The Open Transportation Journal
DOI: 10.2174/0126671212414633260119104836, 2026, 20, e26671212414633
ISSN: 1874-4478
1

RESEARCH ARTICLE **OPEN ACCESS**

Integrating Ensemble Machine Learning and Structural Equation Modeling to Predict and Explain Biking Demand in Kigali: A Data-driven Approach for Sustainable Urban Mobility

Jean Marie Vianney Ntamwiza[1,*], Hannibal Bwire[1] and Alphonse Nkurunziza[2]

[1]*Department of Transportation and Geotechnical Engineering, University of Dar Es Salaam, Dar es Salaam, Tanzania*
[2]*Department of Civil, Environmental and Geomatics Engineering, University of Rwanda, Kigali, Rwanda*

Abstract:

Introduction: Biking-both shared and non-shared-has become a vital component of sustainable urban mobility across African cities like Kigali. Despite this progress, empirical research and demand modeling of biking behavior remain limited. This study predicts and explains biking behavior in Kigali by integrating advanced ensemble machine learning (ML) techniques with structural equation modeling (SEM).

Methods: A dataset of 6,386 observations was compiled by merging survey responses on biking with secondary data on weather and air quality. Both traditional statistical and advanced ensemble models were developed for comparison. The dataset was partitioned into training (70%) and testing (30%) subsets, with correlation-based and model-based feature selection applied. SEM examined latent constructs representing spatial, social-demographics, temporal, environmental, and attitudinal factors.

Results: Ensemble ML models substantially outperformed traditional approaches, with random forest and XGB classifiers achieving the highest predictive performance. The SEM demonstrated good model fit and explained the variance in biking frequency. Perceived station accessibility emerged as the strongest determinant of biking behavior, while temporal and environmental factors indirectly influenced demand patterns.

Discussion: The combination of ML and SEM has revealed a coexistence of accurate prediction and behavioral insight. Accessibility emerged as central to biking uptake, highlighting the potential of station placement and spatial equity. Indirect effects of temporal and environmental conditions highlight the impact of user perceptions in shaping biking demand.

Conclusion: Integrating ensemble ML and SEM provides predictive robustness and behavioral insight. The findings highlight that improving spatial accessibility and adopting adaptive urban planning strategies enhance sustainable biking uptake.

Keywords: Ensemble learning models, Structural equation modeling (SEM), Biking demand prediction, Sustainable urban mobility, Weather factors, Air quality factors.

*Address correspondence to this author at the Department of Transportation and Geotechnical Engineering, University of Dar Es Salaam, Dar es Salaam, Tanzania; Tel: +250788313909; E-mail: ntamwizajeanmarie@gmail.com

Cite as: Ntamwiza J.M.V , Bwire H, Nkurunziza A. Integrating Ensemble Machine Learning and Struc-tural Equation Modeling to Predict and Explain Biking Demand in Kigali: A Data-driven Approach for Sustainable Urban Mobility. Open Transp J, 2026; 20: e26671212414633. http://dx.doi.org/10.2174/0126671212414633260119104836

Received: October 06, 2025
Revised: November 07, 2025
Accepted: November 25, 2025
Published: March 16, 2026

1. INTRODUCTION

The transportation sector is undergoing a transformation toward sustainable and inclusive mobility aligned with the SDGs (United Nations Sustainable Development Goals) and the New Urban Agenda [1]. Among the low-emission alternatives to motorized transport, biking-both shared and non-shared-has emerged as a viable and equitable mode for urban mobility. While bike sharing systems were formally introduced in the Netherlands in 1965 [2], their adoption has accelerated globally and is now gaining momentum in research focused on emerging African cities, such as Kigali, Rwanda [3], Kumasi in Ghana [4], and Quelimane in Mozambique [5].

Despite this, the modeling of biking demand in African cities remains underdeveloped, limiting the ability of policymakers to implement data-driven planning for non-motorized transport infrastructure. Traditional statistical approaches-such as Bayesian Conditional Autoregressive [6], k-nearest neighbors (KNN) and support vector machines (SVM) [7], and multinomial logit models [8] have frequently been used to model bike sharing demand. However, these models often underperform when dealing with nonlinear relationships, high-dimensional feature spaces [9, 10], and multicollinearity [11], which are common in transport datasets. Additionally, ensemble models, such as Random Forest [12], stacking classifiers [13], Extreme Gradient Boosting (XGBoost) [14], and Light Gradient Boosting Machine (LightGBM) [15] combine multiple learners to improve accuracy beyond what a single model can achieve. The above models have achieved a superior performance in demand prediction involving complexes along with high-dimensional data. However, ensemble models typically lack the ability to analyze causal or structural relationships among factors influencing bike sharing demand, which limits their explanatory power compared to theory-driven models like SEM (Structural Equation Modeling) [16].

To bridge the gap, this study is the first research in transportation integrating advanced ensemble machine learning models, including XGBoost, Random Forest, and a Stacking Classifier-with Structural Equation Modeling (SEM) to predict and explain biking demand, which existing research has failed to achieve. This dual-method framework provides both high predictive performance and behavioral interpretability, which machine learning or SEM cannot do alone. The analysis is based on a dataset of 6,386 observations, merging primary survey data on biking demand with secondary weather and air quality variables. This dataset enables robust predictive modeling in a data-scarce context and captures both objective and perceptual drivers of biking behavior.

The findings reveal that ensemble machine learning models significantly outperform traditional approaches, achieving predictive accuracies between 97% and 99%, compared to 42%–82% for conventional models. The SEM results yielded strong model fit indices (RMSEA = 0.036, CFI = 0.932) and explained 77.1% of the variance in biking frequency. Notably, spatial accessibility emerged as the most influential latent factor ($\beta = 0.878$, $p < 0.001$), with perceived station access having the strongest effect ($\beta = 0.995$). While environmental and temporal variables played a critical role in the predictive models, their direct causal pathways in the SEM were more limited. These results contribute actionable evidence to support low-carbon urban transport systems by guiding policy interventions, such as investment in car-free corridors, station accessibility, and seasonally adaptive cycling infrastructure.

The methodological framework used in this research is designed for replicability and scalability across other rapidly urbanizing African cities-such as Nairobi in Kenya, Kampala in Uganda, and Dar es Salaam in the United Republic of Tanzania, where similar data constraints and urban challenges exist. This paper begins by describing the materials and methods, followed by a presentation of the results. It then discusses the key findings and concludes with policy implications.

2. MATERIALS AND METHODS

2.1. Study Area

Rwanda is an East African country with Tanzania to the East, Burundi to the South, Uganda to the North, and the DRC (Democratic Republic of the Congo) to its West. In recent decades, Rwanda has experienced substantial demographic growth and urban expansion. Its population has been growing at an annual rate of approximately 2.86%, while the urbanization rate increased from 18.4% in 2012 to an estimated 35% by 2024 [17].

Kigali, the case study, is the capital of Rwanda with a population corresponding to 1,745,555 on 730 square kilometers and projected to reach 3.8 million by 2035 [18]. Compared to other East African cities, Kigali has a moderate population density, making it more suitable for planning non-motorized transport (NMT) easily. Its demographic characteristics and hilly topography also present promising opportunities for research and implementation of cycling initiatives.

Kigali has developed an NMT master plan to serve as its guiding framework. To date, the city has constructed 215 kilometers of dedicated cycling lanes, with plans to expand the network to 418 kilometers by 2050 [19]. These lanes are physically separated from motor vehicles and pedestrian pathways to enhance safety and encourage cycling. In addition, the city has introduced regular car-free days and established car-free zones, such as the Imbuga City Walk, to reduce greenhouse gas emissions [20]. Moreover, the city launched the first public bike sharing program in 2021 and has encouraged its employees to reduce reliance on motorized transport by providing electric bicycles, free of charge and maintenance [21].

2.2. Sample Strategies

Table **1** summarizes how data were collected from three districts of the city, and the sample size for each district was proportional to its population. A total sample of 6,386.00 was determined, then distributed to those three districts as per Table **1**.

Table 1. Stratified sampling strategy.

Districts of the City	% of Population in Each District	Sample Size for Each District
Gasabo	50.39%	3,218
Kicukiro	28.17%	1,799
Nyarugenge	21.44%	1,369

Table **2** identifies temporal, environmental, geographic, demographic, and behavioral factors influencing biking demand in Kigali. Bike users were categorized as non-shared, bike-sharing, or both. Rainfall, humidity, and poor air quality (high PM2.5, NO_2, O_3) reduced biking activity. Nyarugenge had the highest bike-sharing uptake due to better connectivity and station density, while Gasabo and Kicukiro favored non-shared bikes.

2.3. Weather Data and Air Quality Data

Daily weather and air quality data were collected between 14th December 2022 and 29th February 2024, aligned with the period of survey data collection. Weather data were collected from Meteo Rwanda in Kigali city stations, which included solar radiation intensity, humidity, rainfall, temperature variations, and wind speed. These variables were selected due to their effects on the comfort and safety of the riders [21]. Air quality data were collected from REMA (Rwanda Environment Management Authority) at its Kigali stations, including CO (carbon monoxide), PM2.5 (particulate matter), NO2 (nitrogen dioxide), and CO2 (carbon dioxide). These data were selected to understand how they affect bike-sharing adoption [22].

2.4. Data Pre-processing

Initially, all variables were converted to their appropriate data types. Irrelevant columns were identified through univariate analysis and subsequently removed. To maintain data integrity, missing values and outliers were carefully addressed. Records with more than 10% missing data were excluded, while missing values in numerical variables were estimated using time-based linear interpolation, leveraging the date-time index to fill gaps based on temporal progression [23].

Table 2. Summary of variables.

Variable	Name	Description	
Dependent Variable	**Biking Preference**	- Non-shared bike	
		- Bike sharing	
		- Both shared and non-shared	
Independent Variables	**Category**	**Variable Name**	**Description / Categories**
-	**Temporal**	Day	Days of the week
-	-	Month	Months of the year
-	-	Year	2022 to 2024
-	**Weather & Air Quality**	Relative Humidity	Solar radiation (W/m^2)
-	-	Wind Speed	%
-	-	Radiation	m/s
-	-	Wind Direction	Degrees (°N)
-	-	Cloud Opacity	%
-	-	Max Temperature	°C
-	-	Min Temperature	°C
-	-	Rainfall	mm
-	-	Atmospheric Pressure	hPa
-	-	SO_2, CO, NO_2, O_3, PM10, PM2.5	Air pollutants ($\mu g/m^3$ or ppb)
-	**Geographic**	District	Gasabo, Kicukiro, Nyarugenge
-	**Demographic**	Age	Below 18, 19–24, 25–30, 31–36, Above 44
-	-	Gender	Male, Female, Other
-	-	Household Income	Low, Lower-middle, Upper-middle, High
-	-	Education Level	Primary, Secondary, Tertiary
-	-	Occupation	Student, Employed, Unemployed, Retired
-	**Behavioral / Preferences**	Average Bike Trip Duration	<30 min, 30–45 min, 45–60 min
-	-	Satisfaction Level with Biking Services	Very satisfied, Satisfied, Neutral, Dissatisfied
-	-	Distance to Nearest Bike Sharing Station	<500m, 1–3 km, >3 km
-	-	Influencing Factors	Infrastructure, Convenience, Cost-effectiveness, Safety, Accessibility, Environmental Concerns
-	-	Desired Biking Features	Protected lanes, Signals, Bike racks, and Station availability
-	-	Infrastructure Quality & Accessibility	Excellent, Good, Fair, Poor, Very Poor

2.5. Exploratory Data Analysis

Exploratory data analysis (EDA) was conducted to examine the characteristics, distributions, and patterns within the dataset. Univariate analysis using histograms was applied to visualize the distributions of numerical and categorical variables, revealing, for instance, skewed distributions in temperature and PM10 levels [24]. Bivariate analysis was performed to identify relationships between key variable pairs, highlighting strong correlations between NO2 and CO concentrations [25]. Temporal analysis was employed to explore trends in weather and air quality variables over the study period, showing seasonal variations in temperature, humidity, and pollutant concentrations. Python with scikit-learn was used for data processing and analysis, while AMOS facilitated structural equation modeling.

2.6. Selection of Features

Feature selection was conducted in two stages. First, multicollinearity was assessed using Pearson correlation plots, and for highly correlated pairs (correlation > 0.7), only one feature was retained [26, 27]. Second, iterative model-based procedures were applied to evaluate feature importance, with ensemble models used to identify variables that could mask the contribution of others [28, 29]. Based on these analyses, less informative or over-generalized features-including 'bike sharing influencing factors', 'year', 'biking features', 'PM10', 'wind speed', 'cloud opacity', 'wind direction', 'solar radiation', 'atmospheric pressure', 'gender', 'minimum temperature', 'relative humidity', 'household income', 'education level', 'NO2', and 'SO2'-were excluded from model training.

2.7. Correlation Test

This test was used to assess the correlation between variables, with a strength ranging from 1 to -1: 1 indicates a positive correlation, 0 indicates no correlation, and -1 indicates a negative correlation.

2.8. Modelling

• **Logistic Regression:** This model is known for classification problems, where it models the probability that a given variable belongs to a particular class with values between 0 and 1 of the response variables [30]. This research used a multi-class classification approach in which a binary classifier is trained for each class against all other classes separately; the predicted probabilities were obtained during the multi-class logistic regression model training phase [31].

• **Support Vector Machine (SVM):** This model is famous for classification analysis. The SVM was chosen for its effectiveness with high-dimensional data [32].

• **Random Forest Model**: A random subset of the training data is constructed using each tree and a random subset of all features. In this research, with replacements from the original dataset, a random forest was implemented by taking random samples to create a bootstrapped dataset of the same size. Subsequently, at each split in a decision tree, a subset of features was randomly selected from the complete set of features [33].

• **Stacking Classifier:** Stacking is an ensemble learning technique that combines multiple machine learning models to obtain predictions that are used as input features for a meta-classifier [33]. This research used a stacking classifier to fit several baseline models, including KNN, Naïve Bayes, SVM, and Logistic Regression [34, 35]. Then, the predictions of those classifiers were used to train and generate predictions.

• **Structural Equation Modeling (SEM):**It is a powerful statistical approach used to explore complex relationships among observed and latent variables. By combining aspects of factor analysis and multiple regression within a single framework, SEM allows researchers to test theoretical models that incorporate both measurement and structural components. SEM is particularly prevalent in disciplines such as education and the social sciences because it can model latent constructs and evaluate hypothesized causal pathways [36-38].

Structural Equation Modeling (SEM) was used to evaluate the fit of the proposed model to the observed data [39]. Model fit was assessed using the Comparative Fit Index (CFI), Tucker–Lewis Index (TLI), and Root Mean Square Error of Approximation (RMSEA), with commonly accepted thresholds of CFI/TLI ≥ 0.95 and RMSEA ≤ 0.06 [40-42].

2.9. Model Metrics

The ensemble machine learning models were evaluated through cross-validation, employing the accuracy metric to gauge their predictive performance. The outcome of each fold's accuracy scores was stored, and the mean and standard deviation were printed on the model's effectiveness [43]. In parallel, the structural equation model (SEM) was evaluated using multiple model fit indices, such as the Root Mean Square Error, Comparative Fit Index, Tucker-Lewis Index, Chi-square statistics, and Normed Fit Index. These SEM metrics complemented the machine learning evaluations by validating the structural relationships between latent constructs. Together, the combination of predictive metrics from ensemble learning and explanatory metrics offered a robust framework for understanding and predicting biking demand [40]. Figure **1** shows how each method was used.

3. RESULTS

3.1. Descriptive Statistics

Based on a total of 6,386 observations, the recorded maximum temperature ranged from 25.8°C to 61.2°C, with a mean of 38.4°C, reflecting both the central tendency and variability of this key environmental variable. Rainfall levels were relatively low, with an average of 2.6 mm and a maximum of 21.2 mm, indicating limited precipitation in the study area. In terms of air quality, the average ozone concentration was 20.8, while PM2.5 levels averaged 42.8, peaking at 127.2, signaling considerable air pollution across the study period.

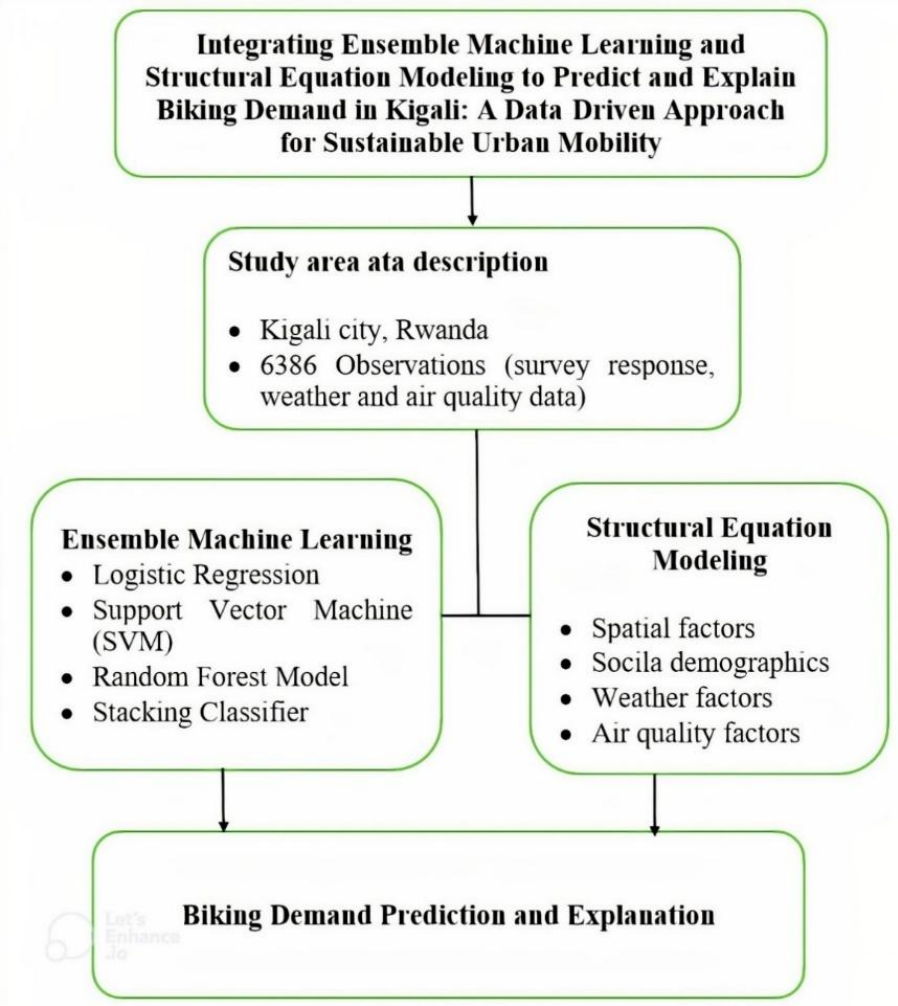

Fig. (1). Conceptual framework.

Table 3. Classification report (logistic regression model).

	Traditional Statistical Models		Ensemble Models		
	Logistic regression	SVM	Random forest	XGB	Stacking classifier
Accuracy	42%	82%	98%	99%	94%
Macro avg (**precision)**	42%	68%	99%	99%	87%
Weighted avg(**precision)**	71%	88%	98%	99%	94%

3.2. Modelling Results

The findings in Table **3** demonstrate that ensemble machine learning models-Random Forest, XGBoost, and Stacking-significantly outperform traditional methods in predicting biking demand. While logistic regression achieved only 42% accuracy and SVM 82%, ensemble models reached up to 99% accuracy and high precision, effectively capturing complex, nonlinear biking behavior. Their reliability makes them valuable for optimizing bike-sharing systems, guiding infrastructure planning, and supporting policies that promote active transport.

3.3. Features Associated with Demand

The Random Forest model results shown in Fig. (**2**) indicate that environmental factors, marital status, age, monthly income, and household number are the most critical factors in predicting biking demand in Kigali. These insights help urban planners and mobility service providers to prioritize air quality management, weather-adaptive strategies, and seasonal planning to better align bike sharing operations with user demand.

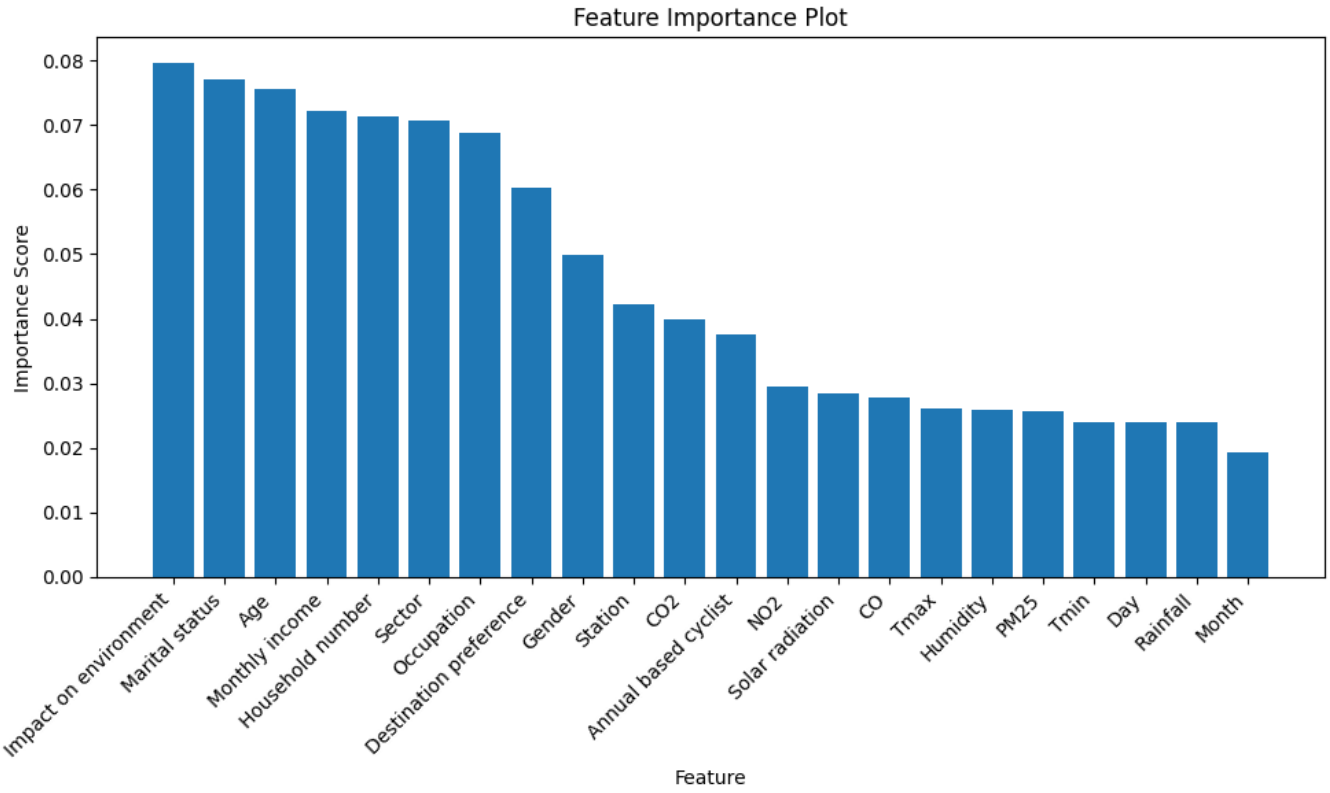

Fig. (2). Feature importance.

The model fit indices (Table **4**) indicate that the structural equation model for biking demand is statistically robust. Despite a significant χ^2 value due to the large sample size (n = 6,386), other indices demonstrate good fit: RMSEA = 0.036, CFI = 0.932, and TLI = 0.910. These results suggest that the model adequately captures the relationships between observed and latent variables influencing biking demand.

The results in Table **5** present the standardized regression weights for the main factors influencing ride frequency, offering key insights into what drives biking demand.

• The most significant predictor is spatial factors (β = 0.878, $p < 0.001$), which show a strong and statistically significant positive relationship with ride frequency. This suggests that proximity to bike sharing stations, road connectivity, and availability of biking infrastructure have a substantial impact on how frequently people use bikes.

• On the other hand, sociodemographic variables, such as household income (β = -0.001, $p = 0.879$), age, and occupation, as well as environmental factors (β = 0.002, $p = 1.000$), and time-of-day ride preferences (β = -0.006, $p = 0.347$) do not show a significant influence on ride frequency.

Table **6** shows the extent to which the model explains variance in key variables, indicating its predictive strength. Ride frequency has a high explained variance ($R^2 = 0.771$), demonstrating strong predictive capacity, particularly due to spatial-related factors. Accessibility and rainfall show exceptionally high explained variance ($R^2 = 0.990$), highlighting their significant influence on biking behavior. Other variables, such as education level and age, exhibit moderate explained variance, suggesting a lesser but still relevant impact.

Table **7** shows how latent socio-demographic factors relate to observed indicators like age, education, and employment in explaining biking demand. Age and education are most influential, but overall socio-demographics have a limited effect on ride frequency. Individuals with primary education bike mostly for work, while those with secondary education exhibit diverse purposes, including employment, studies, unemployment, and retirement.

In Fig. (**3**), the model demonstrates that biking frequency is not determined by a single factor but by a network of interrelated influences, which machine learning cannot provide in feature importance. The integration of latent variables helps capture underlying constructs that are not directly measurable but significantly shape cycling behavior.

Table 4. Model fit indices.

Fit Index	Value	Recommended Threshold	Interpretation
Chi-square (χ^2)	784.440	Low (non-significant ideal)	Acceptable with large sample
Degrees of Freedom (df)	69	-	-
χ^2/df	11.369	< 5 (ideal), < 3 (good)	High, but tolerable
RMSEA	0.036	< 0.08 (acceptable), < 0.05 (good)	Good fit
CFI	0.932	≥ 0.90	Acceptable fit
TLI	0.910	≥ 0.90	Acceptable fit
NFI	0.926	≥ 0.90	Acceptable fit
PCFI	0.706	> 0.50	Parsimonious model
Hoelter (.05)	893	> 200	Strong sample size adequacy

Table 5. Standardized regression weights (main paths to ride frequency).

Path	Standardized Estimate (β)	Significance (P)
Ride Frequency ← Spatial Factors	**0.878**	***
Ride Frequency ← SOCIODEMOGRAPHICS	-0.005	0.526
Ride Frequency ← Environmental Factors	0.002	1.000
Ride Frequency ← Time of Day Ride Pref	-0.006	0.347
Ride Frequency ← Gender	-0.013	0.045
Ride Frequency ← Household Income	-0.001	0.879

Table 6. Squared multiple correlations (R^2 values).

Variable	R^2 (Explained Variance)
Ride Frequency	**0.771 (77.1%)**
Education Level	0.425
Age of respondents	0.372
Employment Status	0.006
ZAccess Perception	0.990
Z_Rainfall Gitega	0.990
Other variables	0.000

Table 7. Latent constructs and their indicators.

Latent Variable	Observed Indicators	Notes
Socio-demographics	Education Level, Age, Employment Status	Strong Age & Education loadings
Spatial Factors	Distance to Station, Infrastructure Quality, Access Perception	Access Perception is very strong (β = .995)
Environmental Factors	Rainfall, Temperature, Sunny Weather	Rainfall strong (β = .995), others weak

The high performance in ensemble machine learning models in Table **8** is broadly consistent with the results reported in previous research. Random Forest model in this research achieved 98% accuracy and a 99.8% AUC, outperforming the accuracies reported in previous studies [44, 45]. In terms of AUC, the Random Forest model in this research outperformed another research [46]. This Support Vector Machine (SVM) model recorded 82% accuracy and 84% AUC, lower compared to a previous study [47]. In terms of AUC, this stacking model outperformed the 90.4% benchmark reported by a recent study [48]. This stacking classifier attained 94% accuracy and a 98% AUC, which compares favorably with the 94.4% and 97% accuracies reported recently [49, 50], respectively. Despite achieving high accuracy in other studies [51, 52], the study successfully incorporated features but failed to report the relationships among constructs, which highlights a limitation of the machine learning approach.

Table 8. Comparison of the findings with existing materials.

Model	Random Forest	SVM	Stacking Classifier
Accuracy	• 90% [44] • 82% [45]	• 99.41% [47] • 97.17% [22]	• 94.4% [49] • 97% [50][REMOVED REF FIELD].
Our Findings	**98%**	**82%**	**94%**
AUC	98.9% [46]	100% [47]	90.4% [48]
Our Findings	**99.8%**	**84%**	**98%**

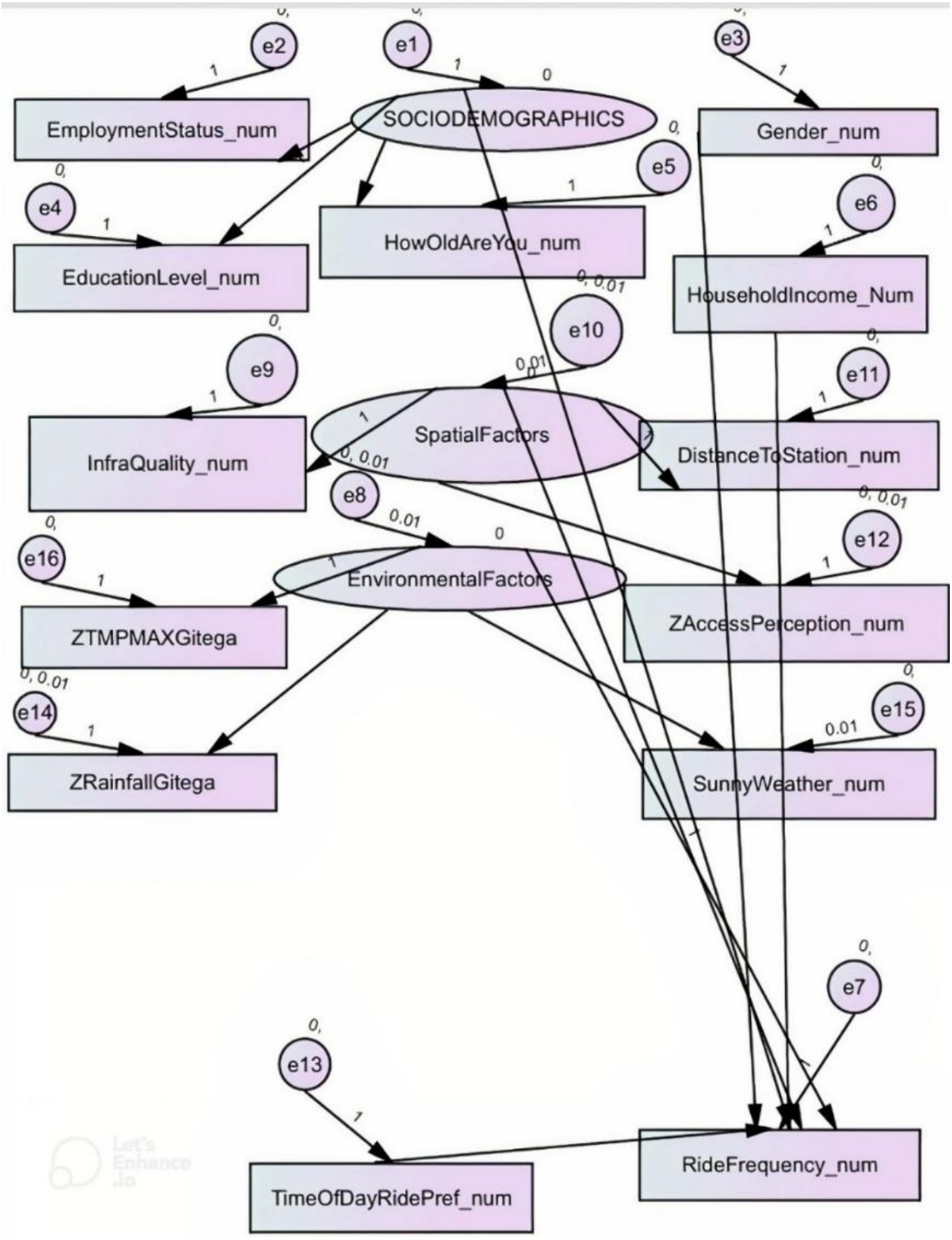

Fig. (3). SEM path diagram for observed variables, latent variables, and error terms.

4. DISCUSSION

• The findings indicated that ensemble machine learning models outperformed traditional statistical models in predicting bike-sharing demand. Logistic Regression achieved 42% accuracy and Support Vector Machine 82%, whereas XGBoost, Random Forest, and Stacking Classifier reached 99%, 98%, and 94%, respectively. This demonstrates the superior ability of ensemble models to capture nonlinear and interactive relationships among variables, along with automatic feature importance ranking, which traditional models cannot achieve.

• Despite their predictive strength, ensemble models failed to reveal causal pathways and latent behavioral constructs, whereas Structural Equation Modeling (SEM) effectively captured them. Feature importance analysis highlighted spatial and environmental factors-particularly distance to bike stations, infrastructure quality, and perceived accessibility-as key determinants of biking demand. High loadings on perceived access and rainfall suggest that accessibility perception and adverse weather conditions strongly influence biking frequency.

• SEM results (RMSEA = 0.036, CFI = 0.932, TLI = 0.910) reinforced the robustness of these relationships, revealing spatial factors ($\beta = 0.878$, $p < 0.001$) as the only significant predictors of ride frequency. This contrasts with prior studies emphasizing sociodemo-graphic influences such as age, gender, and income. The high explanatory power ($R^2 = 0.771$) further indicates that accessibility-related variables, rather than user characteristics, predominantly drive bike-sharing participation in Kigali.

• By combining ensemble machine learning (ML) models with Structural Equation Modeling (SEM), ensemble models captured complex, non-linear relationships and key predictors, while SEM clarified behavioral pathways and causal mechanisms. These findings corroborate prior studies showing that ensemble ML outperforms trad-itional statistical approaches [53, 54]. Unlike previous research, this study identified accessibility and weather as dominant determinants in Kigali, offering a novel, context-specific framework that unites predictive accuracy with interpretive depth-valuable for sustainable urban mobility planning and policy. While these results provide actionable insights, the study is limited by its focus on Kigali and by its reliance on cross-sectional survey data, which may not capture seasonal or long-term behavioral changes.

• Additionally, this research has demonstrated how the integration of AI, behavioral modeling, and SDG-aligned urban planning provides both predictive power and policy insight for advancing sustainable mobility-a practical manifestation of the interdisciplinary sustainability goals explored across the following four studies [55-58].

CONCLUSION

This study demonstrated that ensemble machine learning models, particularly Random Forest and Stacking Classifier, outperform traditional statistical models in predicting bike-sharing demand in Kigali, achieving up to 98% and 99.8% accuracy. These models effectively capture complex, non-linear relationships among biking behavior, spatial factors, and environmental conditions, which traditional models are unable to capture. However, ensemble models alone cannot reveal causal pathways or latent behavioral constructs, which Structural Equation Modeling (SEM) successfully elucidated. SEM results identified spatial accessibility, perceived access, and rainfall as dominant determinants of biking frequency, emphasizing the importance of infrastructure quality and weather sensitivity. By integrating ensemble prediction with SEM-based causal interpretation, this research provides both predictive accuracy and explanatory depth. These findings advance sustainable urban transportation and offer a robust methodological framework linking data-driven demand forecasting with behavioral understanding, guiding evidence-based interventions, such as weather-adaptive operations, infrastructure expansion, and accessibility-focused planning. This approach not only informs local policy in Kigali but also establishes a transferable model for other urban contexts, advancing theory and practice in sustainable mobility.

AUTHORS' CONTRIBUTIONS

The authors confirm contribution to the paper as follows: J.M.V.N.: Written the paper under the supervision of Hannibal Bwire and Alphonse Nkurunziza.

LIST OF ABBREVIATIONS

CFI	=	Comparative Fit Index
CO	=	Carbon Monoxide
CO2	=	Carbon Dioxide
DRC	=	Democratic Republic of the Congo
EDA	=	Exploratory data analysis
KNN	=	K-nearest Neighbors
ML	=	Machine Learning
NMT	=	Non-Motorized Transport
NO_2	=	Nitrogen Dioxide
O_3	=	Ozone
PM10	=	Particulate Matter with particles measuring 10 micrometers
PM2.5	=	Particulate matter with particles measuring 2.5 micrometers
RMSEA	=	Root Mean Square Error of Approximation
SDGs	=	United Nations Sustainable Development Goals
SEM	=	Structural Equation Modeling
SO_2	=	Sulfur Dioxide
SVM	=	Support Vector Machines
TLI	=	Tucker–Lewis Index
XGB	=	Extreme Gradient Boosting

CONSENT FOR PUBLICATION

Not applicable.

AVAILABILITY OF DATA AND MATERIALS

The data supporting the findings of this study are available from the corresponding author [J.M.V.N.] upon reasonable request.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

REFERENCES

[1] E. Cascetta, and I. Henke, "The seventh transport revolution and the new challenges for sustainable mobility", *J. Urban Mobil.,* vol. 4, p. 100059, 2023. [http://dx.doi.org/10.1016/j.urbmob.2023.100059]

[2] J.J. Lin, and J.J. Wu, "Association of bike-sharing service with gentrification in a transit-rich city: A catalyst or an outcome?", *Transp. Res. Interdiscip. Perspect.,* vol. 22, p. 100941, 2023. [http://dx.doi.org/10.1016/j.trip.2023.100941]

[3] H. Bwire, A. Nkurunziza, and A. Nkurunziza, "Enhancing connectivity via GIS-based bike-sharing optimization in Kigali City, Rwanda", *J. Environ. Earth Sci.,* vol. 7, no. 8, pp. 191-206, 2025. [http://dx.doi.org/10.30564/jees.v7i8.10565]

[4] P.K. Alimo, S. Agyeman, A. Danesh, C. Yu, and W. Ma, "Is public bike-sharing feasible in Ghana? Road users' perceptions and policy interventions", *J. Transp. Geogr.,* vol. 106, p. 103509, 2023. [http://dx.doi.org/10.1016/j.jtrangeo.2022.103509]

[5] C.J. Mendiate, J.A. Soria-lara, and A. Monzon, "Identifying clusters of cycling commuters and travel patterns: The case of Quelimane, Mozambique", *Int. J. Sustain. Transport.,* vol. 14, no. 9, pp. 710-721, 2020. [http://dx.doi.org/10.1080/15568318.2020.1774947]

[6] C. Kapuku, S.H. Park, and S.H. Cho, "Modeling the intermodality between public transport and bike-sharing using smartcard trip Chain data", *Int. J. Urban Sci.,* vol. 28, no. 3, pp. 452-478, 2024. [http://dx.doi.org/10.1080/12265934.2024.2312284]

[7] C. Gao, and Y. Chen, Using machine learning methods to predict demand for bike sharing.*Information and Communication Technologies in Tourism 2022,* Springer International Publishing: Cham, Switzerland, 2022, pp. 282-296. [http://dx.doi.org/10.1007/978-3-030-94751-4_25]

[8] M. He, X. Ma, J. Wang, and M. Zhu, "Geographically weighted multinomial logit models for modelling the spatial heterogeneity in the bike-sharing renting-returning imbalance: A case study on Nanjing, China", *Sustain Cities Soc.,* vol. 83, p. 103967, 2022. [http://dx.doi.org/10.1016/j.scs.2022.103967]

[9] A. Ali, R. Jayaraman, E. Azar, and M. Maalouf, "A comparative analysis of machine learning and statistical methods for evaluating building performance: A systematic review and future benchmarking framework", *Build. Environ.,* vol. 252, p. 111268, 2024. [http://dx.doi.org/10.1016/j.buildenv.2024.111268]

[10] H. S. R. Rajula, G. Verlato, M. Manchia, N. Antonucci, and V. Fanos, "Comparison of conventional statistical methods with machine learning in medicine: Diagnosis, drug development, and treatment", *Medicina,* vol. 56, no. 9, p. 455, 2020. [http://dx.doi.org/10.3390/medicina56090455] [PMID: 32911665]

[11] N. Shrestha, "Detecting multicollinearity in regression analysis", *Am. J. Appl. Math. Stat.,* vol. 8, no. 2, pp. 39-42, 2020. [http://dx.doi.org/10.12691/ajams-8-2-1]

[12] X. Yang, "Application of visual multimedia technology in air transportation information service system", *Appl. Math. Nonlin. Sci.,* vol. 9, no. 1, p. 20230959, 2024. [http://dx.doi.org/10.2478/amns.2023.2.00959]

[13] Z. Kun, F. Weibing, and W. Jianlin, "Default identification of P2P lending based on stacking ensemble learning", *2nd International Conference on Economic Management and Model Engineering (ICEMME)* Chongqing, China, 20-22 Nov. 2020, pp. 992-1006. [http://dx.doi.org/10.1109/ICEMME51517.2020.00203]

[14] S.V. e, J. Park, and Y. Cho, "Using data mining techniques for bike sharing demand prediction in metropolitan city", *Comput. Commun.,* vol. 153, pp. 353-366, 2020. [http://dx.doi.org/10.1016/j.comcom.2020.02.007]

[15] X. Han, R. Zhang, and Z. Li, "Research on construction and contribution analysis of demand forecasting model based on GBDT-LightGBM algorithm", *IEEE International Conference on Image Processing and Computer Applications (ICIPCA)* Changchun, China, 11-13 Aug. 2023, pp. 359-364. [http://dx.doi.org/10.1109/ICIPCA59209.2023.10257726]

[16] H. Kang, and J.W. Ahn, "Model setting and interpretation of results in research using structural equation modeling: A checklist with guiding questions for reporting", *Asian Nurs. Res.,* vol. 15, no. 3, pp. 157-162, 2021. [http://dx.doi.org/10.1016/j.anr.2021.06.001] [PMID: 34144201]

[17] I. Gubic, and O. Baloi, "Public open space initiatives for healthier cities in Rwanda", *J. Public Space,* vol. 5, no. 2, pp. 129-146, 2020. [http://dx.doi.org/10.32891/jps.v5i2.1287]

[18] "Population and housing census report", Available from: https://www.statistics.gov.rw/sites/default/files/documents/2025-02/RPHC5%20Thematic%20Report_Agriculture.pdf

[19] "Cycling city Kigali sprints to promote smart and green mobility", Available from: https://healthsojo-africa.org/cycling-city-kigali-sprints-to-promote-smart-and-green-mobility-2/

[20] "Rwanda launches Africa's first public bike-share transport system", Available from: https://www.sisiafrika.com/rwanda-launches-africas-first-public-bike-share-transport-system/

[21] S. VE, and Y. Cho, "A rule-based model for Seoul Bike sharing demand prediction using weather data", *Eur. J. Remote Sens.,* vol. 53, no. sup1, pp. 166-183, 2020. [http://dx.doi.org/10.1080/22797254.2020.1725789]

[22] G. Cao, L.A. Zhou, C. Liu, and J. Zhou, "The effects of the entries by bike-sharing platforms on urban air quality", *China Econ. Quart. Int.,* vol. 3, no. 3, pp. 213-224, 2023. [http://dx.doi.org/10.1016/j.ceqi.2023.09.003]

[23] N.R. Njeri, "Data preparation for machine learning modelling", *Int. J. Comp. Appl. Tech. Res.,* vol. 11, no. 6, pp. 231-235, 2022. [http://dx.doi.org/10.7753/IJCATR1106.1008]

[24] A. S. Rao, B. V. Vardhan, and H. Shaik, "Role of exploratory data analysis in data science", *6th International Conference on Communication and Electronics Systems (ICCES)* Coimbatre, India, 08-10 July 2021, pp. 1457-1461. [http://dx.doi.org/10.1109/ICCES51350.2021.9488986]

[25] K. Guo, L. Lei, H. Song, Z. Ji, and L. Liu, "Co-response of atmospheric NO2 and CO2 concentrations from satellites observations of anthropogenic CO2 emissions for assessing the synergistic effects of pollution and carbon reduction", *Remote Sens.,* vol. 17, no. 5, p. 739, 2025. [http://dx.doi.org/10.3390/rs17050739]

[26] A.G. Dufera, T. Liu, and J. Xu, "Regression models of Pearson correlation coefficient", *Stat. Theory Relat. Fields,* vol. 7, no. 2, pp. 97-106, 2023. [http://dx.doi.org/10.1080/24754269.2023.2164970]

[27] H. Yu, and A.D. Hutson, "Inferential procedures based on the weighted Pearson correlation coefficient test statistic", *J. Appl. Stat.,* vol. 51, no. 3, pp. 481-496, 2024.

[http://dx.doi.org/10.1080/02664763.2022.2137477] [PMID: 38370269]

[28] B. Wickramanayake, C. Ouyang, Y. Xu, and C. Moreira, "Generating multi-level explanations for process outcome predictions", *Eng. Appl. Artif. Intell.,* vol. 125, p. 106678, 2023. [http://dx.doi.org/10.1016/j.engappai.2023.106678]

[29] S. Leem, J. Oh, J. Moon, M. Kim, and S. Rho, "Enhancing multistep-ahead bike-sharing demand prediction with a two-stage online learning-based time-series model: Insight from Seoul", *J. Supercomput.,* vol. 80, no. 3, pp. 4049-4082, 2024. [http://dx.doi.org/10.1007/s11227-023-05593-6]

[30] T. Zhou, K.M.Y. Law, and K.L. Yung, "An empirical analysis of intention of use for bike-sharing system in China through machine learning techniques", *Enterprise Inf. Syst.,* vol. 15, no. 6, pp. 829-850, 2021. [http://dx.doi.org/10.1080/17517575.2020.1758796]

[31] N. Axford, "Logistic regression", *Medsurg Nurs.,* vol. 29, no. 5, p. 353, 2008.

[32] C.E. Widodo, K. Adi, and R. Gernowo, "A support vector machine approach for identification of pleural effusion", *Heliyon,* vol. 10, no. 1, p. e22778, 2024. [http://dx.doi.org/10.1016/j.heliyon.2023.e22778] [PMID: 38268601]

[33] V.E. Sathishkumar, and Y. Cho, "Season wise bike sharing demand analysis using random forest algorithm", *Comput. Intell.,* vol. 40, no. 1, p. e12287, 2024. [http://dx.doi.org/10.1111/coin.12287]

[34] S. Mohapatra, S. Maneesha, S. Mohanty, P.K. Patra, S.K. Bhoi, K.S. Sahoo, and A.H. Gandomi, "A stacking classifiers model for detecting heart irregularities and predicting Cardiovascular Disease", *Healthc. Anal.,* vol. 3, no. 100133, p. 100133, 2023. [http://dx.doi.org/10.1016/j.health.2022.100133]

[35] W. Zhang, H. Li, L. Han, L. Chen, and L. Wang, "Slope stability prediction using ensemble learning techniques: A case study in Yunyang County, Chongqing, China", *J. Rock Mech. Geotech. Eng.,* vol. 14, no. 4, pp. 1089-1099, 2022. [http://dx.doi.org/10.1016/j.jrmge.2021.12.011]

[36] A. Robitzsch, "Model-robust estimation of multiple-group structural equation models", *Algorithms,* vol. 16, no. 4, p. 210, 2023. [http://dx.doi.org/10.3390/a16040210]

[37] Y. Liao, H. Guo, and X. Liu, "A study of young people's intention to use shared autonomous vehicles: A quantitative analysis model based on the extended TPB-TAM", *Sustainability,* vol. 15, no. 15, p. 11825, 2023. [http://dx.doi.org/10.3390/su151511825]

[38] B. Zhou, J. Jin, H. Huang, and Y. Deng, "Exploring the macro economic and transport influencing factors of urban public transport mode share: A bayesian structural equation model approach", *Sustainability,* vol. 15, no. 3, p. 2563, 2023. [http://dx.doi.org/10.3390/su15032563]

[39] M. Maxfield, K. Courtney, S. Assuras, J.J. Manly, and C.S. Widom, "Childhood maltreatment and cognitive functioning from young adulthood to late midlife: A prospective study", *Neuropsychology,* 2025. [http://dx.doi.org/10.1037/neu0001042] [PMID: 41100283]

[40] G. Dash, and J. Paul, "CB-SEM vs PLS-SEM methods for research in social sciences and technology forecasting", *Technol. Forecast. Soc. Change,* vol. 173, p. 121092, 2021. [http://dx.doi.org/10.1016/j.techfore.2021.121092]

[41] H.E. Al Issa, and M.K. Abdelsalam, "Antecedents to leadership: A CB-SEM and PLS-SEM validation", *Int. J. Sustain. Dev. Plan.,* vol. 16, no. 8, pp. 1403-1414, 2021. [http://dx.doi.org/10.18280/ijsdp.160801]

[42] D. McNeish, and M.G. Wolf, "Dynamic fit index cutoffs for one-factor models", *Behav. Res. Methods,* vol. 55, no. 3, pp. 1157-1174, 2022. [http://dx.doi.org/10.3758/s13428-022-01847-y] [PMID: 35585278]

[43] K. Pham, D. Kim, S. Park, and H. Choi, "Ensemble learning-based classification models for slope stability analysis", *Catena,* vol. 196, p. 104886, 2021. [http://dx.doi.org/10.1016/j.catena.2020.104886]

[44] C. Kapuku, S.Y. Kho, D.K. Kim, and S.H. Cho, "Assessing and predicting mobility improvement of integrating bike-sharing into multimodal public transport systems", *Transp. Res. Rec.,* vol. 2675, no. 11, pp. 204-213, 2021. [http://dx.doi.org/10.1177/03611981211045071]

[45] N.S. Hadjidimitriou, M. Lippi, and M. Mamei, "Activity imputation of shared e-bikes travels in urban areas", In: *Machine Learning, Optimization, and Data Science,* Cham: Springer, 2022, pp. 442-456. [http://dx.doi.org/10.1007/978-3-030-95467-3_32]

[46] J.K. Afriyie, K. Tawiah, W.A. Pels, S. Addai-Henne, H.A. Dwamena, E.O. Owiredu, S.A. Ayeh, and J. Eshun, "A supervised machine learning algorithm for detecting and predicting fraud in credit card transactions", *Decis. Anal. J.,* vol. 6, p. 100163, 2023. [http://dx.doi.org/10.1016/j.dajour.2023.100163]

[47] Y.A. Singgalen, "Performance analysis of NBC, DT, and SVM algorithms in data classification visitor reviews of borobudur temple based on CRISP-DM", *Build. Info. Tech. Sci.,* vol. 4, no. 3, pp. 1634-1646, 2022. [http://dx.doi.org/10.47065/bits.v4i3.2766]

[48] N. Kardani, A. Zhou, M. Nazem, and S.L. Shen, "Improved prediction of slope stability using a hybrid stacking ensemble method based on finite element analysis and field data", *J. Rock Mech. Geotech. Eng.,* vol. 13, no. 1, pp. 188-201, 2021. [http://dx.doi.org/10.1016/j.jrmge.2020.05.011]

[49] N. El-Rashidy, S. El-Sappagh, T. Abuhmed, S. Abdelrazek, and H.M. El-Bakry, "Intensive care unit mortality prediction: An improved patient-specific stacking ensemble model", *IEEE Access,* vol. 8, pp. 133541-133564, 2020. [http://dx.doi.org/10.1109/ACCESS.2020.3010556]

[50] N. Nayyer, N. Javaid, M. Akbar, A. Aldegheishem, N. Alrajeh, and M. Jamil, "A new framework for fraud detection in bitcoin transactions through ensemble stacking model in smart cities", *IEEE Access,* vol. 11, pp. 90916-90938, 2023. [http://dx.doi.org/10.1109/ACCESS.2023.3308298]

[51] J.M.V. Ntamwiza, and H. Bwire, "Predicting biking preferences in Kigali city: A comparative study of traditional statistical models and ensemble machine learning models", *Trans. Econ. Manag.,* vol. 3, pp. 78-91, 2025. [http://dx.doi.org/10.1016/j.team.2025.02.003]

[52] J. Singh, R. Singh, P. Ekvitayavetchanukul, P. Singh, M. Diwakar, and M. Avesh, "AI-driven innovations in tunnel construction and transport: Enhancing efficiency with advanced machine learning and robotics", *Open Transplant. J.,* vol. 19, no. 1, p. e26671212383139, 2025. [http://dx.doi.org/10.2174/0126671212383139250618044241]

[53] X. Zhang, and X. Zhao, "Machine learning approach for spatial modeling of ridesourcing demand", *J. Transp. Geogr.,* vol. 100, no. 103310, p. 103310, 2022. [http://dx.doi.org/10.1016/j.jtrangeo.2022.103310]

[54] S. Wang, B. Mo, Y. Zheng, S. Hess, and J. Zhao, "Comparing hundreds of machine learning and discrete choice models for travel demand modeling: An empirical benchmark", *Transportation Res. Part B Methodol.,* vol. 190, 2024. [http://dx.doi.org/10.1016/j.trb.2024.103061]

[55] W.H. Humaish, and A.H. Taher, "The effect of climate on water resources in Iraq using AI", *Sustain. Eng. Innov.,* vol. 7, no. 2, pp. 285-300, 2025. [http://dx.doi.org/10.37868/sei.v7i2.id533]

[56] A. Indaryanto, B.D. Harijadi, and E. Sinaga, "The growing use and impact of artificial intelligence technologies in the tourism industry", *Sustain. Eng. Innov.,* vol. 5, no. 2, pp. 189-204, 2023. [http://dx.doi.org/10.37868/sei.v5i2.id238]

[57] C. García Lirios, J.E. Crespo, J.M. Sepulveda, and T.J.H. Gracia, "Polarization networks around the SDGs in the press from 2020 to 2023", *Herit. Sustain. Dev.,* vol. 6, no. 1, pp. 67-76, 2024. [http://dx.doi.org/10.37868/hsd.v6i1.241]

[58] J.E. Crespo, C. García Lirios, S.S. Vélez Báez, I.C. Rincón

Rodríguez, J.E. Chaparro Medina, V.H. Meriño Córdoba, A. Sánchez Sánchez, C.Y. Quiroz Campas, and M.R. Molina González, "Corporate governance network around social responsibility and activism against the Sustainable Development Goals", *Herit. Sustain. Dev.*, vol. 6, no. 2, pp. 829-844, 2024. [http://dx.doi.org/10.37868/hsd.v6i2.823]