Machine Learning Framework for Urban Green Infrastructure Site Suitability in Mumbai

OPEN ACCESS

Volume: 13

Special Issue: 2

Month: January

Year: 2026

E-ISSN: 2582-0397

P-ISSN: 2321-788X

Citation:
Vishwakarma, Seema, and Gayatri Venkatachalam. "Machine Learning Framework for Urban Green Infrastructure Site Suitability in Mumbai." *Shanlax International Journal of Arts, Science and Humanities*, vol. 13, no. 2, 2026, pp. 76–82.

DOI:
https://doi.org/10.34293/sijash.v13iS2-i3-Jan.10553

Seema Vishwakarma
Assistant Professor, Vidyalankar School of Information Technology
Wadala, Mumbai, Maharashtra, India

Gayatri Venkatachalam
Assistant Professor, SIWS College
Wadala, Mumbai, Maharashtra, India

Abstract
This paper offers a new machine learning framework to analyse site suitability of Urban Green Infrastructure (UGI) in the Mumbai area. Mumbai is one of the megacities of India struggling with environmental stress and limited green space. Through incorporation of spatial data which is GIS-based, the proposed framework integrates machine learning methods, compositing layers like NDVI, land use, PM 2.5, slope, and population density in order to identify the best locations for green infrastructure. The stepwise, flexible solution allows making decisions based on the data without dependence on conventional heuristics. Designed for scalability and interpretability, it supports urban planners in prioritising impactful greening interventions, advancing Mumbai's sustainability and resilience goals while offering a transferable model for other high-density urban regions.
Keywords: Machine Learning, Framework, Infrastructure, Urban Green

Introduction

Rapid urbanisation is changing cities all around the world. It is expected that 68% of the world's population will be living in cities by 2050 [1]. This will increase the pressure on livelihoods, infrastructure, and the climate. Green Infrastructure can be one solution to this—a network of natural spaces integrated with urbanisation that addresses problems such as poor AQI, pollution, ground-level heat, and climatic changes. Parks, gardens, wetlands, and trees provide multiple advantages. Green Infrastructure will also be beneficial in achieving sustainable development goals related to Climate Action and Sustainable Cities. Mumbai is one of the most populous cities of the world and the financial centre of India, and is an apt example of a city with an extreme crunch of Green Infrastructure.

The city has experienced extensive loss of green cover in recent decades. An analysis by the World Resources Institute found that by 2018, about 31% of Mumbai's land area had an average surface temperature above 30.5°C (a 174% increase in "hot" areas compared to 1988) [5], reflecting an intensifying urban heat island effect. At the same time, accessible open space for residents is grossly inadequate: Mumbai offers only on the order of 1 square metre of public green space per person, far below the 10–12 m² per capita recommended by urban planning norms [6]. This, coupled with its great population density, has contributed to environmental problems in

the city. Besides air pollution and heat stress, Mumbai also frequently floods during the monsoon season, in part because of wetlands and floodplains development [7].

It is important that the city increase its green space, improving living standards and strengthening climate resilience. The need to conserve natural spaces and increase green cover in urban areas has been put in focus by recent policy efforts, including Mumbai's climate action plan [8]. However, finding where green infrastructure can be grown in a properly located manner in a densely populated city like Mumbai is a challenging task. Due to limited land area and multitasking of urban land, planners must balance ecological standards against social and logistical factors. To establish the location of new parks and tree plantation GI projects, a systematic, information-driven approach is desirable.

Geographic Information Systems (GIS) and machine learning (ML) offer powerful tools to support this task. While ML can recognise complex trends and enhance decision-making on the basis of large data feeds, GIS facilitates multi-layered spatial, environmental, and urban analysis. Urban planners have the ability to evaluate potential greening interventions across the whole of Mumbai through GIS-ML green infrastructure site suitability analysis. This paper provides a conceptual machine learning framework for Mumbai urban green infrastructure site suitability. First, we examine pertinent research on the application of GIS and ML to site selection and urban green space planning. Next, we offer a unique, step-by-step framework specific to Mumbai, including information on data requirements, preprocessing, and feature engineering and modelling techniques. Lastly, we discuss how the framework impacts sustainable urban development in Mumbai. The objective is to present a theoretical but practical approach that city stakeholders could use to strategically improve Mumbai's green infrastructure in the face of growing urbanisation.

Literature Review

Scholars have recently resorted to the use of GIS-based multi-criteria analysis and machine learning methods to inform the planning of urban green infrastructure. Conventional methods usually use Multi-Criteria Decision Analysis (MCDA) in a GIS setting to assess land suitability for parks or other green land. Joy et al. developed a GIS-AHP (Analytic Hierarchy Process) approach to determine the best locations for urban green parks in Ranchi, India [9]. Various spatial criteria such as NDVI, land use/land cover (LULC), population density, air quality in the form of PM 2.5, and the intensity of urban heat islands were integrated with weights derived by experts to create a suitability map, although more than half of the area was found not suitable [10]. It is important to note that NDVI was the most powerful variable dictating suitability, followed by land use, population pattern, PM 2.5, and UHI advantage [11].

Other existing studies of MCDA using GIS (e.g., a study in Jimma, Ethiopia) also reflected a variety of criteria, including slope, land surface temperature (LST), distances to roads or water, and existing green areas, to map green space potential [12][13]. These studies repeatedly show that using a variety of geospatial layers in suitability models can successfully identify the areas with the greatest need or potential for greening. Whereas the GIS-MCDA method allows a more organised approach to the integration of expert knowledge and spatial information, the subjectivity of weighted criteria and non-changing decision rules can be a limitation.

Current studies have started to supplement and/or substitute the conventional MCDA with machine learning to enhance the quality and objectivity of green space suitability evaluation. Başeğmez et al. (2025) included ML algorithms in the process of selecting sites for sustainable urban green spaces in İzmir, Turkey [14]. Within their model, machine learning was employed to dynamically tune the weights of criteria rather than performing expert evaluation only [15]. They experimented with different models and discovered that a Random Forest algorithm was the most successful one with regard to consistency and predictive power of suitability mapping [15]. The study found that approximately 75% of current green places were not optimally placed based on the model, which meant that much could be done to enhance the distribution of green infrastructure [17].

Another developing use of machine learning in GI planning involves unsupervised learning to divide urban areas into zones and offer specific greening measures. Lin et al. (2025) used spatial multi-criteria evaluation combined with clustering analysis to plan green infrastructure in the Taipei metropolitan region [18]. They assessed each site in terms of different ecosystem services (e.g., heat reduction, stormwater management, carbon sequestration, habitat connectivity) and generated an initial priority map through GIS-based scoring [17]. Thereafter, they clustered similar areas using unsupervised ML and suggested particular GI intervention strategies to each cluster [18][12]. This integration of clustering provided context-sensitive knowledge, revealing that dense urban core clusters needed to be greened differently compared to peri-urban clusters [2].

The use of machine learning to map and predict green infrastructure in cities has also been advanced through remote sensing data. A thorough review by Dobrinć et al. (2025) recorded recent progress in urban GI mapping [4][3]. The review observed a recent tendency to apply deep learning (e.g., convolutional neural networks) to classify or segment urban greenery in high-resolution aerial and satellite images [11]. Among 55 GI mapping studies analysed, 33 used deep learning techniques and 22 used conventional machine learning, indicating increasing desire to use methods capable of automatically learning intricate visual representations of green infrastructure [18]. New sources of data, including multi-spectral Sentinel-2 satellite images and high-resolution commercial imagery, have significantly enhanced the capability to detect vegetation and even single tree canopies in dense urban environments [13][15].

Kooniyara (2025) showed in thesis work on Munich city that a GIS-based ML model trained on geospatial variables could forecast the potential of urban greening to plant trees [16]. A feed-forward neural network trained on a variety of above-ground geospatial variables was used to predict the appropriateness or potential utility of planting new trees at individual sites [7][8]. The ML model had high predictive power ($R^2 = 0.79$ on the test data), and Explainable AI methods (SHAP value analysis) were involved to explain the predictions and comprehend which factors had the greatest impacts on greening potential scores [9].

The literature in general is moving towards more integrated GIS-Machine Learning urban green infrastructure planning. A systematic review of urban GI assessments by Gorjian (2025) found that between 2020 and 2024, researchers have increasingly used data-driven approaches, with methodological advances such as better visualisation and scenario modelling [10]. Simultaneously, gaps remain, particularly inadequate study of social aspects of green infrastructure such as fair access, community desires, and participatory planning [10]. Gorjian (2025) underlines that future work should include residents' views and better definitions of GI benefits to secure outcomes that are based on community needs [2]. These insights highlight the importance of our proposed framework, which aims to exploit the power of machine learning while remaining community-responsive.

Author / Year	Study Location / Type	Method Used	Key Criteria / Data Used	Major Findings	Contribution to GI Planning
Joy et al. (2024)	Ranchi, India	GIS–AHP (MCDA)	NDVI, LULC, population density, PM2.5, UHI	Only 17% area highly suitable; NDVI most influential	Demonstrated MCDA effectiveness for GI suitability
Aiymeku et al. (2024)	Jimma, Ethiopia	GIS–MCDA	Slope, LST, NDVI, distance to roads/water, existing green	Identified resilience-based suitability zones	Showed GIS-MCDA supports climate-resilient GI planning
Başeğmez et al. (2025)	İzmir, Turkey	ML + AHP + WLC + TOPSIS	Multi-criteria with ML-optimized weights	Random Forest best; 75% existing GI suboptimal	ML reduces subjectivity, improves reliability
Lin et al. (2025)	Taipei Basin	GIS scoring + Clustering	Ecosystem services: heat, stormwater, carbon, habitat	Clustered areas for targeted GI strategies	Supports context-specific greening plans
Dobrinić et al. (2025)	Global Review	Deep learning & ML for GI mapping	Sentinel-2, high-resolution imagery, CNNs	33/55 studies used deep learning	Shows shift to AI for vegetation mapping
Kooniyara (2025)	Munich, Germany	Neural network + SHAP	Land use, population, distance to greenery	$R^2 = 0.79$; accurate greening potential prediction	Predictive ML aids future GI planning
Gorjian (2025)	Systematic Review	GIS + AI weighting	Multiple GI evaluation frameworks	Identified gaps: equity, participation	Calls for community-inclusive GI planning
CREDAI–MCHI (2025)	India–UK Comparative Report	Policy Review	Urban green space standards and resilience	Indian megacities far below norms	Shows policy gaps in GI provision

Table 1. Summary of Key Literature on GIS- and ML-Based Urban Green Infrastructure Planning

Figure 1 Illustration of GIS-ML Integration for Urban Green Infrastructure Planning

Proposed Framework

To assess the suitability of sites for urban green infrastructure, a machine learning-based adaptable and concise model used in Mumbai (UGI) is proposed. The GIS-based framework incorporates existing literary methods that have urban-specific characteristics. The framework starts with the gathering and combination of major spatial information on environmental, physical, and socio-urban conditions. Core layers include vegetation and land-use data (NDVI and LULC), climatic and environmental indicators (land surface temperature and PM concentration), topographic attributes (height and gradient), population, road accessibility and public transport, and stocks of available green and open areas like parks, wetlands, and mangroves. Additional helpful layers might contain proximity to water bodies, soil characteristics, and hazard or flood-risk areas. All datasets are geo-referenced to a common coordinate system and coordinated in a GIS environment.

Data Preprocessing and Feature Engineering

After compiling the raw data, preprocessing is performed to ensure data quality and compatibility for analysis. Each spatial layer is checked for errors, missing values, and outliers; noise such as cloud-affected NDVI values is removed, and missing socio-demographic data are filled using interpolation or auxiliary sources. All layers are converted to a common spatial extent and resolution (e.g., 30 m–100 m grid) to allow overlay and efficient city-wide computation, with high-resolution datasets resampled as required.

Input features with different units and ranges are normalised or standardised so that no single variable dominates the model; categorical land-use classes are encoded into suitable numerical forms. Additional spatial features, such as distance to roads, existing parks, or city centres, are generated using GIS tools to represent accessibility and green-space connectivity for suitability analysis.

It may be necessary to mask out or exclude certain areas entirely before modelling. Areas that are absolutely unavailable for conversion (dense built-up downtown, protected heritage sites, water bodies) can be given a suitability score of 0 or removed from the analysis mask so the model does not mistakenly identify them as high potential. In Mumbai, this might include the core business district or already saturated slum areas where interventions would require relocation. All preprocessing decisions and transformations should be documented, yielding a final analytic dataset where each spatial unit (e.g., each grid cell or land parcel) has a vector of features describing its attributes (NDVI, LST, slope, population, etc.). This dataset forms the input for the machine learning model.

Model Setup and Selection

Defining the machine learning approach to predict site suitability involves choosing how the problem is formulated and selecting appropriate algorithms. One option is to frame site suitability as a supervised learning task. For example, a binary classification (suitable vs. not suitable), a multi-class classification (e.g., “unsuitable”, “moderately suitable”, “highly suitable”), or a regression where the target is a continuous suitability score could be used. Supervised learning requires a training dataset with known target values, obtained via expert-identified sample sites, existing well-performing green sites as positive examples, or results from a prior MCDA model as pseudo-labels. For instance, the results of an AHP GIS analysis could serve as initial training targets that the ML model will try to learn and eventually improve upon [15].

Alternatively, an unsupervised or semi-supervised approach could be considered if labelled data is scarce. Clustering algorithms (K-means, hierarchical clustering, etc.) might identify distinct groups of city areas with similar characteristics, which could then be interpreted in terms of suitability or required interventions [10]. However, for a clear “suitability map,” a supervised predictive model is more straightforward.

The framework is model-agnostic, meaning many ML algorithms can be tested to determine which one best characterises the relationship between spatial factors and site suitability. Candidate models include decision-tree-based ensembles such as Random Forest and Gradient Boosting, which handle mixed data

types well and provide interpretable feature importance; neural networks capable of learning complex nonlinear interactions but requiring careful tuning to avoid overfitting; support vector machines suitable for smaller, well-curated datasets; and hybrid approaches that combine ML predictions with optimisation or multi-criteria ranking to enhance practical relevance. The best-performing algorithm is selected based on predictive performance, interpretability, and usability to urban planners.

Model Training and Validation

Using the chosen algorithm, the model is trained on the prepared dataset. Data is divided into training and testing sets (and perhaps a separate validation set, or k-fold cross-validation is performed) to enable assessment of the model's generalisation to other locations in the city. In the case of a supervised approach, the split should be spatially representative, e.g., training on one part of Mumbai and testing on another to check whether the model can predict in other districts. Hyperparameters are optimised when necessary (via grid search, random search) to enhance model performance. The framework emphasises avoiding overfitting through regularisation and early stopping.

It is important to analyse the model after training to determine which features are contributing to the predictions. In the case of tree-based models, this could be feature importance scores or partial dependence plots (e.g., the model may indicate that NDVI and distance to existing parks are the most important predictors). In the case of complex models such as neural nets, interpretability algorithms such as SHAP (Shapley Additive Explanations) or LIME can be used to understand the impact of altering feature values on the suitability prediction [13]. Model performance is validated using the test set with relevant metrics: in a classification context, accuracy, precision-recall, the kappa statistic, or F1-score; for regression, R^2 and RMSE (Root Mean Square Error).

Suitability Mapping and Analysis

The trained model is deployed to generate a city-wide map of green infrastructure site suitability. The model is applied to the entire study area (all grid cells or relevant land parcels in Mumbai) to predict a suitability score or class for each location. The output could be a continuous index (e.g., 0 to 100 scale) or categorical levels (e.g., unsuitable, low, moderate, high suitability), resulting in a GIS raster layer of predicted suitability.

Post-processing of the raw output refines it for decision-making by classifying continuous suitability scores into a small number of categories using percentile or natural-break thresholds, with top-ranked areas highlighted as priority GI sites. Suitability results are presented as thematic GIS maps showing spatial patterns across Mumbai, with clear colour gradients, ward boundaries, and key landmarks. Model outputs are verified by field checks and expert advice to ensure feasibility on the ground. The model can then be refined with updated constraints and additional data.

Decision Integration

The last step is to include the suitability analysis in urban planning and decision-making in Mumbai. Findings are disseminated to city planners, policymakers, and stakeholders in a simplified form. In addition to maps, summary statistics may come in handy, e.g., "X percent of the land area in Mumbai is highly suitable for new green infrastructure, mainly in the suburban periphery."

The framework also facilitates scenario analysis. Planners may question what would happen if the focus is more on the reduction of heat islands. This may be investigated by changing input layers or model parameters and comparing the new suitability map to the baseline. A participatory component is also useful at this stage. Outputs of the model can be shared with community groups or the general population to seek feedback. Local knowledge may note, e.g. that an otherwise highly suitable vacant plot is privately owned and being developed, or that a moderately suitable area is in fact a community playground that can be improved.

Ultimately, the framework is intended to become a decision-support tool. This ML-led system can be used periodically by Mumbai's urban planning department or environment ministry to monitor progress and always find new opportunities for green infrastructure. The model can also be retrained as data quality improves (better resolution sensors, IoT environmental data, etc.), making the framework an organic component of the city's planning process. The output—the suitability map and related analytics—must be directly incorporated into urban development strategies, land-use zoning, and sustainability action plans in Mumbai.

Conclusion

Mumbai is developing at a very rapid pace, and this is associated with numerous environmental challenges. The extent to which the city can create and preserve green spaces in its already densely populated space will determine its future sustainability to a large extent. This paper provides a straightforward yet complete framework that uses GIS and machine learning to find good places to create new green areas in Mumbai, given its limited land area and climate risk.

Machine learning integration brings important benefits compared to conventional planning heuristics. By learning patterns from multi-layered spatial data, a ML-based solution can distinguish non-intuitive areas of opportunity (such as concentrations of heat-trap areas that are bare, or under-utilised government areas prone to flooding that could be converted into wetlands). It can also dynamically balance the many factors (environmental, social, infrastructural) that make a site suitable, avoiding one-size-fits-all regulations. In the case of Mumbai, greening efforts can be focused more precisely, prioritising interventions that will cool hotspots, absorb runoff, clean air pollution, and serve underserved communities. Such targeted planning is essential in a city where there is as little as 1–2 m^2 of green space per capita [6].

The provided framework is interpretable and iterative in nature, ensuring that AI is turned into a transparent decision support tool rather than a black box, so that planners are able to understand, rationalise, and integrate site recommendations as per shifting policy priorities and communal realities. By incorporating participatory inputs with understandable GIS-ML outputs, the approach allows the formulation of consensus between technical analysis and ground-level needs. This model is effectively suited to the Mumbai context and can contribute substantially to sustainable development by enabling the determination of the best suitable places to plant city green infrastructure to cool heat, manage floods, enhance biodiversity, enhance people's health, and adapt to climatic conditions. It can transform various spatial data into meaningful intelligence to develop a resilient and networked green space system that maximises environmental and social benefits, makes national sustainability imperatives possible, and offers a transferable and scalable model for other fast-growing cities in need of a data-driven path towards becoming more livable and sustainable.

References

1. Aiymeku, T. G., Tabor, K. W., Wedajo, G. D., & Roba, Z. R. (2024). Green spaces suitability analysis for urban resilience using geospatial technology.
2. Başeğmez, M., Doğan, A., & Aydın, C. C. (2025). Management of sustainable urban green spaces through machine learning-supported MCDM and GIS integration. Environmental Science and Pollution Research, 32(11), 11466–11487. https://doi.org/10.1007/s11356-025-36367-7
3. CREDAI–MCHI. (2025). Green spaces and climate resilience: British new towns vs Indian megacities. Retrieved from https://mchi.net/green-spaces-and-climate-resilience-british-new-towns-vs-indian-megacities/
4. Dobrinć, D., Miler, M., & Medak, D. (2025). Mapping the green urban: A comprehensive review of materials and learning methods for green infrastructure mapping. Sensors, 25(2), 464. https://doi.org/10.3390/s25020464
5. Gorjian, M. (2025). GIS-based assessment of urban green infrastructure: A systematic review of

advances, gaps, and interdisciplinary integration (2020–2024) [Preprint]. Preprints.org. https://doi.org/10.20944/preprints202508.0786.v1

6. Joy, M. S., Jha, P., Yadav, P. K., Bansal, T., Rawat, P., & Begam, S. (2024). Site suitability analysis of urban green parks in Ranchi city using GIS–AHP based multi-criteria decision analysis. Urbanization, Sustainability and Society, 1(1), 169–198. https://doi.org/10.1108/USS-10-2023-0008
7. Kooniyara, V. P. (2025). Predicting urban greening potentials with artificial intelligence model—a GIS-based machine learning approach for local assessment (Master's thesis, Technical University of Munich).
8. Lin, Z.-H., Laffan, S. W., & Metternicht, G. (2025). Strategically identifying optimal locations for multifunctional green infrastructure: A case study in the Taipei Basin. Land Use Policy, 157, 107654. https://doi.org/10.1016/j.landusepol.2025.107654
9. Zhang, L., Lin, H., & Wang, F. (2022). Individual tree detection based on high-resolution RGB images for urban forestry applications. IEEE Access, 10, 46589–46598.
10. Song, H., et al. (2022). HA-Unet: A modified U-Net based on hybrid attention for urban water extraction in SAR images. Electronics, 11, 3787.
11. Choi, K., et al. (2022). An automatic approach for tree species detection and profile estimation of urban street trees using deep learning and Google Street View images. ISPRS Journal of Photogrammetry and Remote Sensing, 190, 165–180.
12. Jombo, S., Adam, E., & Tesfamichael, S. (2022). Classification of urban tree species using LiDAR data and WorldView-2 satellite imagery in a heterogeneous environment. Geocarto International, 37, 9943–9966.
13. Schmohl, S., Vallejo, A. N., & Soergel, U. (2022). Individual tree detection in urban ALS point clouds with 3D convolutional networks. Remote Sensing, 14, 1317.
14. Jia, J., Cui, W., & Liu, J. (2022). Urban catchment-scale blue-green-gray infrastructure classification with unmanned aerial vehicle images and machine learning algorithms. Frontiers in Environmental Science, 9, 778598.
15. Waheeb, S., Al-Hussein, A., & Hassan, R. (2023). Integrating GIS, remote sensing, and machine learning for urban green infrastructure planning. Remote Sensing Applications: Society and Environment, 32, 100981.
16. Francis, J., Disney, M., & Law, S. (2023). Monitoring canopy quality and improving equitable outcomes of urban tree planting using LiDAR and machine learning. Urban Forestry & Urban Greening, 89, 128115.
17. Akın, A., Çilek, A., & Middel, A. (2023). Modelling tree canopy cover and evaluating the driving factors based on remotely sensed data and machine learning. Urban Forestry & Urban Greening, 86, 128035.
18. Yu, M., Xu, H., Zhou, F., Xu, S., & Yin, H. (2023). A deep-learning-based multimodal data fusion framework for urban region function recognition. ISPRS International Journal of Geo-Information, 12, 468.