Scientific and Social Research, 2026, Volume 8, Issue 7
http://ojs.bbwpublisher.com/index.php/ssr
Online ISSN: 2981-9946
Print ISSN: 2661-4332

Predicting Campus-Space Suitability with Multi-Source Data and Seven Machine-Learning Models

Haopeng Li[1], Mustafa Muwafak Alobaedy[1]*, Mohd Nurul Hafiz Bin Ibrahim[1], Enci Chen[2]

[1]Faculty of Information Technology, City University Malaysia, Petaling Jaya 46100, Malaysia
[2]Information Center, Wenzhou Kean University, Wenzhou, 325060, China

Author to whom correspondence should be addressed.

Abstract: Campus planners still rely on surveys and static indicators that miss high-frequency behavioral change. The study developed a Space–Behavior–Decision framework for Wenzhou-Kean University using 243,062 anonymized half-hourly Wi-Fi records (July 2023–March 2025), road-integration for 30 buildings, and functional labels for 35 access points. Seven regression models were compared. Random forest achieved the highest accuracy (R^2 = 0.9998, RMSE = 0.0025, MAPE = 0.38 %). The decision tree remained nearly as accurate (R^2 = 0.9986, MAPE = 0.47 %) while offering transparency and easy GIS integration, making it preferable for routine planning. Feature importance gave weights of 0.48 (density), 0.32 (integration), and 0.20 (function). The high fit mainly shows the composite index is learnable. Quarterly updating and post-occupancy checks are required before operational use.

Keywords: Campus public space; Spatial suitability; Wi-Fi sensing; Machine learning; Random forest; Decision tree; Explainable AI

Online publication: August 14, 2026

1. Introduction

A campus is more than classrooms. Actual use shifts with timetable, weather, and student habits, yet planning still leans on functional zoning and surveys. Existing Wi-Fi networks already generate continuous occupancy signals. The difficulty is combining those signals with network accessibility and building function into one usable suitability score. This paper examines whether multi-source indicators can reproduce a composite suitability function, how seven regressors compare on accuracy and deployment cost, and how machine-learning weights can revise expert weights, using 243,062 records from Wenzhou-Kean University.

Previous smart-campus studies have demonstrated that Wi-Fi connection logs can reveal fine-grained spatial occupancy patterns without requiring intrusive sensing [1]. Campus-scale Wi-Fi traces have also been used to construct activity-based occupant models for energy and urban applications [2]. The correspondence between Wi-Fi connection counts and camera-based occupancy measurements provides further evidence

that network data can serve as a practical proxy for space use [3]. Recent reviews nevertheless emphasize that occupancy-prediction performance depends on sensing resolution, preprocessing, validation design, and transferability across buildings [4]. Edge-based transfer learning has been explored for classroom occupancy detection, while IoT-enabled smart-building research has highlighted the need to combine sensing, learning, and operational decision-making [5–6]. Anonymous Wi-Fi records have also supported person-count estimation in university-library settings [7]. These studies motivate the present integration of behavioral intensity, spatial accessibility, and functional attributes within one campus-space suitability framework.

2. Data and methods

The study covers July 2023–March 2025. Each of the 243,062 records counts unique MAC addresses at one access point in a 30-minute window. Only aggregated counts are used. As shown in **Table 1**, the analytical dataset is organized into four mutually linked layers rather than treated as a single undifferentiated data stream. The Wi-Fi layer contains 243,062 access-point-by-30-minute observations and provides the dynamic occupancy-density variable D. The spatial-network layer covers 30 buildings and supplies the road-integration variable I, whereas the functional layer maps 35 access-point locations to function-matching scores F. The target layer then assigns one composite suitability value S to every record, enabling the three predictors to be aligned at the same analytical unit.

Table 1. Data layers

Data layer	Scale	Core fields	Role
Wi-Fi records	AP × 30 min; 243,062 obs.	Unique MAC count	Occupancy density D
Spatial network	30 buildings	Road-integration index	Integration I
Functional attributes	35 AP locations	Core functions + public	Function matching F
Target	One value per record	Composite suitability S	Response variable

Road integration (I) and occupancy density (D) were min–max normalized. Function matching (F) was assigned from location–function relation. Original AHP weights (0.30/0.45/0.25) were re-estimated by random-forest importance as 0.32/0.48/0.20. The formula is $S = 0.32I + 0.48D + 0.20F$. Because S is built from its predictors, the experiment tests approximation of a known rule rather than external prediction. The sample was split 70/15/15. Seven models were trained and ranked by R^2, RMSE, MAE, and MAPE under two criteria: pure accuracy and a joint assessment of accuracy, interpretability, and GIS deployment cost. **Table 1** also clarifies the different spatial and temporal scales represented by the predictors. Density varies at half-hour intervals, while integration is relatively stable at the building-network level and function matching changes mainly when space use or access-point classification is revised. Combining these layers allows the model to distinguish a frequently occupied but poorly connected location from a well-connected space that is underused or functionally mismatched. At the same time, because S is constructed from I, D and F, the table makes explicit that the present experiment evaluates the reproducibility of a composite decision rule rather than prediction of an independent external outcome.

The emphasis on transparent spatial decision support is consistent with explainable machine-learning applications in BIM-supported building-performance analysis and interpretable prediction of building performance [8–9]. In addition, the use of road integration is grounded in space-syntax research showing that

spatial configuration affects accessibility, encounter opportunities and the sociability of public spaces [10]. These methodological precedents support the joint evaluation of predictive accuracy, interpretability and GIS deployment cost adopted in this study.

3. Results

As shown in **Table 2**, the seven models form a clear performance hierarchy on the held-out test set. Random forest achieves the highest R^2 (0.9998) and the lowest RMSE (0.0025), MAE (0.0017), and MAPE (0.38%), producing a mean rank of 1.00. AdaBoost records the second-best R^2, RMSE, and MAE, but its MAPE of 0.71% is higher than that of the decision tree. The decision tree therefore ranks third overall while retaining the second-lowest percentage error, an important result for applications in which relative error and rule transparency are both relevant.

Table 2. Test-set performance of the seven models

Rank	Model	Test R^2	RMSE	MAE	MAPE (%)	Mean rank
1	Random forest	0.9998	0.0025	0.0017	0.38	1.00
2	AdaBoost	0.9995	0.0038	0.0028	0.71	2.50
3	Decision tree	0.9986	0.0062	0.0031	0.47	2.75
4	XGBoost	0.9979	0.0078	0.0042	0.63	3.75
5	ANN	0.9876	0.0189	0.0147	2.78	5.00
6	LSTM	0.9449	0.0531	0.0361	8.36	6.00
7	LinearSVR	0.8607	0.0635	0.0469	11.56	7.00

Table 2 further shows that XGBoost remains highly accurate ($R^2 = 0.9979$), although its error values are consistently higher than those of the three leading tree-based models. The ANN maintains an R^2 of 0.9876, but its RMSE and MAPE rise to 0.0189 and 2.78%, respectively. The larger deterioration for LSTM and LinearSVR indicates that additional sequence complexity or a predominantly linear decision boundary does not provide an advantage for this three-variable composite target. Overall, the table supports two distinct conclusions: random forest is the strongest model for numerical fidelity, whereas the decision tree offers a more favorable balance between accuracy, interpretability, and implementation effort. **Figure 1** presents the same comparison graphically and makes the scale of the performance gaps easier to inspect.

Figure 1. Test-set performance of the seven models

Random forest ranked first on every metric. The decision tree's R^2 was only 0.0012 lower and its MAPE only 0.09 points higher, yet it supplies readable thresholds and runs quickly inside GIS. As shown in **Figure 1**, the four panels reveal that the leading models are tightly clustered in R^2 but become more clearly separated when error metrics are examined. Random forest has the shortest bars in all three error panels, confirming that its advantage is consistent rather than dependent on a single indicator. The figure also highlights a useful distinction between AdaBoost and the decision tree: AdaBoost performs better on RMSE and MAE, whereas the decision tree produces a lower MAPE. By contrast, LSTM and LinearSVR show visibly larger errors across the lower panels, confirming that their weaker R^2 values are accompanied by practically meaningful prediction deviations. This visual evidence motivates the metric-specific ranking analysis reported in **Figure 2**.

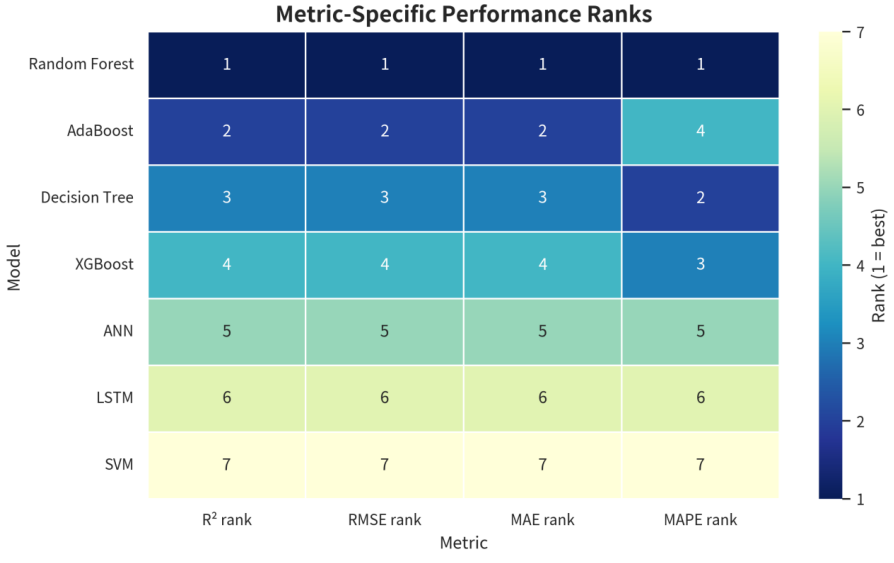

Figure 2. Rank heat map based on R^2, RMSE, MAE, and MAPE

As shown in **Figure 2**, random forest occupies rank 1 for R^2, RMSE, MAE, and MAPE, demonstrating complete consistency across the evaluation criteria. AdaBoost ranks second on R^2, RMSE, and MAE but falls to fourth on MAPE, while the decision tree ranks third on the first three metrics and rises to second on MAPE. XGBoost remains fourth on the absolute-error and fit metrics and third on MAPE. The heat map therefore prevents the overall mean rank from obscuring metric-specific strengths and shows why model selection should not rely on R^2 alone. The stable lower positions of ANN, LSTM, and LinearSVR also confirm that the ranking pattern is not caused by one anomalous metric. After establishing the model hierarchy, **Table 3** examines how the best-performing random forest redistributes importance among the three suitability indicators.

Table 3 compares the original analytic-hierarchy-process weights with the data-derived random-forest importance values. This comparison is used not to replace planning judgment automatically, but to identify where observed behavioral and spatial patterns support or challenge the initial expert weighting scheme.

Table 3. Objective weights from random-forest importance

Indicator	Original AHP	RF importance	Rounded weight
Occupancy density	0.45	0.4801	0.48
Road integration	0.30	0.3219	0.32
Function matching	0.25	0.1980	0.20

As shown in **Table 3**, occupancy density increases from an original weight of 0.45 to an importance value of 0.4801, a gain of 0.0301. Road integration increases from 0.30 to 0.3219, whereas function matching decreases from 0.25 to 0.1980. After rounding, the revised weights remain normalized at 0.48, 0.32, and 0.20. These shifts indicate that actual behavioral intensity contributes more to the learned suitability rule than initially assumed, while fixed functional labels provide less discrimination once density and accessibility are included. The result does not imply that function is unimportant; rather, it suggests that function matching should operate as a contextual constraint instead of the dominant driver of short-term suitability variation. **Figure 3** then evaluates whether the relative model performance remains stable from the validation set to the test set.

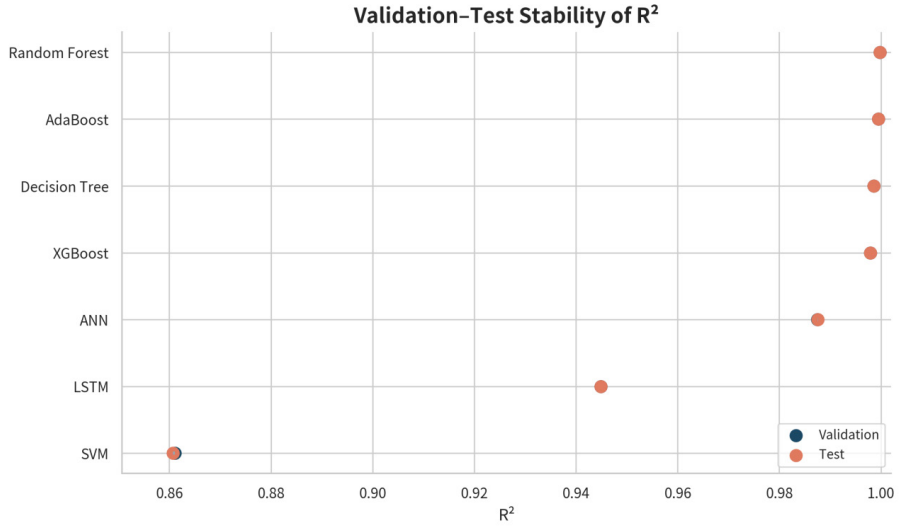

Figure 3. Validation-to-test stability of the seven models

As shown in **Figure 3**, the validation and test R^2 markers are nearly coincident for the four strongest tree-based models, which cluster close to 1.00. ANN also shows a relatively small validation-to-test displacement, although its overall fit is lower. LSTM and LinearSVR remain substantially behind the leading models in both partitions, indicating that their weak test performance is not an isolated split-specific event. The close validation-test agreement supports internal stability under the current random split; however, it should not be interpreted as evidence of temporal transferability because nearby half-hourly records may appear in different subsets. **Figure 4** complements this stability analysis by showing the magnitude and ordering of the final feature weights.

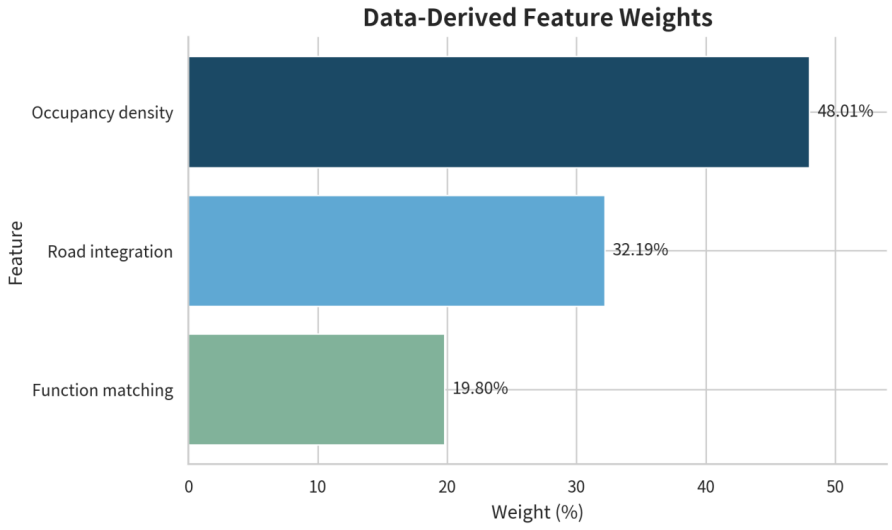

Figure 4. Objective weights derived from random-forest feature importance

Behavioral intensity and network accessibility together account for about 80 % of the learned importance. As shown in **Figure 4**, occupancy density contributes 48.01% of total importance, road integration contributes 32.19%, and function matching contributes 19.80%. Occupancy density is therefore about 1.5 times as influential as road integration and approximately 2.4 times as influential as function matching. Density and integration together account for 80.20% of the model-derived weight, indicating that observed use and network accessibility jointly dominate the suitability score. For planning practice, this hierarchy suggests that interventions should first examine whether a space is actually used and whether it is accessible, and then interpret those findings in relation to designated function. **Table 4** translates the quantitative model comparison into recommendations for different operational contexts.

Table 4. Recommended model by application context

Application context	Recommended model	Rationale
High-precision offline scoring	Random forest	Best values on all four metrics
Planning review & real-time GIS	Decision tree	Near-best accuracy, transparent, lightweight
Robust nonlinear alternatives	XGBoost / AdaBoost	Close performance, higher explanation cost

As shown in **Table 4**, the recommended model changes with the intended application rather than following a single universal ranking. Random forest is appropriate for high-precision offline scoring because

it provides the best values on all four metrics and can be used for periodic campus-wide recalculation. The decision tree is preferred for planning review and real-time GIS integration because its rules can be inspected, communicated to planners, and executed with low computational overhead. XGBoost and AdaBoost remain useful nonlinear alternatives when robustness is required, although their explanation and maintenance costs are higher. The table therefore defines a two-tier deployment strategy: random forest can serve as the high-accuracy reference engine, while the decision tree can support routine, auditable operational decisions. Agreement between the two models may be treated as a high-confidence result, whereas disagreement should trigger field verification or post-occupancy review.

4. Discussion

Tree models dominated because the target is a simple combination of three normalized predictors. The near-perfect R^2 must be read carefully: the target is built from the same variables, so high fit shows learnability rather than external validity. Density now leads among the learned weights, matching the emphasis on observed behavior. Integration remains useful for spotting accessibility–use mismatches. For day-to-day planning, the authors recommend the decision tree; random forest remains the offline high-precision engine. Both should be re-trained quarterly and checked against post-occupancy measures.

The need for operational verification is consistent with evidence that non-intrusive environmental variables can support machine-learning occupancy models but remain sensitive to building context and data quality [11]. Multimodal post-occupancy evaluation offers a suitable external validation route because model-derived suitability can be compared with observed movement, space-use patterns and user experience across campus learning environments [12]. Accordingly, the proposed scores should be treated as decision-support evidence: persistent high-suitability or low-suitability signals can guide inspection priorities, but consequential planning actions should be confirmed through field observations and periodic recalibration.

5. Limitations and outlook

Five limits remain: the target is constructed from the predictors; random splits allow nearby records into both sets; unique-MAC counts are not true person counts; function matching is coarse; and full hyper-parameter details are not preserved. Future work should use independent outcomes, temporal validation, and publish code and model cards.

6. Conclusion

Random forest gave the highest accuracy while the decision tree offered the best balance of accuracy, transparency, and deployment cost. Objective weights of about 0.48 / 0.32 / 0.20 place density first. The framework is suitable for diagnostic mapping and quarterly updating if treated as an approximation of a composite index. Linking model alerts with field observation can support an auditable planning cycle.

Disclosure statement

The authors declare no conflict of interest.

References

[1] Zhao Z, Wang T, Zhang Y, et al., 2023, Geo-Visualization of Spatial Occupancy on Smart Campus Using Wi-Fi Connection Log Data. ISPRS International Journal of Geo-Information, 12(11): 455.

[2] Mosteiro-Romero, M, Miller C, Quintana M, et al., 2023, Leveraging Campus-scale Wi-Fi Data for Activity-based Occupant Modeling in Urban Energy Applications. Journal of Physics: Conference Series, 2600(13): 132008.

[3] Alishahi N, Ouf MM, Nik-Bakht M, 2022, Using WiFi Connection Counts and Camera-based Occupancy Counts to Estimate and Predict Building Occupancy. Energy and Buildings, 2022(257): 111759.

[4] Li T, Liu X, Li G, et al., 2024, A Systematic Review and Comprehensive Analysis of Building Occupancy Prediction. Renewable and Sustainable Energy Reviews, 2024(193): 114284.

[5] Monti L, Tse R, Tang SK, et al., 2022, Edge-Based Transfer Learning for Classroom Occupancy Detection in a Smart Campus Context. Sensors, 22(10): 3692.

[6] Khan I, Zedadra O, Guerrieri A, et al., 2024, Occupancy Prediction in IoT-Enabled Smart Buildings: Technologies, Methods, and Future Directions. Sensors, 24(11): 3276.

[7] Hernando-Cánovas L, Martínez-Sala AS, Sánchez-Aarnoutse JC, & et al., 2025, A Machine Learning Approach for Estimating Person Counts Using Anonymous WiFi Data in a University Library. Sensors, 25(22): 7065.

[8] Shen Y, Pan Y, 2023, BIM-supported Automatic Energy Performance Analysis for Green Building Design Using Explainable Machine Learning and Multi-objective Optimization. Applied Energy, 2023(333): 120575.

[9] Li Y, Zhang H, Shen X, et al., 2025, Interpretable Machine Learning for Predicting and Optimizing Residential Building Performance in Cold Regions. Energy and Buildings, 2025(347): 116321.

[10] Askarizad R, Lamíquiz Daudén PJ, Garau C, 2024, The Application of Space Syntax to Enhance Sociability in Public Urban Spaces: A Systematic Review. ISPRS International Journal of Geo-Information, 13(7): 227.

[11] Banihashemi F, Weber M, Deghim F, et al., 2024, Occupancy Modeling on Non-Intrusive Indoor Environmental Data Through Machine Learning. Building and Environment, 2024(254): 111382.

[12] Guo Y, Sui J, 2025, Post-Occupancy Evaluation of Campus Learning Spaces with Multi-Modal Spatiotemporal Tracking. Buildings, 15(11): 1831.