Site Suitability for Essential Services and Digital Connectivity in India using Machine Learning

Muthu Reshmi K. [1*], Rishitha N.[1], Srishti Gulecha R.[1] and Dr Vani K.[1]

[1]Department of Information Science and Technology, College of Engineering Guindy, Anna University, India

*muthureshmi08@gmail.com

Abstract: *In India's impoverished regions, limited access to essential services and digital connectivity hinders socioeconomic development and escalates inequality. This study addresses these challenges by focusing on the placement of Aadhaar centers and Network towers to enhance service accessibility and connectivity. Aadhaar centers are crucial for providing residents with unique identification. Concurrently, improving network coverage is vital for economic activities, and communication. Site suitability for Aadhar Centres leverages the use of nighttime lights, population, built-up areas, and distance from roads to identify optimal locations. The dataset consisting of these factors for existing and non-Aadhaar center locations are given to machine learning models like Support Vector Classifier, Logistic Regression and XGBoost to estimate the suitability. Support Vector Classifier was the best model with a testing accuracy of 0.890. Using these predicted suitability rates, suitability maps are visualized with zone wise most suitable points. The network towers placement suitability module uses various factors like slope, population, elevation, Land Use Land Cover, distance to roads, and distance to recreational centers are processed for existing and non-tower locations to train and test various machine learning models. Support Vector Classifier, Random Forest Classifier and XGBoost models are used out of which XGBoost was identified as the best model with a testing accuracy of 0.940. The subsequent task involves identifying dead zones where signal coverage is nil. The existing network towers and few locations with most suitability to place new network towers are analyzed and the remaining area which is not covered by both are tagged as dead zones. The Aadhar center placement is done for all zones in Nawada, Bihar and Network Tower Placement Suitability along with dead zones is done for Coimbatore, Tamil Nadu. This study helps policymakers and service providers to enhance infrastructure and service delivery in impoverished regions of India.*

Keywords: *Machine Learning, Socioeconomic Development, Digital Connectivity, Dead Zones*

Introduction

Infrastructure development and access to reliable communication is crucial for fostering economic growth and innovation. Aadhaar is a 12-digit unique identity number that can be obtained voluntarily by all residents of India, based on their biometrics and demographic data (Unique Identification Authority of India, n.d). Aadhaar centers are vital to ensure efficient and accessible enrollment and authentication services for citizens across various regions. By strategically selecting sites, authorities can optimize outreach and service delivery, ensuring that Aadhaar services are readily available to all segments of the population, including those in remote and under-served areas. This contributes to enhancing infrastructure accessibility, and facilitating efficient public service delivery.

Identifying suitable sites for network towers promotes connectivity and supports economic development. Proper placement facilitates connectivity in underserved areas, enhances access to information, and promotes social inclusion. Both play a crucial role in addressing specific societal challenges, contributing to the overall progress towards achieving the SDGs (Goal 9) and fostering a more sustainable and equitable future for all. Determining suitable locations for network towers is crucial for ensuring reliable communication infrastructure, supporting economic development, and fostering innovation, contributing to sustainable growth and advancement in various sectors.

Identifying suitable sites for Aadhaar centers and network towers relies on various environmental, demographic and other factors. By analyzing satellite data, researchers can monitor changes over time and comprehend the needs of the area. Machine learning algorithms can efficiently process and analyze vast amounts of data, enabling the identification of complex patterns and relationships. Using these methods is often more cost-effective and helps in identifying most suitable locations.

The primary objective of this study is to identify suitable sites for placement of Aadhaar centers and network towers using multi-source data and machine learning techniques. Google Earth Engine (GEE) is used to access satellite data to fetch factors like slope, Land Use Land Cover (LULC), population, elevation, and Night-Time Light (NTL) intensity. Additionally, OpenStreetMap is used to extract the Point of Interest (POI) data and road network. Machine learning models are used to identify the optimal locations. Furthermore, zone wise best locations for Aadhaar centers are identified and dead zones are identified for network towers.

Literature Review

Numerous research works have been carried out related to site suitability analysis using various techniques. Various studies have used GIS (Geographic Information Systems) and AHP (Analytic Hierarchical Process) to determine a suitable site for different infrastructures like schools and hospitals. The identification of suitable hospital sites in Rajpur-Sonapur municipality, West Bengal is done using Analytic Hierarchy Process along with Geographic Information System in the study by Halder et al. (2020). Analytical Hierarchy Process, a multi-criteria decision analysis technique which was used to organize the identified criteria into a hierarchy structure to identify suitable locations for recommending suitable sites for schools in district of Melaka Tengah a part of Malacca, Malaysia (Mustaffa et al., 2021).

The study (Santanu Kumar Misra et al., 2015) provides a comprehensive overview of methodologies used in determining suitable locations for urban development using Geographic Information Systems and Multi-Criteria Evaluation (MCE). The review emphasizes the flexibility and simplicity of these techniques in analyzing potential urban development sites. Over the past decade, extensive research has been conducted utilizing various approaches to assess land suitability. Key methods include the use of ArcGIS software to analyze multiple thematic layers created from satellite data, considering factors such as elevation, slope, land use, and proximity to infrastructure. Weights are assigned to each factor based on their importance, and these criteria are combined using GIS tools to generate final suitability maps. The review emphasizes the integration of environmental and infrastructural factors in the evaluation process, providing a robust framework for urban planners. The use of advanced techniques such as Fuzzy-AHP and DEMATEL further enhances the precision of site suitability analysis, allowing for more informed decision-making in urban development projects.

Machine learning techniques have also been studied for suitability analysis. A method to identify suitable areas for hospitals within the Gaza Strip in Palestine using machine learning approaches is developed (Almansi et al., 2021). Hospital inventory points were taken and 15 factors were extracted. Support Vector Machine (SVM), Multilayer Perceptron (MLP), and linear regression (LR) models were used for generating suitability maps. Based on areas under the ROC curve, the MLP model yielded a prediction accuracy of 84.90%, SVM of 75.60%, and LR of 64.40%. Machine learning approaches are utilized to advise the spatial location of solar power projects in China (Sun et al., 2023). Twenty-one factors in all—including land cover, slope, elevation, proximity to water resources, surface annual mean air temperature, land cost, ease of transportation, distance to residential area and power grid, population density, electricity consumption, solarGHI, and rooftop density—are analyzed. The models that were trained and verified were Multi-Layer Perceptron (MLP), Random

Forest (RF), and extreme gradient boosting (XGBoost). Five categories were included in the suitability map that was subsequently created: very low, low, moderate, high, and very high. At a spatial resolution of 1km, probability maps of PV installation were created. Out of the three ML models, the RF model had the highest Area Under Curve (AUC) values (0.78) for the training and validation datasets.

The study proposed by Tayal et al. (2020) provides a detailed methodology for analyzing and optimizing mobile network coverage in Uttarakhand, India, using Geographic Information System (GIS) tools and datasets. It outlines the collection and analysis of various datasets, including ASTER GDEM for topographical information, census data for population distribution, and existing tower locations from BSNL office. The study incorporates mosaicing of ASTER GDEM images, block-wise population data analysis, and mapping of existing Base Transceiver Station locations with coverage areas. The analysis includes creating buffer zones around existing towers to identify overlapping and dead zones. The methodology aims to suggest optimal tower locations by removing dead zones and minimizing overlap, and improving mobile network coverage in the region. Thus, the proposed methodology can effectively identify suitable locations for new towers, enhancing mobile network coverage and quality of service.

Veronesi et al. (2017) introduces an innovative approach to improve GIS-based Multi-Criteria Decision Analysis (MCDA) by automating the weight selection process for various criteria. The methodology incorporates multiple geographic and environmental factors such as land use, proximity to infrastructure, and environmental impacts, utilizing advanced algorithms to enhance objectivity and consistency in weight determination. The study's case application in transmission tower site selection demonstrated that automated weight selection reduces subjectivity, yielding reliable and consistent results. The automated method produced a suitability map, accurately identifying optimal transmission tower locations, thus enhancing the efficiency and robustness of spatial decision support systems. This approach proves effective in practical applications, leading to more reliable decision-making in geographic and spatial planning.

Methodology

This section outlines the methodology employed to assess site suitability for Aadhaar centers and Network Towers.

a. Aadhar Centre Site Suitability:

To identify suitable locations for Aadhaar enrollment centers, this section details the methodology employed, which leverages a spatial analysis framework incorporating nighttime lights, population density, and proximity to roads. Figure 1 shows the module diagram for this process.

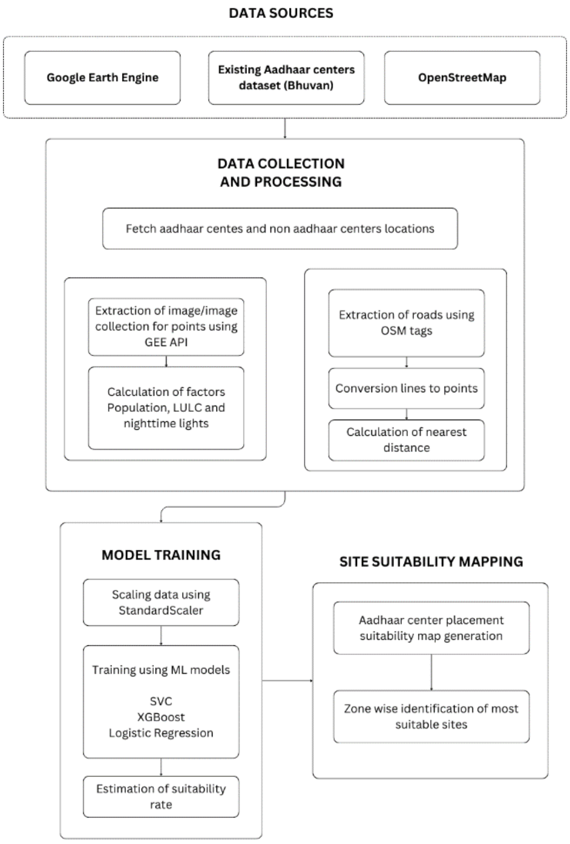


Figure 1: Architecture for Site suitability for placement of Aadhaar Centers

Firstly, the existing Aadhaar center locations are fetched for the Kanchipuram district using a dataset provided by Bhuvan (2023). Non-Aadhaar center locations are randomly sampled. There are about 164 existing Aadhaar centers in Kanchipuram and 200 non-Aadhaar center locations are taken.

The features are extracted for these points to build the dataset. The avg_rad band of VIIRS images is used to fetch nighttime light data and the population is obtained from WorldPop Global Project Population Data. Utilizing Google's Dynamic World V1 dataset, the most recent LULC type is obtained. Distance to roads, and Distance to recreational centers extraction begins by importing an area of interest which is Coimbatore in QGIS software. Using Open Street Map (OSM), the roads, and are fetched after which clipping is done. Conversion to points is done and finally, distance is calculated using the distance to hub tool.

Support Vector Classifier (SVC), XGBoost, and Logistic Regression (LR) models are trained with the extracted features. The dependent variable is Aadhaar center/not. The data is split 80-20 for training and testing. The best model is selected and saved after model training.

A shapefile of the area and the best model are used as inputs to construct suitability maps. Latitude and longitude are extracted at high resolution points across the region. The trained model is then leveraged to estimate a suitability score for each location. Finally, a grid is created and the suitability scores are interpolated to produce a map where each location has a predicted level of suitability.

In addition to the dataset used for training and testing, the Nawada region is selected to test a new area. Using QGIS and satellite imagery, the characteristics are first extracted. After scaling, the saved model is loaded and used to forecast each point's appropriateness. Lastly, a tiff file is prepared and used for visualization utilizing the dataframe containing the suitability values. Zone wise most suitable points are fetched and displayed.

b. Network Tower Site Suitability:

Network tower placement suitability is done by the procedure highlighted in Figure 2. OSM, Google Earth Engine, and the existing tower data are utilized to extract various factors like Slope, Population Density, Elevation, LULC, Distance to Roads, and Distance to Recreational centers which are processed to train various machine learning models and produce suitability maps in the Area of Interest (AOI).

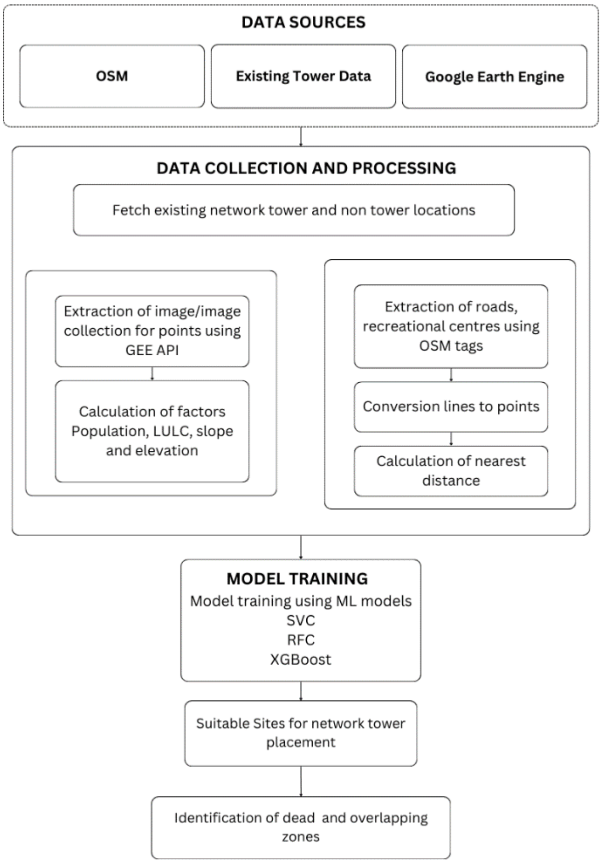


Figure 2: Architecture for Network tower placement Suitability

The existing network towers which are around 1587 and non-network tower points around 2000 are taken for training in Coimbatore, a city in the state of Tamil Nadu in India. Coimbatore encompassing roughly 500 square kilometers is taken as the study area. This choice offers a unique blend of urban and rural environments, with the foothills of the Western Ghats introducing variation in terrain. Coimbatore's status as a developing city experiencing infrastructure growth makes it an ideal place for network placement optimization.

For each point, the following features - Elevation, Slope, LULC, Population Density, Distance to roads, and Distance to recreational centers are extracted. Slope, elevation and population are extracted from the Google Earth Engine Datasets - NASA SRTM Digital Elevation 30m, WorldPop Global Project Population Data datasets respectively. For every point in the dataset, a point geometry is created and slope, elevation and population are retrieved.

The algorithm for LULC extraction involves several steps. First, a dataset with coordinates from Coimbatore is taken. For each coordinate, a point geometry is created, and a filter is defined for the region of interest and time period. The Region of Interest (ROI) and time period are specified as Coimbatore and 2023 - 2024. Next, an image collection is retrieved using the 'DYNAMICWORLD V1' dataset, sorted by time, and converted to a list. Then, the list of images is iterated over, the 'label' band is selected, and a reduce region operation is applied to extract the LULC type for each point. Finally, the calculated values are stored in a CSV file. Distance to roads, and Distance to recreational centers are extracted using QGIS software using the same method proposed for Aadhar Centre placement suitability.

After feature extraction is done, the following machine learning models - Support Vector Classification, XGBoost, and Random Forest classification are used. The data is scaled using Standard Scaler before training and testing. The dataset is split as 80-20 for training and testing. Support Vector Classification finds the separating hyperplane that best margins data points of different classes. XGBoost is a powerful ensemble learner that combines weak decision trees into a strong learner for improved classification and regression. Random Forest is an ensemble learning method using multiple decision trees for robust classification. Classification evaluation metrics like accuracy, precision, recall and F1 Score are used to evaluate these models.

The best model is chosen to visualize the network tower suitability map for the whole Coimbatore region. The most suitable points of a region are taken and the custom ranges are drawn after which the dead zones. It considers factors like desired signal strength, existing towers' real-world coverage, and areas with high user demand. It then virtually positions new towers and adjusts their reach to maximize coverage while meeting the minimum signal requirement. Finally, any remaining uncovered areas are flagged as dead zones, and adjustments can be made to tower placement or range to improve coverage.

The UI is developed with the help of QGIS, qgis2webplugin, HTML, CSS, JavaScript. qgis2web is a plugin that is used to export the vector and raster layers easily to a web interface.All the influencing factors for the site suitability are visualized along with the final results. ReactJS is used for navigation.

Results and Discussion

This section presents the findings of the analysis regarding site suitability for both Aadhaar centers and network towers. It also discusses the results and their implications for optimal placement strategies.

a. Aadhar Centre Site Suitability:

The findings for Aadhaar site suitability mapping are detailed here. Three models are trained and tested with the dataset of Kanchipuram region. 20% of the dataset is used for testing and the corresponding accuracy, precision, and recall values for every model are tabulated in Table 1 using the formulas given in Figure 3. 5-fold cross validation was done for every model.

$$\begin{aligned} precision &= \frac{TP}{TP + FP} \\ recall &= \frac{TP}{TP + FN} \\ F1 &= \frac{2 \times precision \times recall}{precision + recall} \\ accuracy &= \frac{TP + TN}{TP + FN + TN + FP} \\ specificity &= \frac{TN}{TN + FP} \end{aligned}$$

Figure 3: Classification Evaluvation Metrics-Formulas

Table 1: Evaluation Metrics - Aadhaar Centers site suitability

Model	Support Vector Classifier	Logistic Regression	XGBoost
Training Accuracy	0.945	0.872	0.935
Testing Accuracy	0.890	0.849	0.863
Training Precision	0.931	0.883	0.899
Training Recall	0.945	0.822	0.961
Testing Recall	0.914	0.829	0.914
Training F1-score	0.923	0.851	0.929
Testing F1-score	0.889	0.841	0.865

The hyper-parameters used for SVC model are as follows - the C parameter controls the penalty for misclassification of data points and is set to 10. The Kernel parameter determines the type of kernel function used to transform the input data into a higher-dimensional space where it can be linearly separated and is set to 'rbf' which stands for Radial Basis Function. The gamma value is set to 1. It can be inferred that SVC has a better test accuracy with 89%.

Additional testing is done for Nawada region. Figure 4 shows the extracted features for Nawada to predict suitable sites for the placement of Aadhaar centers. Figure 4(a) shows the nighttime light intensity for Nawada, while Figure 4(c) depicts the Population density and LULC is shown in Figure 4(d). Distance to roads is shown in Figure 4(b).

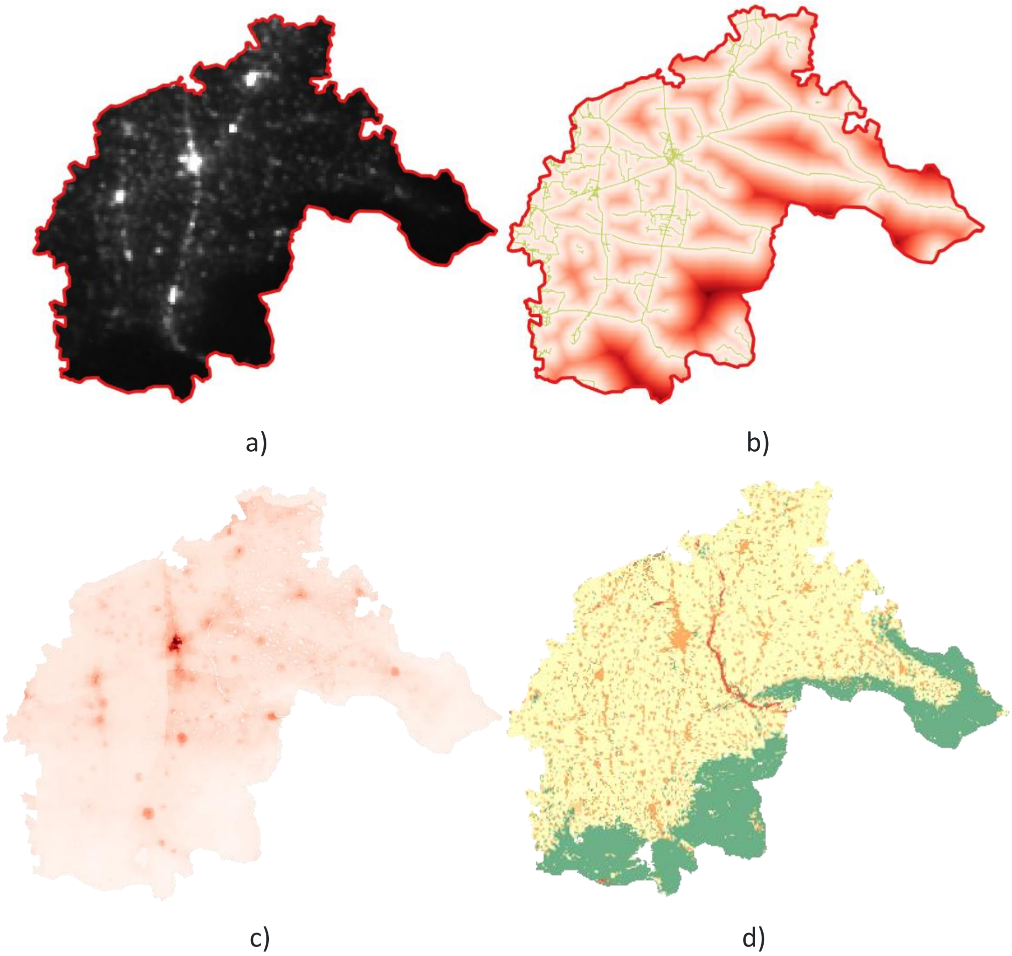


Figure 4: Extracted features for Aadhaar centers suitability – Nawada

Using these features, a suitability map is generated as shown in Figure 5. The regions are classified into 5 categories as shown. Zone wise most suitable points for Aadhaar centers in Nawada is shown in Figure 6.

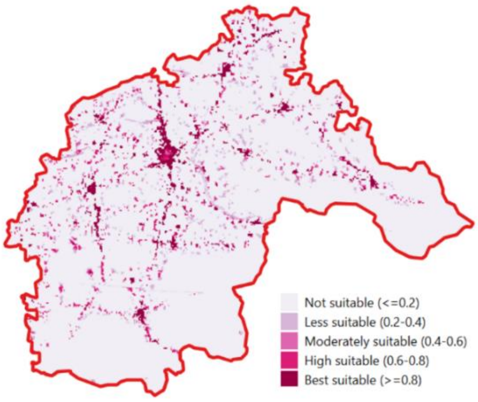


Figure 5: Suitability map for placement of Aadhaar centers in Nawada

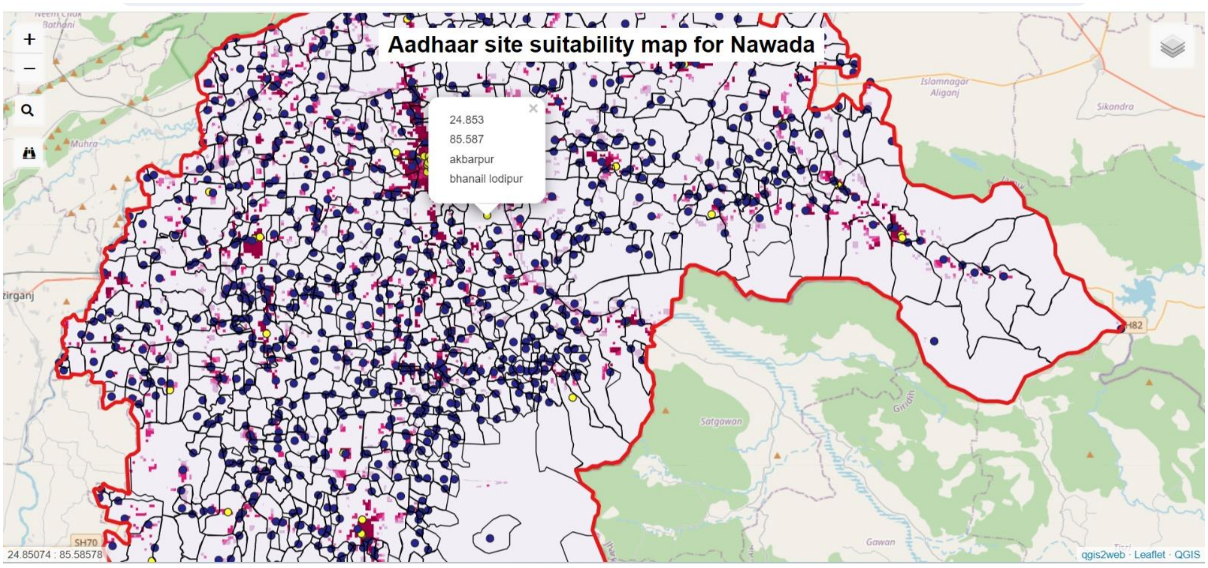


Figure 6: Zone wise suitable points for Aadhaar centers in Nawada

b. Network Tower Site Suitability:

The findings for network site suitability are detailed here. For testing, the points were taken for the whole Coimbatore region. The features obtained are visualized in Figure 7. Elevation, Slope, Population Density, Distance to Roads, Distance to Recreational centers, and LULC are shown in Figures 7(a), 7(b), 7(c), 7(d), 7(d) and 7(f) respectively.

The ML models - SVC, RFC, and XGBoost are trained and the best one SVC is used for visualizing the whole area of Coimbatore. The evaluation metrics are tabulated in Table 2 using the formulas given in Figure 3.

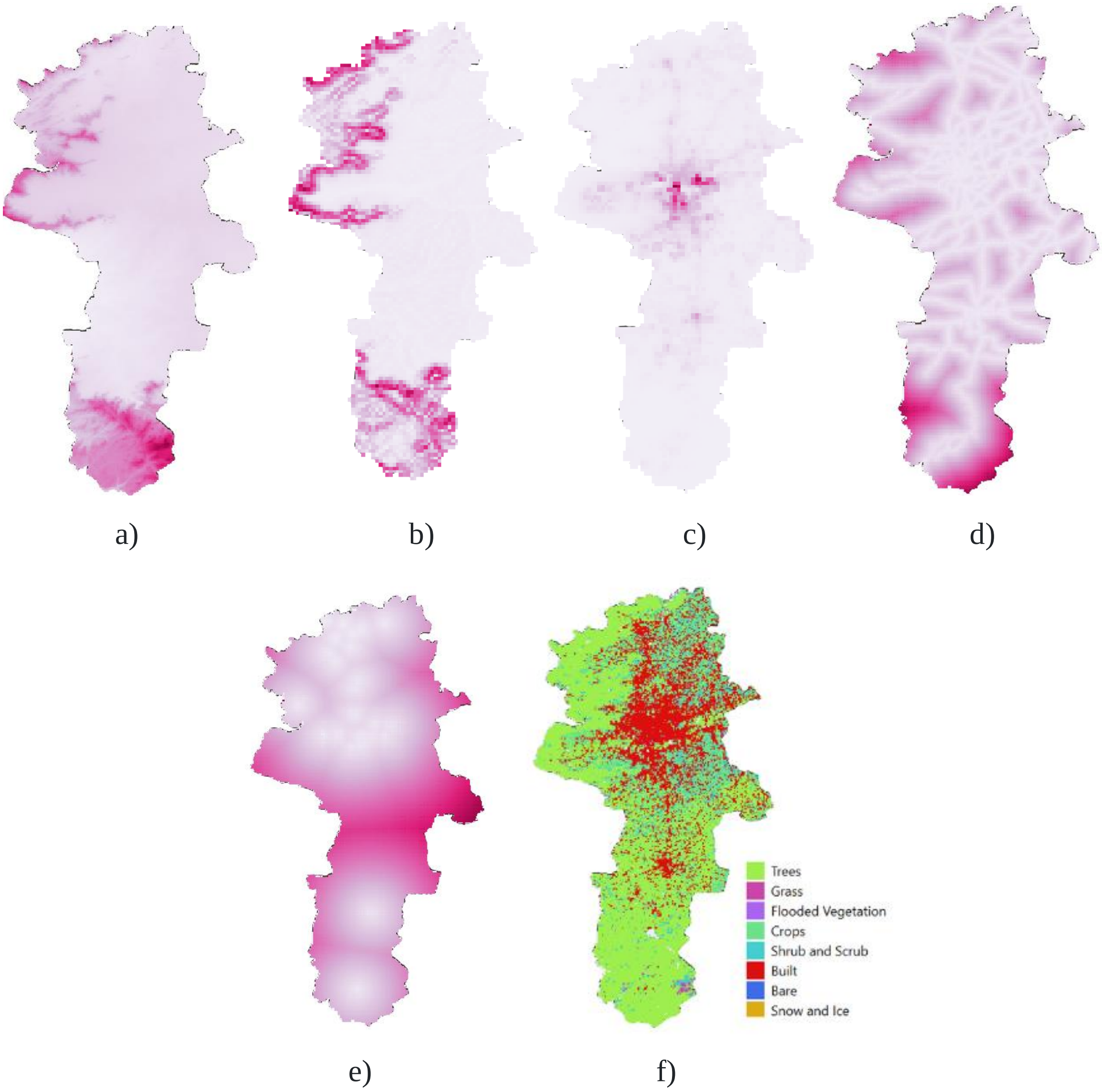


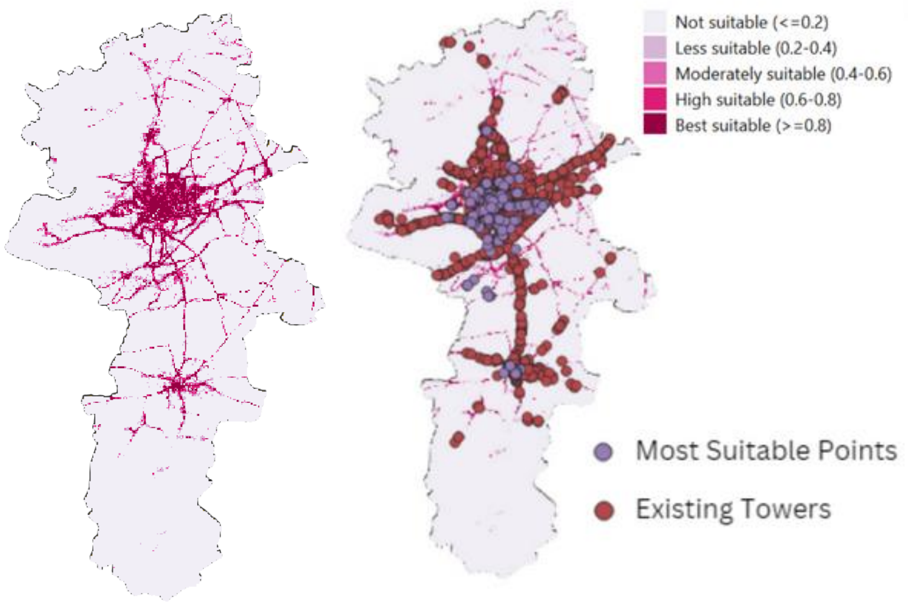
Figure 7: Coimbatore Features - Network Tower Site Suitability

The hyper-parameters used are as follows -n_estimators is set to values 50, 100, and 200 which specifies the number of boosting rounds. While learning_rate which controls the step size during gradient descent is set to 0.01, 0.1, and 0.2. And max_depth which is set to 3, 5 and 7 is used to determine the maximum depth of each decision tree.

Table 1: Evaluation Metrics – Network Tower Site Suitability

Model	XGBoost	Support Vector Classifier	Random Forest Classifier
Training Accuracy	0.990	0.895	0.977
Testing Accuracy	0.940	0.902	0.938
Training Precision	0.984	0.854	0.960
Training Recall	0.937	0.879	0.926
Testing Recall	0.994	0.911	0.988
Training F1-score	0.937	0.918	0.943
Testing F1-score	0.989	0.882	0.974

In training, the XGBoost model resulted in an accuracy of 0.990, precision of 0.984, recall of 0.994, and F1 Score of 0.989. Similarly, testing values were 0.940, 0.937, 0.937, and 0.937 for accuracy, precision, recall, and F1 score respectively. The visualization of the suitability map is shown in Figure 8(a). The existing network towers and top locations for placing new network towers are shown in Figure 8(b).



a) b)

Figure 8: Suitability Map for Placement of Network Towers – Coimbatore

The dead zones for a small region of Coimbatore are shown in Figure 9. The green buffers represent existing network towers. The Yellow buffers represent the new towers that can be placed. The red region is the dead zone that is not covered by either.

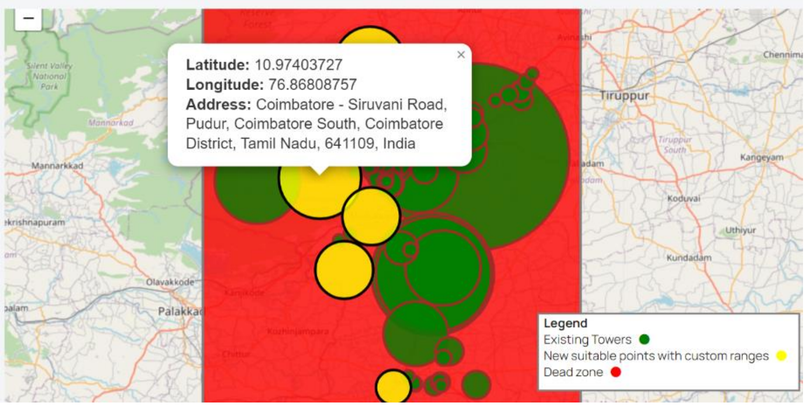


Figure 9: Dead Zones after placing old and new network towers in a small region of Coimbatore

c. User Interface:

Figure 10 visualizes the zone-wise best points for placement of Aadhaar centers in Nawada in the User Interface. For every village/town, the best point can be fetched while hovering over it.

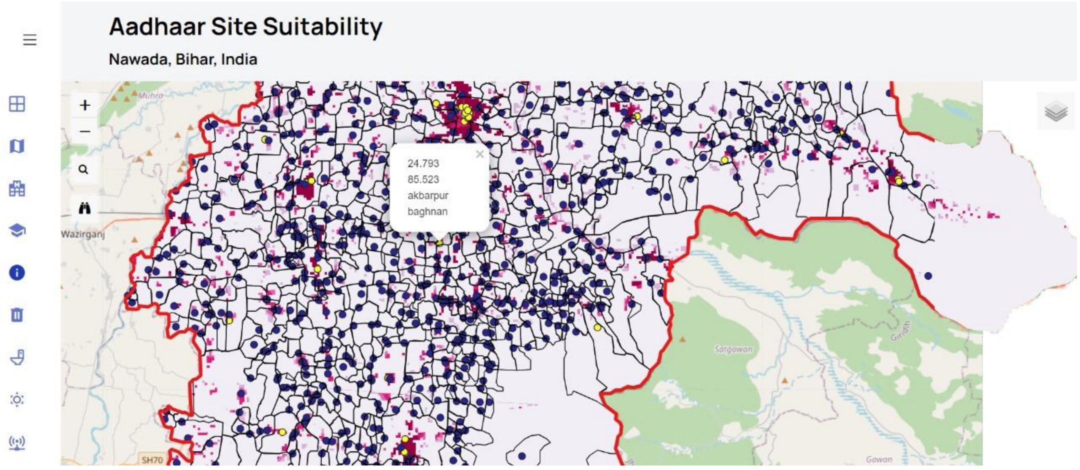


Figure 10: Aadhar Site Suitability for Nawada in User Interface

Figure 11 shows the network tower site suitability in Coimbatore in the User interface. On hovering a point the latitude, longitude along with the address of that coordinate is displayed.

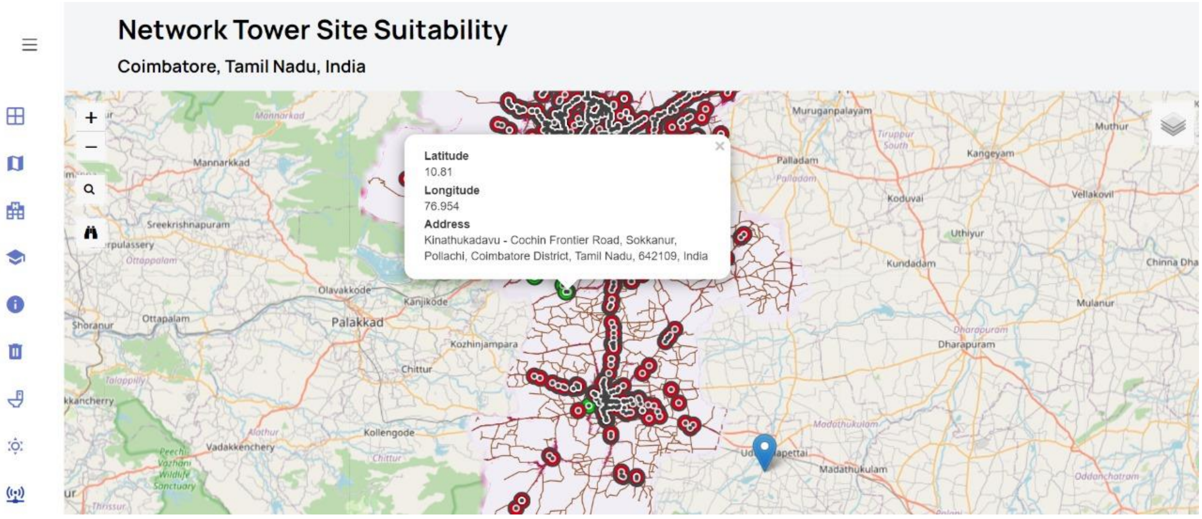


Figure 11: Network Tower Site Suitability for Coimbatore in User Interface

Conclusion and Recommendation

The study demonstrates the use of multi-source data and machine learning approaches for site suitability mapping. It highlights how Geographic Information System (GIS) data can be used to gather a wide range of information to extract various features, generate suitability maps and perform analysis for achieving major sustainable development goals for the improvement of poverty zones and the overall well-being of the society. The system was developed using tools and technologies such as Google Earth Engine (GEE) API, OpenStreetMap API, Google Colaboratory Notebook, and QGIS.

Infrastructure plays a vital role in sustainable development. Thus, is it necessary to identify the most suitable locations for development. Hence, for suitability mapping, the required features are extracted using which the dataset is built and machine learning models are trained. Finally, the best model is chosen for additional testing and generation of site suitability maps and to perform analysis to find the most suitable points.

The Aadhaar center site suitability module helps to identify best sites for placement of Aadhaar centers in every village/town. It makes use of OpenStreetMap and Google Earth Engine to fetch the influencing factors - NTL, population density, LULC, and distance to roads. SVC, LR, and XGBoost machine learning models are used for training. Among the models SVC is found to be the best model with an accuracy of 0.945, precision of 0.931, recall of 0.945 and F1-score 0.923 during training. It has an accuracy of 0.89, precision of

0.864, recall of 0.914 and F1-score 0.889 during testing. Hence chosen for Aadhaar center site suitability mapping in Nawada.

The network tower placement module focuses on identifying optimal locations for network towers, essential for establishing robust communication infrastructure. Leveraging data from various sources such as OpenStreetMap and Google Earth Engine, factors like terrain slope, population density, elevation, land use, and proximity to key amenities like distance to roads and recreational areas are analyzed. Machine learning techniques, particularly the XGBoost model, are employed to generate suitability maps within the area of interest. With high accuracy and precision, the XGBoost model proves effective in pinpointing suitable tower locations and facilitating informed decision-making for network infrastructure deployment and expansion with an accuracy of 0.990, precision of 0.984, recall of 0.994, and F1 Score of 0.989. Similarly, testing values were 0.940, 0.937, 0.937, and 0.937 for accuracy, precision, recall, and F1 score respectively. The dead zones for a small region of Coimbatore are found after placing the existing network towers and new towers at locations with more suitability with customized ranges.

Incorporating dynamic, real-time data and multi-objective optimization can further refine site selection strategies. Collaborating with urban planning initiatives and expanding research to rural areas will help ensure comprehensive coverage and connectivity. The work can also be extended for placement of other essential infrastructure. Incorporating deep learning approaches can improve the handling of complex and large-scale datasets.

References

Almansi, K. Y., Shariff, A. R. M., Abdullah, A. F., & Ismail, S. N. S. (2021). Hospital Site Suitability Assessment Using Three Machine Learning Approaches: Evidence from the Gaza Strip in Palestine. Applied Sciences, 11(22), 11054. https://doi.org/10.3390/app112211054

Bhuvan. (n.d.). Retrieved May 1, 2024, from https://bhuvan.nrsc.gov.in/hackathon/iisf2023/

Halder, B., Bandyopadhyay, J., & Banik, P. (2020). Assessment of hospital sites' suitability by spatial information technologies using AHP and GIS-based multi-criteria approach of Rajpur–Sonarpur Municipality. Modeling Earth Systems and Environment, 6(4), 2581–2596. https://doi.org/10.1007/s40808-020-00852-4

Misra, S. K., & Sharma, S. (2015). Site suitability analysis for urban development: A review. The International Journal on Recent and Innovation Trends in Computing and Communication, 3(6), 3647-3651. https://doi.org/10.17762/ijritcc.v3i6.4511

Mustaffa, A. A., Bahiah, B., Rasib, A. W., & Saifullizan, M. B. (2021). GIS and Multi-criteria Analysis for School Site Selection (Study Case: Malacca Historical City). International Journal of Integrated Engineering/International Journal of Integrated Engineering, 13(4). https://doi.org/10.30880/ijie.2021.13.04.022

Sun, Y., Zhu, D., Li, Y., Wang, R., & Ma, R. (2023). Spatial modelling the location choice of large-scale solar photovoltaic power plants: Application of interpretable machine learning techniques and the national inventory. Energy Conversion and Management, 289, 117198. https://doi.org/10.1016/j.enconman.2023.117198

Tayal, S., Garg, P. K., & Vijay, S. (2017). Site suitability analysis for locating optimal mobile towers in Uttarakhand using gis. In Proceedings of the 38th Asian Conference on Remote Sensing—Space Applications: Touching Human Lives, ACRS (pp. 2-7).

Unique Identification Authority of India. (n.d.). Retrieved May 1, 2024, from https://uidai.gov.in/

Veronesi, F., Schito, J., Grassi, S., & Raubal, M. (2017). Automatic selection of weights for GIS-based multicriteria decision analysis: site selection of transmission towers as a case study. Applied Geography, 83, 78–85. https://doi.org/10.1016/j.apgeog.2017.04.001