes well, producing highly accurate predictions across validation sets in all folds, demonstrating its robustness against overfitting. The low standard deviation further suggests that classification performance remains consistent across different training and testing subsets. Thus, the proposed GIS and machine learning framework for assessing ward-level public transport accessibility is effective.

TABLE 6. CROSS-VALIDATION PERFORMANCE

Evaluation Metric	Value
Random Forest Accuracy	96.67%
Mean Cross-validation Accuracy	99.00%
Standard Deviation	3.00%
Number of Folds	10

VI. CONCLUSION AND FUTURE WORK

This study presented a machine learning and geographic information system framework for assessing public transport accessibility in Chennai City at the ward level. GIS techniques were used to analyze spatial data, such as ward boundaries, population, and bus stops, to derive accessibility indicators. Six indicators were derived, including population, population density, bus count, bus density, 400 m service coverage, and 800 m service coverage. The Random Forest classification model was used to classify accessibility across 200 municipal wards using these indicators.

The spatial analysis using GIS revealed variations in public transport accessibility across wards. Service-coverage analysis showed that the 800 m walking-distance buffer was considerably more accessible than the 400 m service area, while the accessibility classification revealed a number of wards with relatively low public transport availability. These results offer useful spatial insights to help identify underserved areas and improve transportation planning going forward.

The Random Forest classifier demonstrated a very high predictive performance, with an overall classification accuracy of 96.67% on the test dataset. Furthermore, the model achieved a mean accuracy of 99.00% in 10-fold cross-validation, with a standard deviation of 3.00%, indicating robustness. Feature-importance analysis revealed that the most influential factor for accessibility was the 400 m service coverage, followed by the 800 m service coverage and bus density. These findings underscore the importance of spatial accessibility indicators in assessing public transport availability.

The framework offers practical decision-support for transportation planners, municipal authorities, and policy-makers, combining spatial analysis with predictive machine learning. This approach enables easy identification of accessibility gaps and supports evidence-based planning for better public transport services and sustainable urban mobility.

While this study analyzes bus transport accessibility only at the ward level, several avenues for further research remain. Future studies could incorporate other transport modes, such as metro rail, suburban rail, and multimodal transport networks, to evaluate accessibility more comprehensively. Incorporating time-varying service frequency, passenger demand, traffic congestion, and real-time traffic information may further improve model performance. Comparing additional machine learning and deep learning algorithms, such as XGBoost, LightGBM, CatBoost, and neural network models, could also provide further insights into accessibility prediction. These extensions would contribute to the development of intelligent, data-driven transport systems capable of supporting sustainable urban planning and smart-city applications.

References

1. Adeniran, A. O., Adeniran, A. A., Ogieva, M. O., & Ogwuche, G. (2026). Adoption of Intelligent Transport Systems (ITS) in urban transportation planning. In Discover Global Society (Vol. 4, Number 1). Discover. https://doi.org/10.1007/s44282-025-00312-3

2. Wang, Y., Ram, S., Currim, F., Dantas, E., & Sabóia, L. (2016). A Big Data Approach for Smart Transportation Management on Bus Network. In Proc. IEEE Int. Smart Cities Conf. (ISC2) (pp. 1–6). IEEE. https://doi.org/10.1109/ISC2.2016.7580839

3. Citizen Consumer and Civic Action Group (CAG). (2024). Why Chennai's MTC App Is Failing Commuters: A CAG Study. The Times of India. https://timesofindia.indiatimes.com/city/chennai/why-chennais-mtc-app-is-failing-commu ters-a-cag-study/articleshow/115819361.cms

4. Gadepalli, R., Mahendru, S., & Verma, N. (2021). Roadmap for Data-Driven Transformation of Indian Bus Systems. Changing Transport/GIZ. https://www.changing-transport.org/publication/roadmap-for-data-driven-transformation-of-indian-bus-systems/

5. Government of Tamil Nadu. (2025). Chennai MTC Performance Data - 2020-21 to 2024-25. CKAN/OpenCity. https://data.opencity.in/dataset/chennai-mtc-performance-data

6. Zannat, K. E., & Choudhury, C. F. (2019). Emerging Big Data Sources for Public Transport Planning: A Systematic Review on Current State of Art and Future Research Directions. Journal of the Indian Institute of Science (Vol. 99, Issue 4, pp. 601–619). Springer. https://doi.org/10.1007/s41745-019-00125-9

7. Krúpová, S., & Koman, G. (2026). The Role of IoT and Big Data in Smart City Transport Management: Lessons Learned from Case Studies in Slovakia and the EU.

Transportation Research Procedia (Vol. 93, pp. 948–959). Elsevier. https://doi.org/10.1016/j.trpro.2025.12.027

8. Loia, F., Perillo, C., & Gravili, G. (2026). Unleashing the power of digital twin and big data as a new frontier for smart mobility: An ecosystem perspective. Big Data Research (Vol. 43, 100576). Elsevier. https://doi.org/10.1016/j.bdr.2026.100576

9. Tzika-Kostopoulou, D., Nathanail, E., & Kokkinos, K. (2024). Big data in transportation: a systematic literature analysis and topic classification. Knowledge and Information Systems (Vol. 66, pp. 5021–5046). Springer. https://doi.org/10.1007/s10115-024-02112-8

10. Zhang, Z., Jin, C. J., Song, Y., & Li, D. (2026). AI-related papers in the transportation field: Statistics and evolutionary trends. Transportation Research Interdisciplinary Perspectives (Vol. 35, 101803). Elsevier. https://doi.org/10.1016/j.trip.2025.101803

11. Oraon, B. (2025). Chennai Transfer Friction Index – Thesis Research Live Project. SPA Vijayawada. https://github.com/bharatoraon/Chennai-Transfer-Friction

12. Jiang, W., & Li, J. (2022). Big Data for Traffic Estimation and Prediction: A Survey of Data and Tools. Applied System Innovation (Vol. 5, Issue 1). MDPI. https://doi.org/10.3390/asi5010023

13. Hassan, M., Mahin, H. D., Al Nafees, A., Paul, A., & Shraban, S. S. (2025). Big data applications in intelligent transport systems: a bibliometric analysis and review. Discover Civil Engineering (Vol. 2, 49). Springer. https://doi.org/10.1007/s44290-025-00205-z

14. NITI Aayog. (2015). National Urban Transport Helpline (NUTH) Operations Document. Sustainable Urban Transport Project (India). https://www.niti.gov.in/sites/default/files/2023-02/NUTH_Operations_Document.pdf

15. Son, H., et al. (2024). Leveraging Advanced Technologies for (Smart) Transportation Planning: A Systematic Review. Sustainability (Vol. 17, 2245). MDPI. https://doi.org/10.3390/su17022245

16. Nikitas, A., Michalakadakou, K., Nielsen, I. C., Itoh, E., & Dadashi, A. (2020). A Sustainable and Equitable Smart Mobility Strategy for the Future. In Sustainability (Vol. 12, Issue 13). MDPI. https://doi.org/10.3390/su12135476

17. Batty, M., Axhausen, K. W., Giannotti, F., Pozdnoukhov, A., Bazzani, A., Wachowicz, M., & Portoli, Y. (2012). Smart cities of the future. In The European Physical Journal Special Topics (Vol. 214, pp. 481–518). Springer. https://doi.org/10.1140/epjst/e2012-01703-3

18. Arsenio, E., Martens, K., & Di Ciommo, F. (2016). Sustainable urban mobility plans: bridging climate change and equity targets? In Research in Transportation Economics (Vol. 55, pp. 30–39). Elsevier. https://doi.org/10.1016/j.retrec.2016.04.008

19. Anda, C., Erath, A., & Fourie, P. J. (2017). Transport modelling in the age of big data. In International Journal of Urban Sciences (Vol. 21, Issue 1, pp. 19–42). Taylor & Francis. https://doi.org/10.1080/12265934.2017.1281150

20. Lian, Y., Zhang, G., Lee, J., & Huang, H. (2020). Review on big data applications in safety research of intelligent transportation systems and connected/automated vehicles. In Accident Analysis & Prevention (Vol. 146, 105711). Elsevier. https://doi.org/10.1016/j.aap.2020.105711

21. Abirami, S., Pethuraj, M., Uthayakumar, M., & Chitra, P. (2024). A systematic survey on big data and artificial intelligence algorithms for intelligent transportation system. In Case Studies on Transport Policy (Vol. 17, 101247). Elsevier. https://doi.org/10.1016/j.cstp.2024.101247

22. Ma, X., Liu, C., Wen, H., Wang, Y., & Wu, Y. J. (2017). Understanding commuting patterns using transit smart card data. In Journal of Transport Geography (Vol. 58, pp. 135–145). Elsevier. https://doi.org/10.1016/j.jtrangeo.2016.12.0014

23. Tao, S., Corcoran, J., Mateo-Babiano, I., & Rohde, D. (2014). Exploring Bus Rapid Transit passenger travel behaviour using big data. In Applied Geography (Vol. 53, pp. 90–104). Elsevier.

https://doi.org/10.1016/j.apgeog.2014.06.008

24. Pelletier, M. P., Trépanier, M., & Morency, C. (2011). Smart card data use in public transit: a literature review. In Transportation Research Part C: Emerging Technologies (Vol. 19, Issue 4, pp. 557–568). Elsevier. https://doi.org/10.1016/j.trc.2010.08.003

25. Bwambale, A., Choudhury, C. F., Hess, S., & Iqbal, M. S. (2019). A framework for combining disaggregate travel survey data and aggregate mobile phone data for trip generation modelling. In Proc. 15th World Conference on Transport Research. Mumbai, India. https://doi.org/10.1016/j.cstp.2019.03.001

26. Kieu, L. M., Bhaskar, A., & Chung, E. (2015). A modified density-based scanning algorithm with noise for spatial travel pattern analysis from smart card AFC data. In Transportation Research Part C: Emerging Technologies (Vol. 58, pp. 193–207). Elsevier. https://doi.org/10.1016/j.trc.2015.01.012

27. Yigitcanlar, T., Butler, L., Windle, E., Desouza, K. C., Mehmood, R., & Corchado, J. M. (2020). Can building “artificially intelligent cities” safeguard humanity from natural disasters, pandemics, and other catastrophes? In Sensors (Vol. 20, Issue 10). MDPI. https://doi.org/10.3390/s20102988

28. Guan, S., Yun, J., Zhang, Q., & Aurora, A. (2016). Intelligent Transportation System Contributions to the Operating Efficiency of Urban Traffic. In Journal of Intelligent & Fuzzy Systems (Vol. 31, pp. 2213–2220). IOS Press. https://doi.org/10.3233/IFS-162204

29. Luo, D., et al. (2019). IoT-based dynamic dispatching and route optimization for public transport systems. In IEEE Internet of Things Journal (Vol. 6, Issue 6). IEEE. https://doi.org/10.1109/JIOT.2019.2941011

30. Williams, S., White, A., Waiganjo, P., Orwa, D., & Klopp, J. (2015). The digital matatu project: using cell phones to create an open source data for Nairobi’s semi-formal bus system. In Journal of Transport Geography (Vol. 49, pp. 39–51). Elsevier. https://doi.org/10.1016/j.jtrangeo.2015.10.005

31. S. Jeyalaksshmi and R. M. Ganesh, "Adaptive Duty-Cycle Scheduling using Bi-Directional Long Short-Term Memory (BiLSTM) for Next Generation IoT Applications," 2022 International Conference on Computing, Communication and Power Technology (IC3P), Visakhapatnam, India, 2022, pp. 75-80, doi: 10.1109/IC3P52835.2022.00024. keywords: {Adaptive scheduling;Adaptive systems;Processor scheduling;Telecommunication traffic;Traffic control;Prediction algorithms;Energy states;IoT;Artificial Intelligence;AIoT;BiLSTM;JSI.