

**A COMPARATIVE ANALYSIS OF MACHINE LEARNING  
APPROACHES FOR SITES SUITABILITY OF  
BROADBAND TOWERS**



**By**

**GHULAM HASNAIN**

**A thesis submitted in partial fulfillment of the requirements for  
the degree of Master of Science in Remote Sensing and  
Geographical Information Systems**

**Institute of Geographical Information Systems  
School of Civil and Environmental Engineering  
National University of Sciences and Technology  
Islamabad, Pakistan**

**July 2023**

## THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Ghulam Hasnain (Registration No. MSRSGIS 00000327485), of Session 2020 (Institute of Geographical Information systems) has been vetted by undersigned, found complete in all respects as per NUST Statutes/Regulation, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature: 

Name of Supervisor: Dr Javed Iqbal

Date: 6-Jul-2023

Signature (HOD): 

Date: 6-Jul-2023

**Dr. Javed Iqbal**  
Professor & HOD IGIS, SCEE (NUST)  
H-12, Islamabad

Signature (Principal & Dean SCEE): 

Date: 03 AUG 2023

**PROF DR MUHAMMAD IRFAN**  
Principal & Dean  
SCEE, NUST

# National University of Sciences & Technology

## MASTER THESIS WORK


We hereby recommend that the dissertation prepared under our supervision by: Mr. Ghulam Hasnain (Reg # 00000327485) Titled: “A Comparative Analysis of Machine Learning Approaches for Sites Suitability of Broadband Towers” be accepted in partial fulfillment of the requirements for the award of MS degree with (B+) grade.

### Examination Committee Members


1. Name: Ms. Quratulain Shafi

Signature: 

2. Name: Mr. Waqas Ahmad

Signature: 

Supervisor's Name: Dr. Javed Iqbal

Signature: 

Date: 26 Jun 2023



Head of Department


**Dr. Javed Iqbal**

Professor & MOD IGIS, SCEE (NUST)  
H-12 Islamabad

Date: 03 AUG 2023

26/6/2023  
Date

**COUNTERSIGNED**

  
Principal & Dean SCEE  
PROF DR MUHAMMAD IRFAN  
Principal & Dean  
SCEE, NUST

# **Dedication**

To

My Lovely Family

For their unwavering support, care, and motivation throughout my academic journey.

## **Academic Thesis: Declaration of Authorship**

I, **Ghulam Hasnain**, declare that this thesis and the work presented in it are my own and have been generated by me as the result of my own original research.

### **A COMPARATIVE ANALYSIS OF MACHINE LEARNING APPROACHES FOR SITES SUITABILITY OF BROADBAND TOWERS**

I confirm that:

1. This work was done wholly by me in candidature for an M.S. research degree at the National University of Sciences and Technology, Islamabad.
2. Wherever I have consulted the published work of others, it has been clearly attributed.
3. Wherever I have quoted from the work of others, the source has always been cited.
4. I have acknowledged all main sources of help.
5. Where the work of thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.
6. None of this work has been published before submission. This work is not plagiarized under the H.E.C. plagiarism policy.

Signed: ..........

Date: ...06/July/2023.....

## **Acknowledgement**

I am thankful to my Creator Allah Subhana-Watala to have guided me throughout this work at every step and for every new thought which You setup in my mind to improve it. Indeed, I could have done nothing without Your priceless help and guidance. Whosoever helped me throughout the course of my thesis, whether my parents or any other individual, was Your will, so indeed none be worthy of praise but You.

I am profusely thankful to my beloved parents who raised me when I was not capable of walking and continued to support me throughout every department of my life.

I would also like to express special thanks to my supervisor Dr. Javed Iqbal for his help throughout my thesis and for the Research Methodology course which he has taught me. I can safely say that I haven't learned any other subject in such depth than the ones which he has taught.

I would also like to say special thanks to Mr. Waqas for his tremendous support and cooperation. Each time I got stuck in something; he came up with the solution. Without his help I wouldn't have been able to complete my thesis. I appreciate his patience and guidance throughout the whole thesis.

I would also like to thank Miss Quratulain shafi, for being on my thesis guidance and evaluation committee and express my special thanks for the help. I am also thankful to Mr. Faqir Hussain and Miss Tayyaba Sana for their support and cooperation.

Finally, I would like to express my gratitude to all the individuals who have rendered valuable assistance to my study.

## TABLE OF CONTENTS

Certificate.....	i
Dedication.....	ii
Academic Thesis: Declaration of Authorship.....	iii
Acknowledgement.....	iv
List of Figures.....	vii
List of Tables.....	viii
List of Abbreviations.....	ix
ABSTRACT.....	x
CHAPTER 1 INTRODUCTION.....	1
1.1 Role of geoinformatics in telecommunication.....	4
1.2 Role of machine learning in GIS.....	6
1.3 Research gap.....	10
1.4 Motivation.....	11
1.5 Relevance to national needs.....	12
1.6 Objectives.....	12
CHAPTER 2 MATERIALS AND METHODS.....	13
2.1 Study area.....	13
2.2 Methodology.....	14
2.2.1 Data acquisition.....	14
2.2.2 Data standardization.....	24
2.3.2 Data modelling.....	24
2.3.2.1 Random Forest (RF).....	24
2.3.2.2 Support vector machine (SVM).....	26
2.3.2.3 Extreme gradient boosting (XGBoost).....	27
CHAPTER 3 RESULTS AND DISCUSSIONS.....	28
3.1 Random Forest.....	28
3.1.1 Accuracy.....	28
3.1.2 Relative variable importance.....	29
3.1.2 Sites suitability for broadband towers.....	29
3.1.3 Predication quality.....	29
3.2 Support vector machine.....	30
3.2.1 Accuracy.....	30
3.2.2 Relative variable importance.....	30
3.2.3 Sites suitability for broadband towers.....	30
3.2.4 Predication quality.....	31
3.3 XGBoost.....	31

3.3.2	Relative variable importance.....	31
3.3.2	Sites suitability for broadband towers.....	31
3.3.3	Predication quality.....	32
3.4	Assessment of predicted accuracy.....	32
3.5	Comparisons of models.....	37
CHAPTER 4 CONCLUSION AND RECOMMENDATIONS .....		40
4.1	Conclusion.....	40
4.2	Recommendations .....	41
References.....		42
APPENDICES .....		50
	Appendix – 1. Code for land use classification in google earth engine. ....	50
	Appendix – 2. Code of geospatial machine learning.....	51



## List of Figures

Figure 1. Study area map showing Islamabad with its elevation.....	15
Figure 2. flowchart of methodology adopted in the study.....	15
Figure 3. showing datasets maps of all variables.....	23
Figure 4. Accuracy comparison of classifiers on training and test data .....	34
Figure 5. BTS sites suitability using RF classifier.....	34
Figure 6. BTS sites suitability using SVM classifier.....	35
Figure 7. BTS sites suitability using XGBoost.....	35
Figure 8. showing the key matrices of all three models. ....	36
Figure 9. showing the comparison of models.....	39

## List of Tables

Table 1 showing dataset used in this study.....	16
Table 2 showing references for parameters selection.....	16
Table 3 showing data preprocessing of all factors.....	20
Table 4 all attributes (factors) data in BTS data layer .....	20
Table 5 Summary table of all classifiers.....	36

## List of Abbreviations

<b>Abbreviation</b>	<b>Explanation</b>
ML	Machine learning
GML	Geospatial machine learning
BTS	Base transceiver station
GPS	Global positioning system
RF	Random Forest
SVM	Support vector machine
XGBoost	Extreme gradient boosting
AUC	Area under curve
ROC	Receiver operating characteristic

## ABSTRACT

The demand for high-speed wireless communication services is increasing due to their wide applicability in daily life which requires smart planning to provide seamless coverage. Geospatial technologies play a vital role in planning by providing valuable insights into the physical and geographical aspects of a given area, but still the telecom sector does not utilize the power of geospatial machine learning approaches. The objective of the study was to (1) propose suitable sites for Base Transceiver Station (BTS) towers using Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost) and Random Forest (RF) classifier and (2) compare the models based on recall, accuracy, specificity, and Area Under the Curve (AUC). The land-use, geology, population, proximity to roads, bulk density and slope were used as exploratory variables. Models were trained on 70% data using 10-fold cross validation technique while 30% data was used for model validation using the python programming environment in ArcGIS. The results showed that the XGBoost comes with the accuracy of 97%, RF of 89%, and SVM of 57% for proposing suitable BTS sites. Relative variable importance showed that population, proximity to roads, slope, and land-use were identified as the most important exploratory variables, whereas bulk density and geology were recognized as the least relevant ones. The findings of validation matrices concluded that XGBoost is the best performing model for proposing suitable sites for BTS towers with the classification quality of 0.97%. The study reveals that these valuable insights are helpful for the telecom operators to implement XGBoost for proposing sites to increase signal strength and coverage for users.

# CHAPTER 1

## INTRODUCTION

Telecommunication services are essential for modern society and have a wide range of applications and benefits. It allows people to communicate with each other over long distances, whether by phone, email, or other means. This enables individuals, businesses, and organizations to stay connected and share information and ideas. It provides the infrastructure that enables people and organizations to connect to the internet and access a vast array of information and resources (Jean et al., 2018). It contributed to the global economy by enabling businesses to operate across the globe, indentifying new potential markets, and increase productivity. It provides the bases for the development of information technology (IT) sector ,boosting the precision agriculcture , smart villages , telehealth and distance learning which contribute in the development of the country. (Ullah et al., 2019). Telecom infrastructure are very important and its services plays a signifcant role to keep organizations, amd individuals stay connected and keeps businesses running. Telecommunication infrastructure is complex, it required a lot of capital and planning. Planning for telecommunication infrastructure is challenging due to several factors, including cost, regulatory constraints, and technological advancements. The installation of new base transceiver station (BTS) tower, infrastructure, such as fiber optic cables, towers, and other equipment, can be costly. BTS towers are the important part of telecom industry as it can provide coverage to large area without physically in contact with all the serving area using different frequency bands.It can provide extensive network coverage to specific regions, including , urban, indoor and rural areas. These towers are strategically positioned to ensure that a wide range of areas receive uninterrupted network coverage. They are crucial for individuals and businesses relying on wireless communication for their daily activities. By strategically placing towers, network operators can ensure that

mobile phone users have access to reliable and continuous network coverage. The location of BTS towers are chosen strategically to ensure high quality coverage to a wide range of area. Its basically design to cover maximum population and its coverage distance is vary according to population density of the serving area (Gomez et al., 2019). The towers are designed to handle a significant amount of data traffic and have the capacity to manage multiple connections simultaneously, making them particularly valuable in densely populated areas where network congestion is a concern. The maintaining and upgrading existing tower infrastructure also be expensive. To cater this issues different technological advancements contribute to manimize the cost and get maximum profit from it (Jia et al., 2020).

Gomez et al (2020) stated that fixed telephony, even mobile voice services have become a necessity as a variable that influences growth. The next generation broadband technologies plays a positive role on the economy of the region and helps the economic activities to flourish.

Jiwu et al., (2018) worked on base transceiver station (BTS) site selection and concluded that sites selection of BTS tower is one of the most important parts of wireless broadband service planning. Appropriate site location is an essential factor in determining the better coverage performance, signal strength and telephone traffic distribution.

Identifying suitable sites can be determined by using different technologies. Akeem Babatunde (2017) worked on analyzing telecommunication towers using GIS and GPS techniques and observed that these technologies plays vital role in the planning and management of telecommunication towers.

Saikhom et al., (2016) adopted a geospatial approach for identifying optimal sites for broadband towers. Five contributing factors including land use, elevation, existing towers, and soil datasets are processed using proximity and viewshed analysis. Multi-criteria spatial

modeling using a fuzzy membership-based overlay approach is used to identify the suitable location.

Rofii et al., (2016) worked on calculating the coverage areas of base transceiver station and eliminating the existing towers in the overlapping coverage zones using fuzzy clustering means and particle swarm optimization. The aim of the study is to count and geographically locate the existing telecom infrastructure and analyze the growth of telecom towers in the study area. It was analyzed by using particle swarm optimization (PSO) and fuzzy clustering techniques to produce the computational intelligence. They concluded that the PSO techniques are useful to accelerate the achievement of convergence. Merger technique and BTS sharing can reduce the number of towers in the same location.

Arthur et al., (2019) discussed the tower sharing policy and analyzed the sites of deployed telecom infrastructure in Ghana. They focused on identifying the towers of all eight operators working in the study area. Performed overlay analysis on coverage prediction of each operator to check the overlapping zones and towers that can be shared to other operators. They resulted that the cost of analyzing the cost and benefits of infrastructure sharing helps in optimizing the network and removing congestion and minimizing the cost. They concluded that infrastructure sharing policy can save 44.61% of capital for all operators while sharing common towers. It decreases in overall cost allocation for capital expenditure (CAPEX) and Operating expenses (OPEX) of the telecom infrastructure.

Casier et al (2006) worked on cost allocation for capital expenditure (CAPEX) and Operating expenses (OPEX) for a network service provider. Telecom infrastructure is costly, and it requires a handsome amount of money for installation, implementation and maintaining of network. To efficiently use the investment in both CAPEX and OPEX can make it sustainable. They used top-down and bottom-up cost modeling approaches to allocate fair capital for both categories and

concluded that it depends on the nature of the services. Some expenses are seen as direct cost and directly allocated for that because but it varies from 10 to 50% in the real telecom environment.

### **1.1 Role of geoinformatics in telecommunication**

Geoinformatics is an emerging technology which address and solves the problem of geography, geosciences and related branches of technology. Geographic information system commonly known as GIS is an important tool that facilitates in planning, implementation, monitoring and live tracking of projects in different sectors. It facilitates in environmental management decision making, land use planning, suitable sites analysis, infrastructure mapping, asset management, optical fibre route planning and it is the best approach to discover the optimum zones and sites. GIS is now a standard technology applied throughout the telecoms industry. (Broek et al., 2019). 80% telecom infrastructure data is geographic in nature and GIS can process big data based on conditional criteria to analyze geo-telco data and visualize, manage, and automate the workflow for moving towards the smart decision making (Sobral et al., 2018).

Geospatial technologies help to make the planning process easy, telecommunication and GIS are different fields but they can be combined together to make useful workflows to automate the processes and create the opportunities to ease the planning process for telecom infrastructure.

Narbaev et al., (2021) implemented GIS to modernize the telecom networks on the basis of identifying the demographic patterns with the help of GIS. They analyze the geographic data of households, educational institutions, enterprises and organizations and their demography to provide network and internet coverage, improve the efficiency of fiber optic cables. They developed a six-step methodology which is based on cartography techniques. Creating



large-scale digital maps using statistics and clutter datasets. Moreover, They concluded that geospatial demographic gridded and census data become an important tools for performing reserach related to demography and it helps to estimate the requiement of an area for a particular service.

Asassafeh el al., (2020) worked on developing a new approach for choosing the optimal sites for broadband towers. They classify the contrbuiting factors that contributed in sites selection for telecom towers and assigning the weights to perfrom weighted overlay analysis in ArcMap software. They conclude that this approach is good for chosing the BTS sites and while using multiple socio economic factors.

Avikal et al., (2021) worked on multi criteria decision making (MCDM) for selecting suitable sites for installiation for BTS towers. They found that the predicting factors and their importance plays an important role for telecom towers. They used four factors which were analyzed through MCDM and analytic hierarchy process (AHP). They concluded that running and maintenance cost are showing highest importance the most important criteria and noise pollution showing its least variable importance.

Khilare et al., (2021) worked on techniques for supply chain management and time & cost saving approach for BTS towers. Relative important index (RII) method was used to perform analysis on 16 datasets that are collected using questionnaire surveys. They concluded that RII method reduces the time & cost of the construction of ground base cellular tower.

Amiri (2021) established location-allocation model using Genetic Algorithm (GA) for BTS anntannas. The Delphi method is implemented in MATLAB simulator to optimize the locations for BTS towers. And then them GA is performed on it. They concluded that the integration of GIS along with genetic algorithm resulted in faster convergence, effective reduction the number of antennas, and significantly reduction of the service costs.

Premarathne et al., (2021) proposed suitable sites for establishing BTS tower, while considering 08 parameters. population density, existing tower locations, proximity to roads, land use, reservations, slope and educational institutes. Weighted sum techniques was used along with the weights that are assigned based on expert opinion to perform analysis in Arcmap. The sites suitability maps of tower are generated based on the resultant suitability levels of weighted sum analysis. They concluded that this methodology can help to identify zones for establishing new BTS towers.

Tayal et al., (2017) worked on locating optimal positions for telecom towers and perform site suitability analysis in Uttarakhand using GIS. They used eight factors including administrative boundary, roads, rivers, forest cover, digital elevation model (DEM), block-wise population, existing served area, and location of the existing tower are prepared using GIS software. Buffer analyses are used to calculate the coverage range of the existing tower and highlighted the dead zones and overlapping areas. They concluded that this methodology is useful for proposing suitable sites and it helps in removing dead and overlapping coverage zones.

## **1.2 Role of machine learning in GIS**

Chen et al., (2018) discussed the machine learning methods, their applications in remote sensing and GIS, . Its major application in supervised machine learning, unsupervised learning, semi-supervised learning, and reinforcement learning. They reviewed several specific machine learning techniques and their applications in remote sensing and GIS, including artificial neural networks, decision trees, multi layer perceptron (MLP) and support vector machines. They discussed the pros and cons of machine learning in geospatial technologies, including the need of big training data and the difficulty of interpreting the results of complex models.

Zhang et al., (2020) used machine learning techniques for traffic prediction from a geospatial perspective. They begin by discussing the importance of traffic prediction and the challenges

it poses, including the need to incorporate both static and dynamic factors and the complexity of modeling human behavior. They reviewed several specific machine learning techniques and their applications in traffic prediction, including linear regression, artificial neural networks, support vector machines, and extreme gradient boosting. They also discuss the use of geospatial data and techniques such as spatial-temporal data mining and GIS-based modeling in traffic prediction.

Machine learning is useful technology to automate the sites suitability analysis and its implementation is growing rapidly due to its accuracy and efficiency. These modelling techniques are more robust than using single classifier. (Mokarram et al., 2015)

Yang et al., (2015) combined machine learning with web GIS for evaluating hotel sites which are reliable, unbiased, the location of the hotel is very important. They presented a new automated web GIS approach for evaluating potential sites for proposed hotel properties and developed a new toolbox which was hotel location selection and analyzing toolset (HoLSAT). They used multiple factors including location of existing hotels, stars, accessibility etc and then perform machine learning on it to propose the potential sites and concluded that this methodology shows good results and it is useful for the evaluation of hotel locations.

Arabameri et al., (2021) used geospatial machine learning approaches for modeling groundwater potential. Classification and regression tree (CART), random subspace (RS) with the multilayer perceptron (MLP), and naïve Bayes tree (NBTree) machine learning techniques are implemented on spatial data. The primary data was collected using field survey approaches of 205 spring locations with 14 contributing factors which are curvature, elevation, aspect, slope, terrain surface texture, lithology, proximity to faults, fault density, rainfall, distance to streams and land use/land cover are used as model inputs and segmented the data into 70:30 ratio for training and validation of models. The model outputs are evaluated on 06 statistical

matrices which are accuracy, f-score, Kappa, ROC curve, sensitivity and specificity. They concluded that multilayer perceptron is the best model with the accuracy .933 on the validation datasets.

Wang et al (2020) worked on sites selection using machine learning and empirical approach for digital signage in Beijing. They focus on providing the new sites selection model for accurate signage locations that integrates the spatial information of multiple factor data and combine empirical locations with the machine learning models for proposing locations for signages. Huff model is used as the empirical approach to calculate the spatial accessibility, machine learning for identifying potential locations and overlay analysis for obtaining the deployable sites. They concluded that the proposed methodology has higher accuracy for sites selection amongst all models. It improves the accuracy and efficiency up to a certain extent.

GIS helps to identifying areas of poor coverage, potential business zones, and sites for installation of base transceiver station (BTS) towers, planning, implementation and monitoring of optical fibre cable (OFC) routes and nodes, BTS connectivity, identifying the line of sight barriers along with mapping the skyscraper, building heights, dead coverage zones, overlapping coverage zones, optimizing OFC network infrastructure, and predicting future capacity requirements. (Narbaev et al., 2021).

With the advancement in technology, Machine learning and predictive modelling are widely applicable in many disciplines including GIS. Geospatial machine learning (GML) is an emerging field which is the subset of machine learning that uses spatial data and geospatial techniques to analyze it from multiple perspectives and make predictions about the geographic phenomena. It is used in planning processes, geo-business intelligence, churn prediction and it assumes as a powerful tool for sites suitability analysis, as it allows for the integration of both

spatial and non-spatial data to create models that can predict the suitability of a site for a particular use (Al-Ruzouq et al., 2019).

Jiang et al., (2017) discovered geospatial intelligence machine learning frameworks for searching ranking algorithms. They examined that most search engines only focus on few parameters such as popularity and release date which failed to consider multi-dimensional users' preferences including geographic information, which resulted in inconvenient and bad experiences for the users. A number of ranking features were identified by using machine learning to segment data into different parts based on user, system architecture for using machine learning with knowledge-based approaches for ranking feature is proposed to combine software with machine learning to automatically learn a ranking function. They resulted that machine learning approach outperforms the rest of approaches and it helps to improve the search ranking in the geospatial context for the users.

Shirzadi et al., (2018) worked on geospatial machine learning approaches for landslide susceptibility mapping. They focused on introducing a novel machine learning algorithm of bagging (BA), rotation forest (RF), alternating decision tree (ADTree) based on the multiboost (MB), and random subspace (RS) based on two scenarios with different samples size and resolutions for spatial predictions. The models are evaluated on different statistics and ROC curves, the data segmented into 60 and 40% with the spatial resolution of 10 and 20 meters to train and validate the models. They concluded that this novel geospatial machine learning approaches are important and useful alternative tools to assist decision makers and planners to manage the tasks of managing landslide-prone areas.

GML is a powerful tool for performing site suitability analysis with enhanced accuracy and the power of learning from hidden patterns. It works on different data formats as it has capabilities and allow the integration of spatial and non-spatial data to create, train and validate models

that can predict the suitability of a site for a particular use. These models can be used to support decision-making and to identify the most suitable location for a project, it is a time-saving approach along with the easy deployment over the world wide web. It can be open source so its running cost is minimal to the commercial software that provides site suitability analysis facilities (Walter et al, 2022).

The above-mentioned pieces of the literatures disclosed that many studies on broadband tower site suitability were conducted worldwide using different geospatial approaches. However, a comparative assessment of site suitability using machine learning approaches has not been carried out for broadband towers.

### **1.3 Research gap**

The literature that are discussed above provide valuable insights on the importance and implementation of geospatial technologies for planning of broadband towers. The findings demonstrate that identifying suitable sites for BTS towers is essential for seamless coverage and for eliminating the dead zones. Despite the multitude of studies identifying suitable sites for broadband towers using different geospatial analysis approaches, there remains a lack of a comprehensive and integrated geospatial machine learning approach that considers the multiple socio-economic and environmental factors to propose suitable sites from a smart planning perspective. Additionally, the current literature overlooks the comparison of different machine learning approaches and checking the classification quality from multiple angles. Therefore, the current study aimed to make a comparative analysis between three machine learning models which are Random Forest (RF), Xgboost, and Support Vector Machine (SVM) by implementing on multiple geospatial datasets. The results will be displayed on the map. The model prediction performance will be checked using the area under curve (AUC) and results are evaluated based on specificity, the AUC, sensitivity and cross-validation.

## 1.4 Motivation

Many countries are facing unavailability of good quality of broadband coverage and even voice coverage are not available in some areas. Like other countries, Pakistan which is one of the developing countries is facing the same issues. As it spread over the large area with different topography, the uneven distribution of population in the region and one of the top ten populated country of the world. The demands of broadband services are increasing in the regions due to its consumption in different sectors and its highly important for distance learning , remote jobs, technologies busienss and tele-health. It consider as the significant contributor in the economic growth of the country. To cater the high demand for seamless broadband coverage , different countries adopted different stratagies to fulfill the demand such as network sharing , Artificial intelligence (AI) based planning system, but Pakistan is still using traditional pratices which consume a lot of resources in terms of insfrastructure , human resources and equipments. The geospatial machine learning is popular to provide the support to the planner to make the planning process easy and fast with the automation of different stage and the capabilities of machine learning models. It is the trending technology that started changing the world with its processing speed and self-learning capability. The working on geospatial machine learning models for broadband towers sites suitability will reduce the time of planning and in deployment phase which helps to provide the coverage faster to the usersa and it also contributed to the economy. The reason for selection of this topic is to implement multiple geospatial machine learning (GML) models to identify suitable sites for next generation broadband tower and suggest the best working model with higher level of accuracy along with other important metrics which can work on open source data and planner can get insight in no time while sitiing in the office. The machine learning models is not just facilitates in site suitability but also its cost effective and time-saving and ease the lives of the telcom planners.

## **1.5 Relevance to national needs**

Telecommunication is one of the major contributing sectors in Pakistan economy. According to the Pakistan Bureau of Statistics (PBS) Survey report 2017-18, it supports the economy and accounts for 26% of gross domestic product (GDP) of the country. Its contribution is growing annually to the GDP. It works as a catalyst in other sectors such as agriculture for precision farming, automated pesticide spraying, drone mapping and crop health identification. It is important for tele health sector which is of the major pain point in country due to major population lives in rural areas and accessibility to the health facilities are minimal and it can help to bridge this gap. Pakistan is world fifth most freelancer producer country in the world and good seamless broadband coverage is one the basic needs for freelancers' community. Workers from many areas of the country are providing the services in different parts of the world. Apart from direct impact, it provides opportunities to Pakistani youth to excel in information and technology sector throughout the world in the form of remote workers. High-speed broadband connectivity is one of the primary requirements for communication in the world so it's very important to design a model that gives quick results. The aim of this study is making network planning easy and flexible which contribute to achieve the goals of digital Pakistan initiatives. Its cost and time saving approach, so it reduces the project implantation timeline.

## **1.6 Objectives**

The objectives of this research are given below:

1. To proposed suitable sites for broadband towers using machine learning approaches.
2. To evaluate the resultant sites suitability of each model based on sensitivity, specificity, area under curve (AUC) and cross-validation.



### MATERIALS AND METHODS

#### 2.1 Study area

- 2 The study area used for conducting this research is Islamabad, the capital city of Pakistan, which is in the northern part of the country as shown in figure 01. It has 2,001,579 estimated population as per the 6<sup>th</sup> census conducted in 2017. It has a total area of 906 km<sup>2</sup> which are divided into 05 zones with the mean elevation of 506 m. it is planned city which was built in 1960's. It is surrounded by Rawalpindi, Punjab from the south, east and west side, while Haripur district, Khyber Pakhtunkhwa from the north side. The city is well-developed with modern infrastructure, and it is known for its natural beauty and a well-planned layout. It's one of the popular tourist destinations, known for its scenic beauty, historic monuments, and cultural heritage. The city also has a growing economy with several industries and businesses operating in the city. It has the reputation of business minded capital. The demand of high-speed broadband connectivity is one of the major priorities of its citizens. So, fulfil this gap in the markets different operators come to avail this opportunity. So apart from competitive telecommunication market potential in the study area, the overall, it is a well-developed and modern city with a strong economy, a well-developed telecommunications infrastructure, and a growing number of businesses and industries. Islamabad has a well-developed telecommunications infrastructure with several telecom companies providing services in the city, including PTCL, Zong, Jazz, Telenor, and Ufone. These companies offer various packages and plans for both prepaid and post-paid customers. In addition, there are several international and local internet service providers (ISPs) that offer broadband and fibre internet services in Islamabad.

## **2.2 Methodology**

The research methodology is based on three major phases which are depicted in flowchart diagram which shown in figure 2. The first major portion is data acquisition from different sources and transform it into a geographic data format, the second one is preprocessing of that dataset and making it ready for running machine learning models on it. The third step is to segment the data which is adopted to divide data for models training and models validation purposes, different factor optimization and k-fold cross validation techniques are applied to train and validate machine learning models. Moreover, the final part is to compare the results of each model based on different metrics and then propose the best model for broadband tower sites suitability analysis.

### **2.2.1 Data acquisition**

This research work is data intensive and to find out the suitable sites for BTS tower, it required data, and its acquisition is the primary step. Data acquisition refers to the process of gathering relevant data from authentic sources which used for answering the research questions ad helps in achieving research objectives. To achieve the objectives of present study, multiple contributing factors are considered for performing analysis which are shown in table# 01. To answer the question of why these datasets are used in the study is one the major question which is answered thorough reviewing of existing literature related to tower site suitability was conducted to determine the reasons for using these datasets in the study. These datasets were chosen based on references from various research papers, as highlighted in Table 02. The focus of the study is on identifying factors related to tower site suitability. The selection of datasets is informed by the existing literature in the field using different research papers served as a reference for selecting the factors included in the datasets.

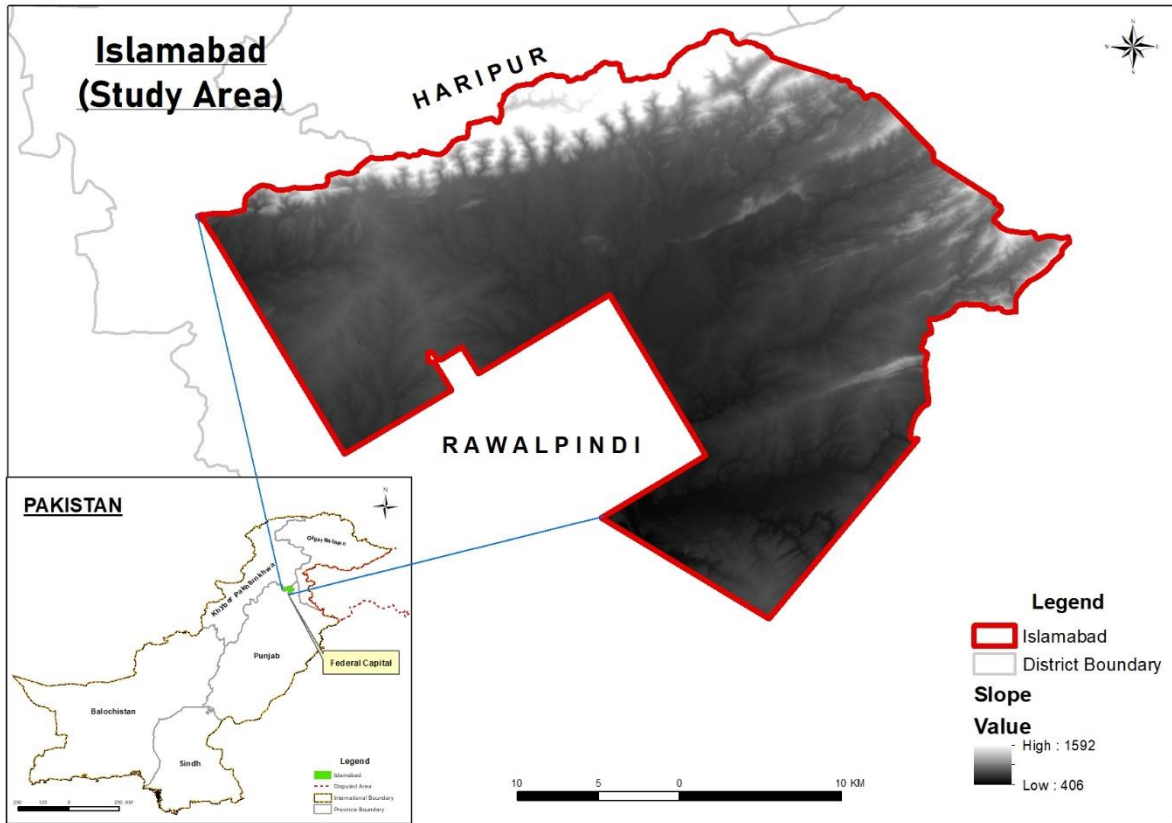


Figure 1. Study area map showing Islamabad with its elevation.

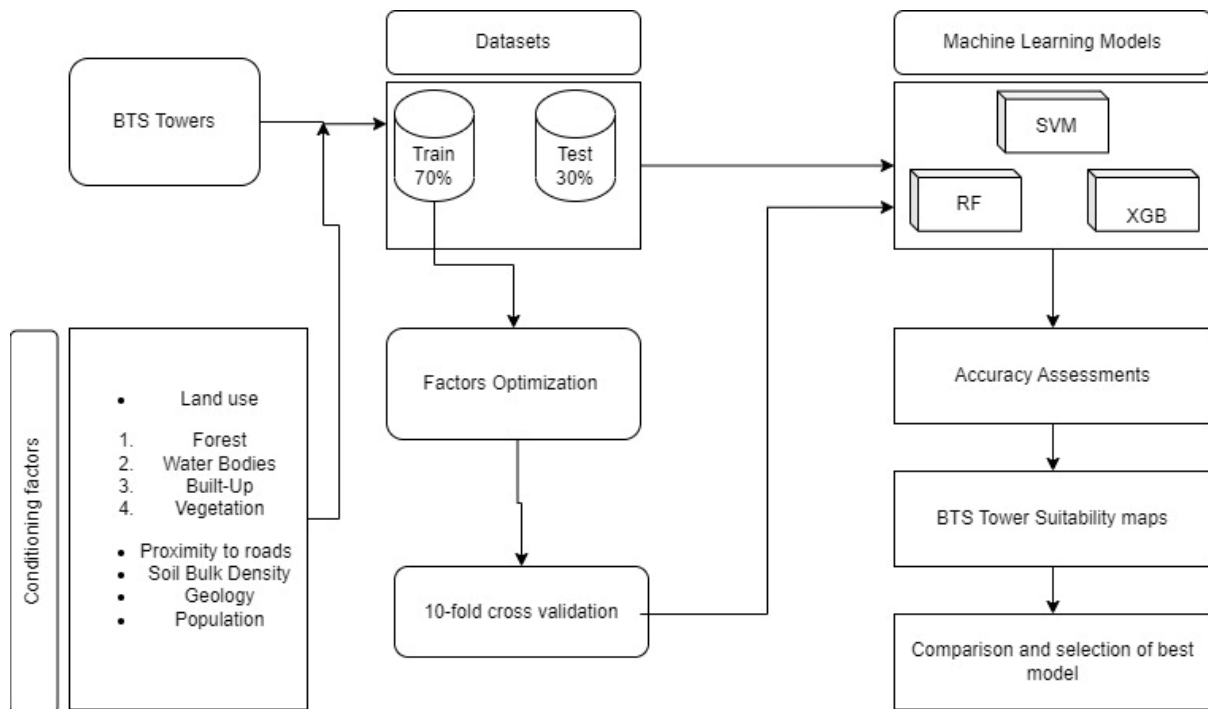


Figure 2. flowchart of methodology adopted in the study.

Table 1 showing dataset used in this study.

<b>S. No</b>	<b>Data</b>	<b>Source</b>	<b>Data Type</b>
1	Land use Data	Extracted from Landsat 8	Raster
2	Slope	SRTM Digital Elevation Model	Raster
3	Road Network	Open Street Maps	Vector
4	Existing Tower	Universal Service Fund Co	Vector
5	Soil Bulk Density	ISRIC World soil information	Raster
6	Geology	Geological survey of Pakistan	Vector
7	Population	Land Scan population data	Raster

Table 2 showing references for parameters selection.

<b>Data used</b>	<b>Premarathne et al (2021)</b>	<b>Saikhom et al (2018)</b>	<b>Tayal et al (2017)</b>
Land use	✓	✓	✗
Slope	✓	✓	✓
Population	✓	✓	✓
Existing tower	✓	✓	✓
Roads Network	✓	✗	✓
Geology	✗	✓	✗
Bulk Density	✓	✗	✓

### **2.2.1.1 Landuse data**

Landuse data is one of the key factors for performing analysis. It is referring to information that categorizes and describes the several types of land use within the study area. It provides insight into how the land is utilized by humans for various purposes. Its data are extracted from landuse classification on satellite imagery, aerial photographs the digitizing the existing landuse maps. Different software's provides the tools for performing landuse classification including Qgis, ArcGIS and Google Earth Engine (GEE). GEE is basically a newer approach that doesn't needs to follow the process of downloading satellite imagery from different sources and then perform analysis on it. It provides access to all the data in a single planform and runs classification on cloud server which has high speed to process imageries in short time. In the present study the landuse classification is performed in Google Earth Engine (GEE) on Landsat 08 imagery. The Landsat 08 imagery is based on different bands which are basically stored the response of objects with different wavelengths with spatial resolution of 15, 30, 60 and 120 meters. The red, green, and blue bands of OLI imagery that was captured in 2022 having cloud cover >10% are used for classification. 120 training samples are used against each class and classify it into five classes using supervised classification. The classes which are forest, water bodies, vegetation, barren land, and built-up areas. The classification resulted 187,322,400 sq. meters areas in vegetation class, 125,763,200 sq. meters barren land, 452,387,600 sq. meters built-up area, 149,798,700 sq. meters forested area and 9,272,500 sq. meters water are present in the study area. The quality of classification is checked in GEE and exported with the spatial resolution of 30 meters. The resultant raster output is transformed into universal transverse Mercator (UTM) projection zone 43N with the spatial resolution of 30 meters and the radiometric resolution of 32 bits. The error matrix is also run-on classification to check the accuracy and quality of classification results. The final output landuse dataset map is shown in figure 03(a).

### **2.2.1.2 Population data**

The population of an area helps to decide the making of business cases and installation of broadband towers. It is the information about the numbers of people residing within a particular potential coverage range of towers. It helps in smart planning, strategic network optimization and decision making for coverage and improving service quality. The population data are obtained from various sources which are census data reports, population density grid and survey etc. Many organizations are working on population data preparation by combining various sources. Land scan is one of them the It produced the population data by combining the census data along with spatial data and imagery analysis technologies on global scale to produce the gridded datasets. The population density data was acquired from Landscan global population database that is the industry standard for global population distribution with a spatial resolution of 250 meters. It can be downloaded from landscan website. This gridded dataset is downloaded with the spatial resolution of 250 meters, reprocessed it and extract the study area in ArcMap. The reclassification techniques are applied and classify it into 5 classes for better understanding of population density and reproject the raster into UTM zone 43N with the spatial resolution of 30 meter and radiometric resolution of 32 bit.

### **2.2.1.3 Bulk density**

Bulk density of soil data acquired from international soil reference and information Centre (ISRIC). It refers to the mass of a material per unit volume, typically expressed in grams per cubic centimeter ( $\text{g}/\text{cm}^3$ ). It is a measure of how densely packed or compacted a substance. It is an essential parameter in various fields, environmental science, construction, and mining. It helps in understanding soil compaction, soil fertility, water retention, and nutrient availability. Bulk density is the mass of bulk solid that occupies a unit volume of a bed, including the volume of all interparticle voids. Considering that a powder is really a particle gas mixture with both interparticle spaces and intraparticle voids, three classes of bulk density have become

conventional: aerated, poured, and tap. Different types of soil have different typical ranges of bulk density, and variations can indicate differences in soil properties and health. It is calculated by dividing the mass of the particles by their total volume. The formula indicates the soil's ability to behave as a structural support, a diffusion zone, a water source, and aeration zone. Porosity indicates the volume fraction of void space or air space inside a material. Volume determination is relative to the amount of internal or external pores present in the powder. It is available on ISRIC website as a gridded dataset that is recorded with different depths. The data which is used in this study is calculated from the weight of soil in a  $30 \text{ cg/cm}^3$ . The downloaded data is transformed into UTM zone 43N and exported as raster data with the spatial resolution of 30 m and radiometric resolution of 32-bits. Its values range from 1360 to  $1720 \text{ cg/cm}^3$ .

#### **2.2.1.4 Road network**

The roads dataset is extracted from OpenStreetMap (OSM). The data are available in the OSM format. The roads and highways data are downloaded in OSM format and then converted into shapefile using global mapper. It contains all types of roads including walking, cycling tracks and attribute information of a lot of records with missing. To cater to this issue, the data cleaning techniques were applied to remove the unwanted data and make it ready for further analysis. The multi-ring buffer analysis was performed on the cleaned data to create different zones according to national highway authority (NHA) rules and analysis requirements. NHA construction is not allowed in the 45 feet's zone of both sides of central line of the single road and 0.5 kilometer from the motorway interchange. As per the with the distance of 15 m, 45m, 90m and 120 meters on both sides of the roads. The buffer analysis output is transformed into UTM zone 43 N and then transformed it into raster data with the spatial resolution of 30 x 30 meters and with the radiometric resolution of 32-bit. The final proximity to roads data is ready for analysis and its map representation is shown in figure 03(f).

Table 3 showing data preprocessing of all factors.

<b>Factors</b>	<b>Spatial resolution</b>	<b>Radiometric resolution</b>	<b>Pixel grid</b>
Geology	30 Meters	32 bits	1798 x 1214
Land-Use	30 Meters	32 bits	1798 x 1214
Slope	30 Meters	32 bits	1798 x 1214
Bulk Density	30 Meters	32 bits	1798 x 1214
Proximity to Roads	30 Meters	32 bits	1798 x 1214
Population	30 Meters	32 bits	1798 x 1214

Table 4 all attributes (factors) data in BTS data layer

<b>BTS site</b>	<b>Lat</b>	<b>Long</b>	<b>Land use</b>	<b>Geology</b>	<b>Slope</b>	<b>Proximity to road</b>	<b>Bulk density</b>	<b>Population</b>
1	33.703	72.978	7	6	496	5	1519	10660
2	33.651	72.953	4	4	538	2	1534	4987
3	33.654	73.041	3	3	548	1	1535	11366
4	33.640	73.187	1	1	490	3	1543	3203
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
n	n	n	n	n	n	n	n	n



### **2.2.1.5 Geology**

The geological datasets are extracted from geological map of Pakistan, which is available on the internet, georeferenced it uses image to maps technique with the affine transformation in ArcGIS software. The rock types present in the study are undifferentiated Paleozoic rock, Triassic metamorphic and sedimentary rock, tertiary sedimentary rock, quaternary sediments, Paleogene sedimentary rocks and Neogene sedimentary rock. Paleozoic rocks, also known as undifferentiated rocks, are composed of sedimentary and metamorphic rocks formed between 541 and 252 million years ago. These rocks are not specifically classified, and these are diverse including sedimentary, igneous, and metamorphic rocks that formed in that era. Triassic metamorphic and sedimentary rocks were formed during 252 to 201 million years ago known as Triassic period. As rocks transform under high pressure and heat, metamorphic rocks are formed, whereas sedimentary rocks are formed as sediment accumulate and stratify. Tertiary sedimentary rocks were formed during the tertiary period, which spanned from about 66 to 2.6 million years ago. These rocks are primarily sedimentary in nature and may include formations such as sandstones, shales, and limestones. Quaternary sediments represent the most recent geological period, which extends from about 2.6 million years ago to the present. These sediments include a wide range of materials deposited during the Quaternary Period, such as alluvium, glacial deposits, and volcanic ash. Paleogene sedimentary rocks were formed during the Paleogene Period, which occurred approximately 66 to 23 million years ago. These rocks are predominantly sedimentary and may include sandstones, shales, and coals. Neogene sedimentary rocks were formed during the Neogene Period, which spans from about 23 to 2.6 million years ago. These rocks are primarily sedimentary and can consist of sandstones, siltstones, and conglomerates. Manually digitize the geological types of the study data and its attributes are updated accordingly and reproject the output it into UTM Zone 43N using

transformation and projection technique. It transforms it into raster format with the spatial resolution of 30 meters and radiometric resolution of 32-bits.

#### **2.2.1.6 Slope**

The slope dataset was downloaded from opentopography.org which provide digital elevation model (DEM) data that was captured and processed under shuttle radar topography mission (SRTM) with the 30 x 30 meters spatial resolution in Georeferenced Tagged Image File Format (Geo TIFF) format. DEM provides insights about the height of the earth surface from the mean sea level while slope represents the steepness or inclination of the terrain. It is calculated in elevation over a given distance. In this study the shuttle radar topography mission (SRTM) DEM acquired, and slope values are calculated based on neighboring elevation values. It ranges from 0 to 90 degrees, then classify into five different classes using natural break data classification method. 0 indicates the flat slope while increasing in value indicates the steep slope and 90 degrees representing vertical or near-vertical slopes. Once you have the slope data, it can be used for various purposes such as terrain analysis, line of sights, shadows, modeling, and slope stability assessment, among others.

#### **2.2.1.7 Broadband towers**

The broadband tower dataset is acquired from Universal service fund (USF) Co, in that was available in tabular format. This data is based on all BTS towers installed by all four operators Jazz, Zong, Telenor and Ufone. They installed the antennas on existing towers installed by other operators. BTS towers were shown adjacent to each other, so it required data cleaning. The data cleaning techniques performed to eliminate duplicate values based on geometry, remove the operators and sites ID information and store as comma delimited values (CSV) format, so it can easily be transformed into geospatial data. Tabular data converted into geographic data in point features.

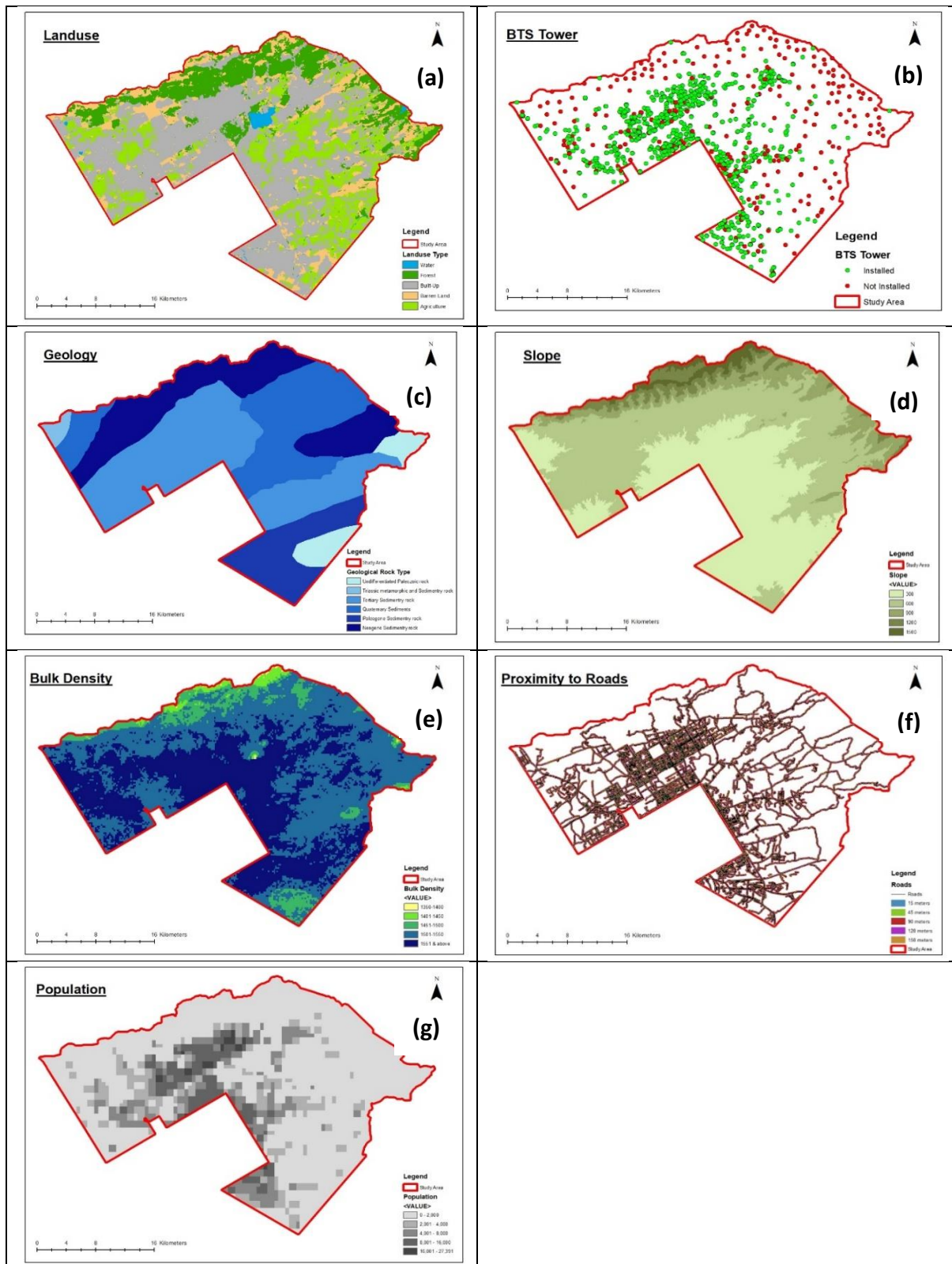


Figure 3. showing datasets maps of all variables.

### **2.2.2 Data standardization**

After data preprocessing, datasets are standardized in same format with the spatial resolution of 30 x 30 meters. The radiometric resolution is standardized to 32 bits floating point values. The extent of all the raster is standardized to 1798 x 1214-pixel grid. The projection transformations of all the raster's into UTM zone 43N. All the layers are converted into American standard code for information interchange (ASCII) for transforming it into 2-D arrays. Machine learning analysis is performed on numeric data and arrays are the best ways to handle data using different data science libraries in python programming language. The BTS tower data and all contributing factors values are extracted using Multivalued to point function in ArcMap stored in the attribute table of BTS dataset as shown in table 04. The finalized representations of all datasets are shown in figure 03.

### **2.3.2 Data modelling**

The prepared datasets are used as input to 03 machine learning models which are support vector machine (SVM), random forest (RF) and XGBoost (XGB). The total data samples which were 1400 in total contains both suitable site's locations with the class value of "1" and the unsuitable sites location with the class value of "0". The data segmentation approach is applied on it and divided the total data into two segments with the ratio of 70% and 30 % using data partitioning technique. 70% data. Models are trained using training data with the help of implementing the 10-k fold validation techniques and get the mean values of accuracy and standard deviation of the statistical matrices while 30% dataset are used for model validation. The code for implementation of machine learning is shown in appendix 02.

#### **2.3.2.1 Random Forest (RF)**

Random forest models can solve both classification and regression tasks. It is a popular machine learning algorithm that has been widely used in spatial data analysis. In spatial data

analysis, random forests are often used for land use classification, spatial interpolations, and spatial regression. It divides the datasets into different parts that's used by multiple decision trees, and then combines the output to develop a more accurate and robust model.

As a decision tree, it utilizes supervised machine learning algorithms to solve problems such as classification and regression. Based on feature-based splits, the decision trees show predictions in a tree-like structure. There are 3 components of a decision tree which are a root node, a decision node, and a leaf node. A root node is the point where the population start dividing, decision nodes are the points where that population splits and the node where no further data splitting is possible known as leaf node. It begins at the root node and ends at the leaves with the decision. The selection of nodes based on Gini index and entropy. Likewise, this algorithm will try to find the lowest Gini index among all the splits possible and then choose the feature as the root node. If the Gini index is the lowest, split impurities are measured by entropy as well.

The major advantages of random forest in spatial data analysis are its ability to handle high-dimensional data with many variables, such as remote sensing and environmental data.

With the RF algorithm, multiple decision trees are built with randomly selected subsets of training data, reducing the correlation between trees and increasing forest diversity by building multiple decision trees with randomly selected features. All predictions from the decision trees are aggregated during predictions by the random forest algorithm to make final prediction. It can incorporate various types of remote sensing data, such as spectral bands, vegetation indices, and texture features, to improve classification accuracy.

### **2.3.2.2 Support vector machine (SVM)**

Support vector machine (SVM) is an algorithm widely used for classification and regression analyses. It is supervised learning techniques that are well-suited to binary classification problems, where the goal is to separate data points into two binary classes.

SVM algorithms are based on finding the boundary that divides two classes of data and maximize their margin, which is a decision boundary. It finds the hyperplane that best separates the two classes of data by maximizing the amount of margin between them. It searches for the hyperplane that produces the greatest margin between the hyperplane and the nearest points in each class.

Due to its ability to process high-dimensional data with nonlinear relationships between variables, support vector machines have gained increasing attention in the field of spatial data analysis. In spatial data analysis, SVM is often used for classification and regression tasks. Spatial data often contains many variables, such as topographic and environmental data, that can make modeling difficult. SVM can handle these high-dimensional datasets and can identify the most important variables for prediction. SVM's kernel functions, such as radial basis function and polynomial kernel, can also help to capture complex relationships between variables.

Its working is based on different steps, firstly it predicts the classes between 1 and -1. Like all other machine learning algorithms, it converts business problems into an optimization problem. SVM classifier uses the loss function known as the hinge loss function to find the maximum margin while optimizing problem. Mathematical concepts of calculus and partial derivatives are used to optimize the weights and regularization parameters are used to eliminate the loss function in the final classification and perform classification using Support Vectors, soft margin, hard margin, and different kernels.

### **2.3.2.3 Extreme gradient boosting (XGBoost)**

Extreme gradient boosting (XGBoost) is a powerful machine learning algorithm that is used for regression and classification analyses. To create a more accurate and robust model, multiple weak learners are combined, typically decision trees. The results of each decision tree are used as variables to boost the accuracy of the next decision tree to make it a more robust and accurate model. In XGBoost algorithm, the gradients of the loss function are calculated, then used to update the model weights. The gradient based update is performed iteratively on each decision tree before adding the next one and updated with a gradient descent optimization method. This approach helps the algorithm to focus on the most challenging data points and make more accurate predictions.

XGBoost has been found to outperform other machine learning algorithms, such as logistic regression, CART, naïve bayes, K-Nearest Neighbors algorithm Multilayer perceptron and Support Vector Machine. It can incorporate various types of remote sensing data, such as spectral bands, texture features, and topographic indices, to improve classification accuracy. Moreover, XGBoost can handle class imbalance and provide probabilistic estimates of class membership, which is useful for uncertainty analysis.

XGBoost is a powerful and versatile algorithm that has shown promising results in various spatial data analysis tasks. Its ability to handle high-dimensional data, handle missing data, and provide important feature measures. makes it an attractive option for researchers and practitioners working with spatial data.

### RESULTS AND DISCUSSIONS

The results of machine learning techniques for predicting the sites suitability of BTS towers are presented in this chapter. A comparison is made between random forests (RF), support vector machine (SVM) and extreme gradient boosting classifiers to determine their suitability for broadband tower sites. The detailed presentation of the experimental results for each method are presented. The comparative analysis is conducted to compare and discuss the results obtained from each classifier. To ensure a robust experimental evaluation, the tests should be expanded and repeated across multiple exploratory variables using cross-folding techniques. To achieve this, 10 cross-fold(k) are applied on all exploratory parameters, each consisting of features from 900 actual suitable sites (positive examples) and 500 unsuitable sites (negative examples) with the random ratio of 70%, and 30% of the data was used for training, and testing, respectively for all three selected models (SVM, RF And XGBoost).

#### 3.1 Random Forest

##### 3.1.1 Accuracy

Random forest works well in learning from the input data, with a 96% accuracy rate on training data, indicating that it works well in learning from input data. It segmented the data into different decision trees and execute trees simultaneously. This accuracy calculates the mean accuracy at training levels. Moreover, the trained model runs on validation data to test how well it can classify the data without classification labels. Its accuracy on the validation data is 89.78% which decreases from training level accuracy. This predicated accuracy of random forest model is considered as good machine learning accuracy as it can accurately classify 89%



of the time which is reliable for implementation. The accuracy of random forest at both training and validation level are shown in figure 04.

### **3.1.2 Relative variable importance**

Random forest resulted the relative variable importance from all six variables which shows that proximity to road is 77%, population 82%, slope 68%, and landuse has 78% relative importance while geology has 29% and bulk density has 26% relative importance. It indicates that geology and bulk density are the least important criterions for broadband sites suitability and landuse, population, slope and proximity to roads are more important criterions for performing sites suitability analysis for broadband towers. The graph of relative variable importance is shown in figure 8(d).

### **3.1.2 Sites suitability for broadband towers**

The models predicted the results in the form of continuous raster layers with the digital number (DN) values that range between 0 and 1. The suitability level for broadband towers is calculated by dividing the data into four divisions with the suitability level which are highly suitable, moderately suitable, marginally suitable, and unsuitable. It resulted in 150 sites in highly suitable class, 87 sites in moderately suitable class, and 79 sites in marginally and 104 sites are in unsuitable class from total 420 sites. The proposed suitable sites for broadband towers resulting from random forest classification are shown in figure 05.

### **3.1.3 Predication quality**

Receiver operating characteristic (ROC) curve is used to check the quality of classification results. RF resulted with the area under curve (AUC) score is 0.89 which indicates that 89% of the time, the model correctly classifies the sites into their suitability level class as shown in figure 8(a).

## **3.2 Support vector machine**

### **3.2.1 Accuracy**

Support vector machine (SVM) model scored a 60% accuracy while training on training datasets. Its accuracy on training is considered the random results in learning from the training inputs. It shows modest effectiveness in learning from the input data. Moreover, the model accuracy decreases while running the trained model on the testing data to validate its performance. It poorly performed on the validation dataset with the accuracy of 57%. It indicates that it did not perform well in predicting outcomes from the unlabeled data in case of sites suitability for broadband towers. The results suggest that the SVM model is less successful in reaching high accuracy at both training and validation datasets as shown in figure 04.

### **3.2.2 Relative variable importance**

SVM resulted the relative variable importance from all six variable shows that proximity to road is 79%, population 84%, slope 72% relative importance are highly important along with geology 33% and bulk density 23% are showing least relative variable importance as shown in figure 8(e).

### **3.2.3 Sites suitability for broadband towers**

The suitability level for broadband towers is calculated by dividing the data into four divisions with the suitability level which are highly suitable, moderately suitable, marginally It is based on DN values and SVM resulted 57 sites in highly suitable class, 49 sites in moderately suitable class, 47 sites in marginally and 266 sites are in unsuitable class from total 420 sites. Suitable, and unsuitable. The proposed suitable sites for broadband towers resulting from SVM classification are shown in figure 06.

### **3.2.4 Predication quality**

The ROC curve showed that the AUC score is 0.58, indicating that the model randomly classified sites into their respective suitability level class 58% of the time which is more generic and not consider as good classification model for this site suitability problem. The ROC curve is shown in figure 08(b).

## **3.3 XGBoost**

### **3.3.1 Accuracy**

The findings revealed that the accuracy of XGBoost model on the training data is 88% which shows that it performed very well in learning from the input data. It showed a significant capacity to generalize, with an accuracy of 97% on the testing dataset, implying that the model may successfully predict outcomes for fresh data. The results show that the XGBoost model achieves excellent accuracy in both training and testing datasets, showing its potential for usage in a variety of applications where precision is critical. The study illustrates the XGBoost algorithm's utility in predictive modelling tasks, as well as its capacity to attain high accuracy in both training and testing datasets shown in figure 04.

### **3.3.2 Relative variable importance**

The relative variable importance of used parameters that resulted by the XGBoost show that proximity to road, population, slope and landuse are more important variables for broadband sites selection with the relative importance of 74%, 79%, 67% and 78% respectively. Moreover, geology and bulk density are less contributed with the relative variable importance of 28% and 23% respectively as shown in figure 8(f).

### **3.3.2 Sites suitability for broadband towers**

The suitability level for broadband towers is calculated by dividing the data into four divisions with the suitability level which are highly suitable, moderately suitable, marginally It is based

on DN values and SVM resulted 57 sites in highly suitable class, 49 sites in moderately suitable class, 47 sites in marginally and 266 sites are in unsuitable class from total 420 sites. Suitable, and unsuitable. The proposed suitable sites for broadband towers resulting from SVM classification are shown in figure 07.

### **3.3.3 Predication quality**

ROC curve was utilized to evaluate the performance of an XGBoost classification model to check under fitting and overfitting of the model. The ROC curve indicated an AUC of 0.97, which demonstrates that the model was able to accurately classify sites into their respective suitability level class with an accuracy of 97%. This suggests that the XGBoost classification model is a near-perfect tool for predicting site suitability. The suitability for broadband towers resulted from the XGBoost model on the validation data is presented in figure 08(c).

### **3.4 Assessment of predicted accuracy**

To accurately measure the predicted accuracy, it is important to consider the difference between observed values and predicted values. Predicted values are generated through modelling based on training data. It measures how well the model fits the training samples and doesn't necessarily reflect predicted accuracy. To achieve the true predicted accuracy, it's necessary to assess the difference between the predicted values for new data, such as validation data and the observed values.

Random forest has 0.89% of AUC value which shows its classification quality was very good to classify the datasets into their classes and its ROC curve are shown in figure 08(a). Support vector machine resulted in 0.58% which shows randomize results and it struggle to classify datasets into their respective classes. The ROC curve is shown in figure 08(b). XGBoost has 0.97% AUC value, and which shows its performance is excellent while classifying the datasets highly accurately into their classes and its ROC curve are shown in figure 08(c).

In the representation of the assessment of the models, the following terms are used to explain the results:

True positives (TP): The cases in which the classifier predicted the suitable BTS site, and the actual data sample's class also has BTS site yes, formally.

$$TP = Pc \cap Pe \quad \text{equation ----- (1)}$$

True negatives (TN): The cases in which the classifier predicted the unsuitable site and the actual data sample's class was also don't have BTS site, formally

$$TN = Nc \cap Ne i \quad \text{equation ----- (2)}$$

False positives (FP): The cases in which the classifier predicted the suitable site, and the actual data sample's class was unsuitable BTS site, formally

$$FP = Pc \cap Ne i \quad \text{equation ----- (3)}$$

False negatives (FN): The cases in which the classifier predicted the unsuitable site, and the actual data sample's class was suitable site, formally

$$FN = Nc \cap Pe \quad \text{equation ----- (4)}$$

For the evaluation of the results, the following metrics are used:

- **Accuracy:** The metric used to evaluate the performance of a machine learning model, calculated as the ratio of correctly predicted samples to the total number of input samples.

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} = \frac{|TP|}{|Pc|} + \frac{|TN|}{|Nc|} \quad \text{equation ----- (5)}$$

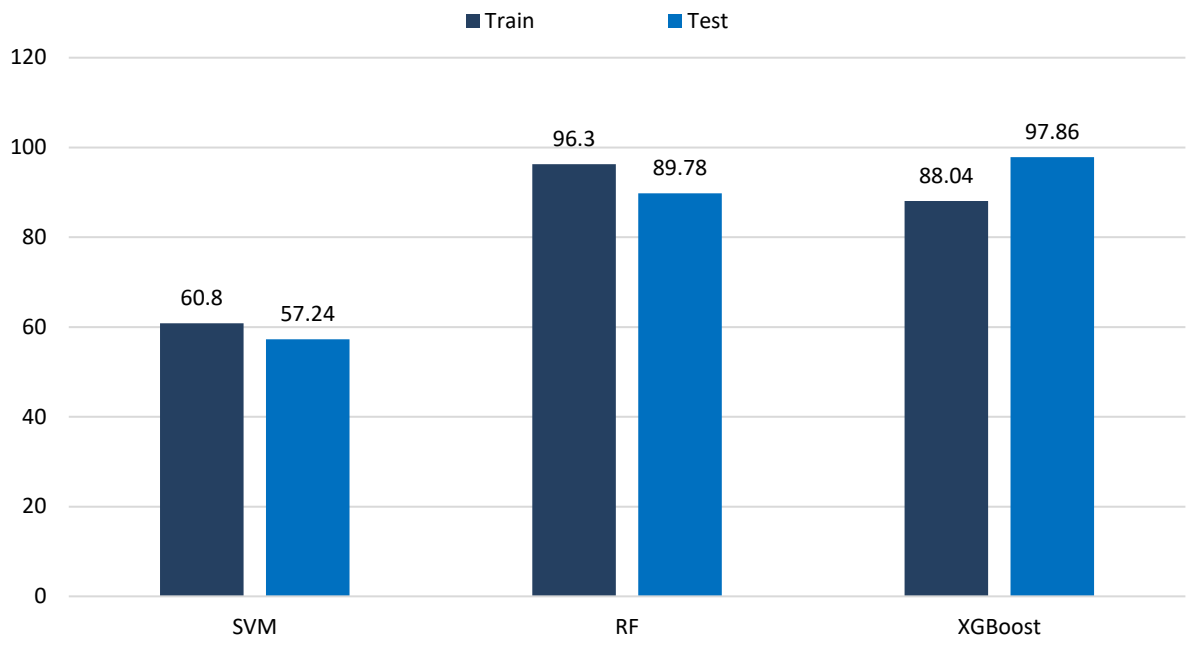


Figure 4. Accuracy comparison of classifiers on training and test data

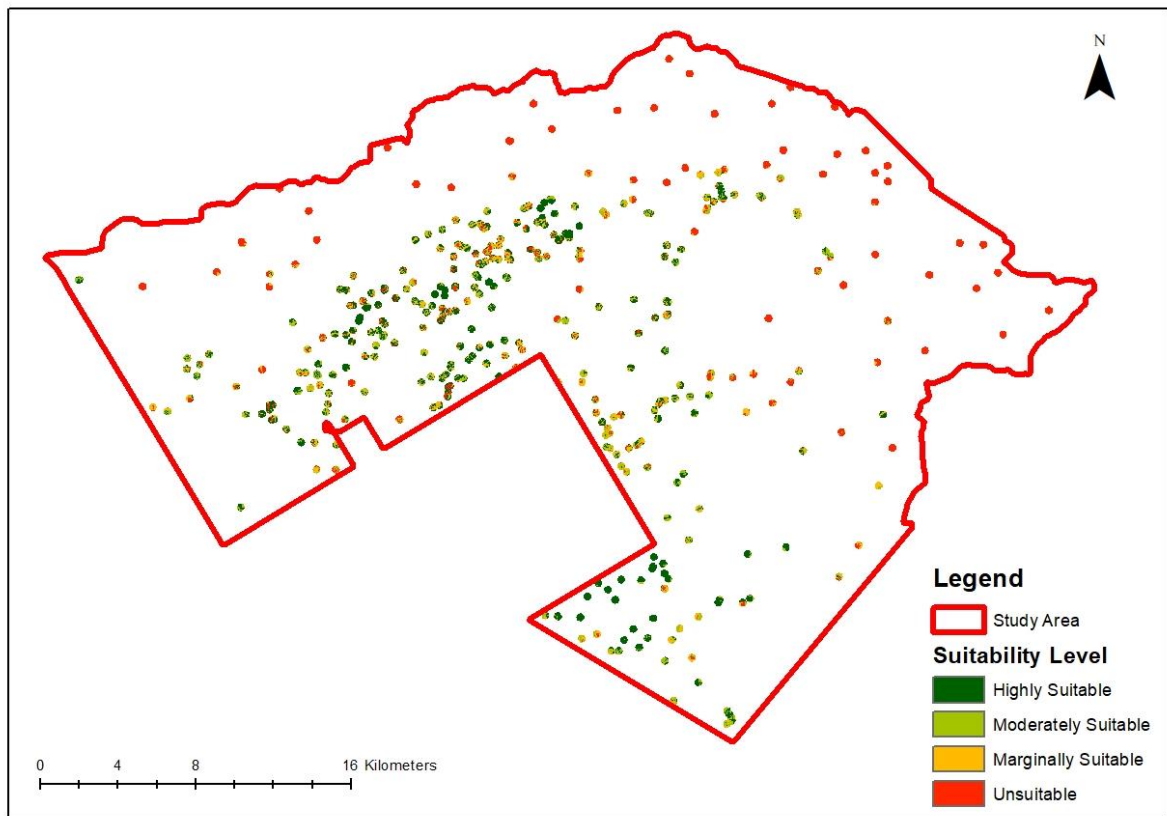


Figure 5. BTS sites suitability using RF classifier.

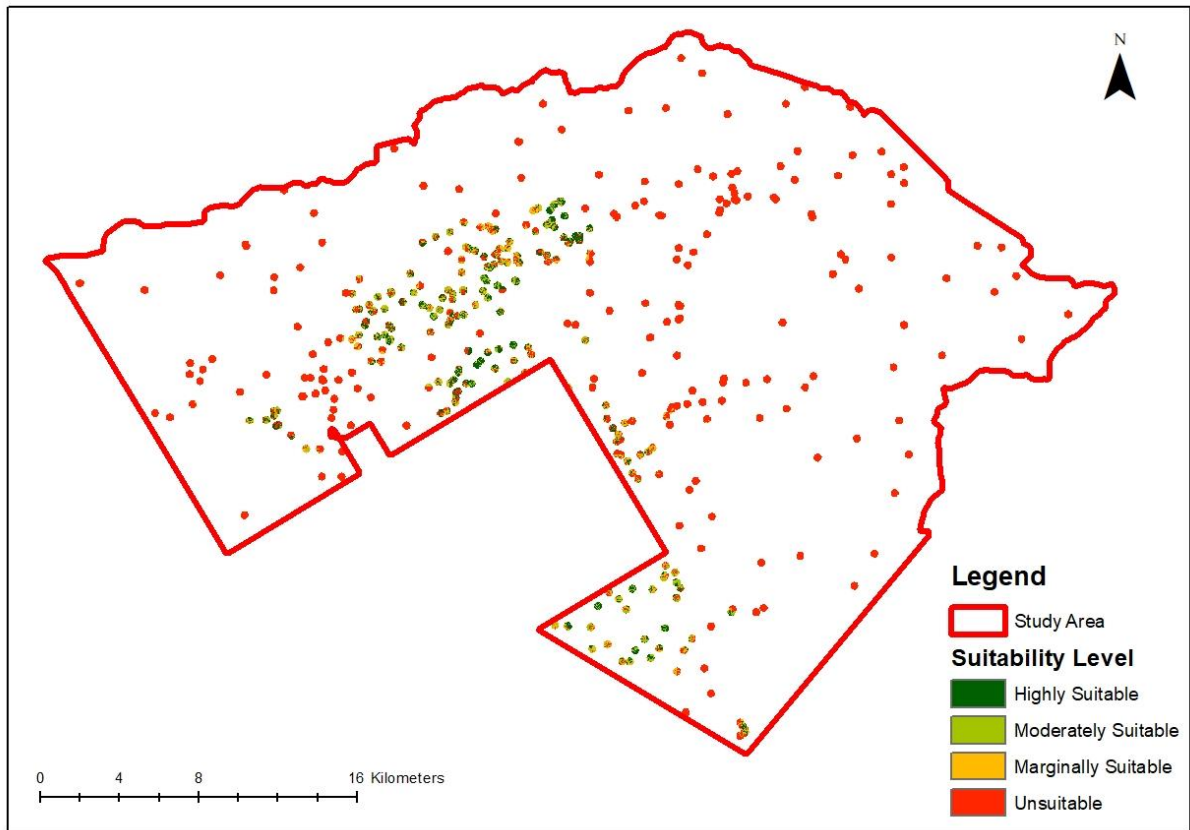


Figure 6. BTS sites suitability using SVM classifier.

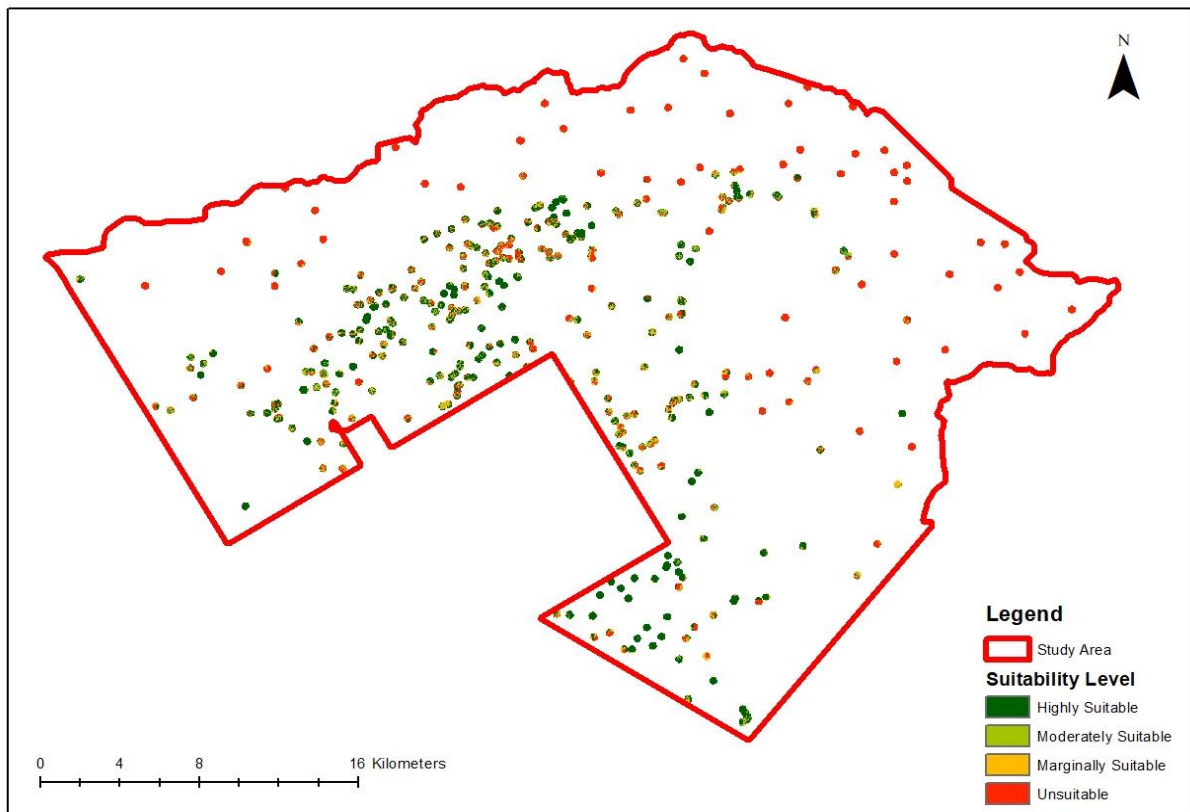


Figure 7. BTS sites suitability using XGBoost.

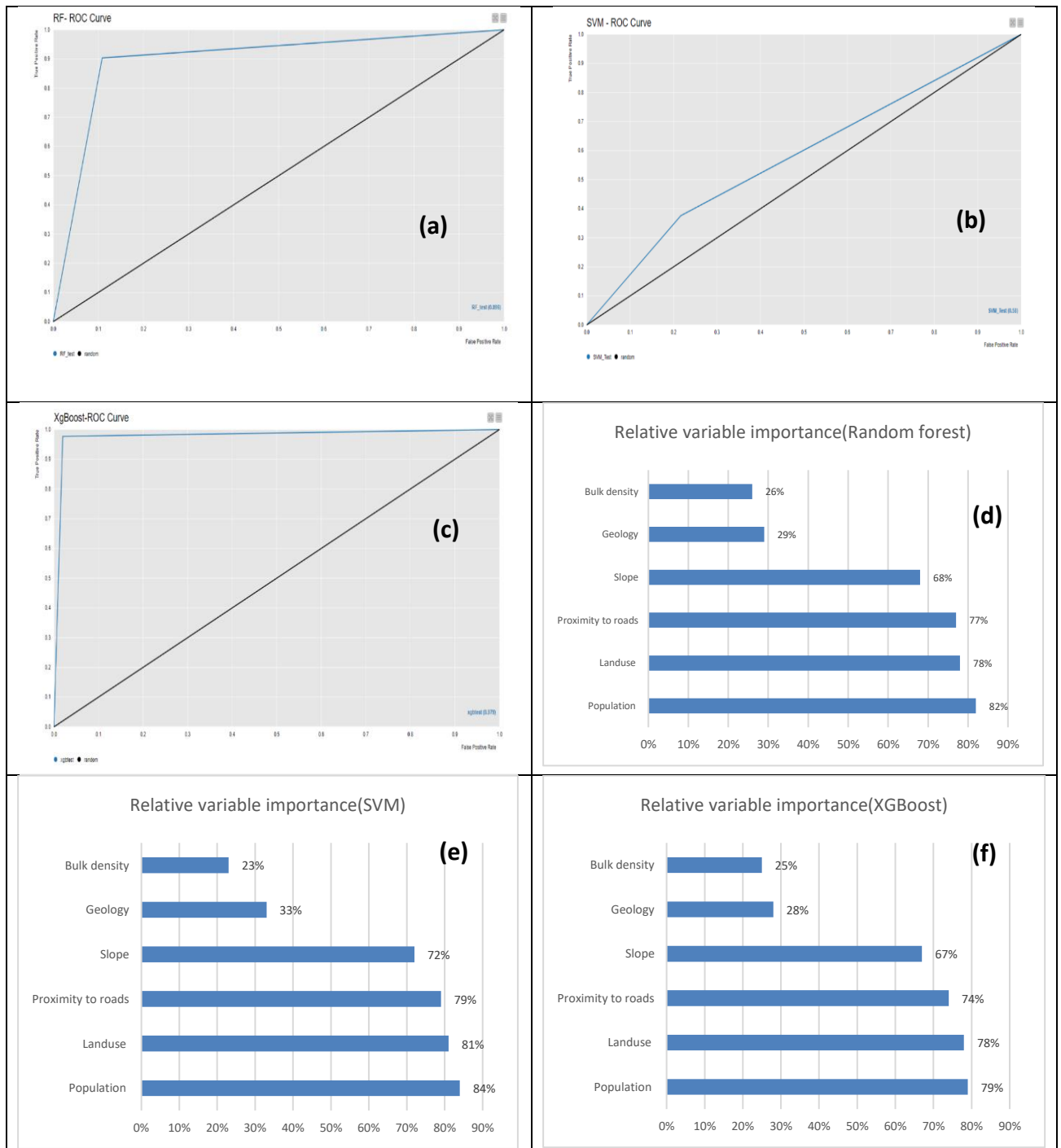


Figure 8. showing the key matrices of all three models.

Table 5 Summary table of all classifiers

<b>Classifier</b>	<b>Accuracy</b>	<b>Precision</b>	<b>Recall</b>	<b>Specificity</b>	<b>f-measure</b>	<b>Cohen's Kappa</b>
<b>SVM</b>	57.24	0.595	0.579	0.579	0.558	15.70%
<b>RF</b>	89.786	0.898	0.898	0.898	0.898	79.50%
<b>XGB</b>	97.86	0.978	0.978	0.978	0.978	95.70%



- **Specificity:** It is a statistical measure that indicates how well a binary classification model can correctly identify negative samples, by measuring the proportion of false positive results with respect to all negative samples.

$$Specificity = \frac{False\ Positive}{False\ Positive + True\ Negative} = \frac{|FP|}{|N_e^i|} \quad \text{equation ----- (6)}$$

- **Precision:** The metric used to measure the accuracy of positive predictions made by a machine learning model, calculated as the ratio of true positive predictions to the total number of samples predicted as positive by the model.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} = \frac{|TP|}{|P_c|} \quad \text{equation ----- (7)}$$

- **Recall:** It also known as sensitivity, is a statistical measure that evaluates the ability of a machine learning model to correctly identify positive samples, calculated as the ratio of true positive predictions to the total number of actual positive samples, regardless of whether they were predicted as positive or negative.

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative} = \frac{|TP|}{|P_e|} \quad \text{equation ----- (8)}$$

- **Cohens Kappa:** The statistic is utilized for evaluating the consistency of ratings between multiple or single raters on qualitative items, known as inter-rater and intra-rater reliability, respectively. It applies to categorical data.

$$Kappa(k) = \frac{P_o - P_e}{1 - P_e} \quad \text{equation ----- (9)}$$

### 3.5 Comparisons of models

For comparing the results models the summary table method and graphs are utilized to show results. The overall summary statistics based on multiple metrics of all used classifiers were presented in table 05. It shows the average metrics applied to all datasets for each classifier after optimization.

Figure 11 shows the accuracy of proposed machine learning models on testing datasets which indicates the SVM doesn't work well and XGBoost is the best performing model and with appropriate parameterization and training, can aid in identifying solutions to site suitability issues by improving classifier accuracy.

Figure 12 showing the precision values of each model which tells the quality of suitable sites for broadband tower predicted by the model while figure 16 sensitivity which measure of how well a machine learning model can detect positive instances along with figure 17, specificity (recall) values measures the proportion of true negatives that are correctly identified by the model and figure 18 representing the Cohen's kappa percentage of each classifier which shows the classification quality of each model.

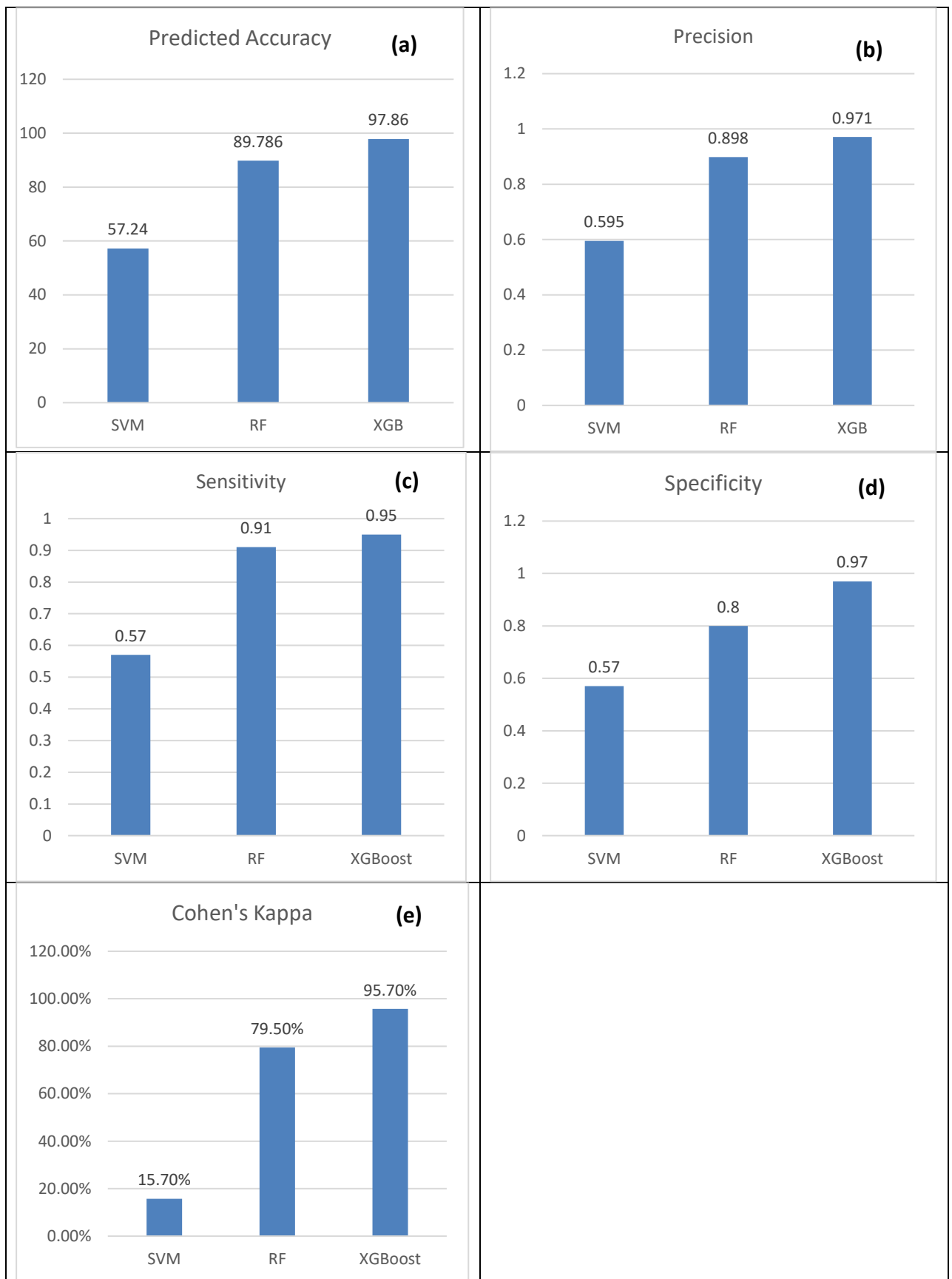


Figure 9. showing the comparison of models.

### CONCLUSION AND RECOMMENDATIONS

#### 4.1 Conclusion

Geospatial machine learning is an incredibly good tool for broadband sites suitability analysis. Among all the three models, XGBoost had the greatest accuracy of 97% on validation data, RF gives 89% accuracy, and SVM gives the lowest accuracy among all three models with 57% accuracy. XGBoost has the greatest model prediction quality with an AUC curve of 0.97%, which is considered as excellent model performance while RF works as second high performing classification model with the AUC curve of 0.89% which indicates that it has good model performance, and SVM doesn't work well for BTS sites suitability, and it gives random prediction results with AUC curve of 0.58%. The kappa value of SVM is 15.7% which indicates slight agreement while RF has 79.5% and XGBoost has 95.7% which is interpreted as almost perfect agreement which indicates that its classification results.

Population, proximity to roads, slope, and land use are identified as the most important exploratory variables, whereas bulk density and geology were recognized as the least relevant ones for BTS sites suitability. These insights can assist telecommunications businesses in selecting the best BTS tower location to increase signal strength, and coverage for users.

These findings can assist telecommunication companies in selecting the optimal BTS tower's location to improve signal strength and coverage for their users. The effectiveness of XGBoost and Random Forest classification models in predicting site suitability and provides valuable insights for telecommunication companies to make informed decisions about BTS tower locations.

## **4.2 Recommendations**

The study emphasizes how essential it is to solve the problem of BTS sites suitability using gridded datasets. This issue might affect the growth of telecom services for people if it is not resolved. The results show that there is need of urgent adoption of geospatial machine learning by telco sector for saving time and resources. We recommend that telecommunication businesses use XGBoost or Random Forest models to identify suitable BTS tower locations to increase signal strength and coverage for their users. These models have proven to be highly accurate and reliable in predicting broadband sites suitability. The important variables identified by the study, including population, proximity to roads, slope, and land use, should also be considered when selecting the best BTS tower location. Additionally, telecommunication companies may benefit from further research to identify other potential variables that could impact site suitability.

## REFERENCES

1. Arrive, T. J., Feng, M., Yan, Y., & Chege, S. M. (2019). The involvement of telecommunication industry in the road to corporate sustainability and corporate social responsibility commitment. *Corporate Social Responsibility and Environmental Management*, 26(1), 152-158.
2. Chen, Y., & Ma, J. (2018). A review of machine learning methods in remote sensing and GIS. *Remote Sensing*, 10(8), 1287. <https://www.mdpi.com/2072-4292/10/8/1287>
3. Jokar Arsanjani, J., & Delavar, M. R. (2015). Machine learning techniques in geospatial data mining: A review. *ISPRS International Journal of Geo-Information*, 4(4), 2311-2336. <https://www.mdpi.com/2220-9964/4/4/2311>
4. Liu, X., & Ma, J. (2019). A review of machine learning techniques for land use and land cover classification. *ISPRS International Journal of Geo-Information*, 8(12), 495. <https://www.mdpi.com/2220-9964/8/12/495>
5. Zhang, L., Wu, C., & Chen, Y. (2020). A review of machine learning techniques for traffic prediction: A geospatial perspective. *ISPRS International Journal of Geo-Information*, 9(8), 448. <https://www.mdpi.com/2220-9964/9/8/448>
6. Ullah, I., Mirza, B., Kashif, A. R., & Abbas, F. (2019). Examination of knowledge management and market orientation, innovation, and organizational performance: Insights from telecom sector of Pakistan. *Knowledge Management & E-Learning: An International Journal*, 11(4), 522-551.
7. Gómez-Barroso, J. L., & Marbán-Flores, R. (2020). Telecommunications and economic development – The 20th century: The building of an evidence base. *Telecommunications Policy*, 44(2), 101904.

8. Jia, C., & Zhang, C. (2020). Joint optimization of maintenance planning and workforce routing for a geographically distributed networked infrastructure. *IISE Transactions*, 52(7), 732-750.
9. Munkhbat, U., & Choi, Y. (2021). GIS-based site suitability analysis for solar power systems in Mongolia. *Applied Sciences*, 11(9), 3748.
10. Al-Ruzouq, R., Shanableh, A., Yilmaz, A. G., Idris, A., Mukherjee, S., Khalil, M. A., & Gibril, M. B. A. (2019). Dam site suitability mapping and analysis using an integrated GIS and machine learning approach. *Water*, 11(9), 1880.
11. Narbaev, S., Abdurahmanov, S., Allanazarov, O., Talgatovna, A., & Aslanov, I. (2021). Modernization of telecommunication networks based on studying demographic processes using GIS. In *E3S Web of Conferences* (Vol. 263, p. 04055). EDP Sciences.
12. De Vries, W. T. (2021). Trends in the adoption of new geospatial technologies for spatial planning and land management in 2021. *Geoplanning: Journal of Geomatics and Planning*, 8(2), 85-98.
13. Rofii, F., Siswanto, D., Hunaini, F., Kafy, A. R., & Indah, J. T. B. Minimising the number of Tower of Base Transceiver Station by Considering the Coverage Area using Fuzzy Clustering Means and Particle Swarm Optimization.
14. Amiri, F. (2021). Optimization of facility location-allocation model for base transceiver station antenna establishment based on genetic algorithm considering network effectiveness criteria (Case Study North of Kermanshah). *Scientia Iranica*.
15. Premarathne, B. M. N., Bamunusinghe, B. K. A. C., & Malavipathirana, C. G. (2021). An Analysis of Suitable Location for Establishing Telecommunication Tower at General Sir John Kotelawala Defence University, Southern Campus.
16. Asassafeh, J., Akkaya, M., & AlTarawneh, M. (2020). A Comprehensive Dynamic Approach for Selecting the Optimal Position of Telecommunication

- Towers. *International Research Journal of Innovations in Engineering and Technology*, 4(11), 26.
17. Casier, K., Verbrugge, S., Meersman, R., Van Ooteghem, J., Colle, D., Pickavet, M., & Demeester, P. (2006). A fair cost allocation scheme for CapEx and OpEx for a network service provider. In *Proceedings of CTTE2006, the 5th Conference on Telecommunication Techno-Economics*.
  18. Asassfeh, J. A., AlTarawneh, M., & Samson, F. (2018). Integrative Model for Quantitative Evaluation of Selection Telecommunication Tower Site. *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, 16(3), 1158-1164.
  19. Avikal, S., Singhal, R., Sajwan, R., Tiwari, R. K., & Singh, R. (2020). Selection of best power supply source for telecom towers in remote areas. *International Journal of Mathematical, Engineering and Management Sciences*, 5(5), 913.
  20. Khilare<sup>1</sup>, S. U., Ghorpade, K. H., & Deshmukh, S. S. (2021). Effective Techniques for Time & Cost Reduction in Cellular Ground Base Tower & Their Supply Chain Management. *Interpretation*, 3242, 2603.
  21. Kweku, J. (2019, March). Analysis of co-location of telecommunication infrastructure in Ghana. In *2019 International Conference on Computing, Computational Modelling and Applications (ICCMA)* (pp. 72-728). IEEE.
  22. Zhang, L., Wu, C., & Chen, Y. (2020). A review of machine learning techniques for traffic prediction: A geospatial perspective. *ISPRS International Journal of Geo-Information*, 9(8), 448.
  23. Chen, Y., & Ma, J. (2018). A review of machine learning methods in remote sensing and GIS. *Remote Sensing*, 10(8), 1287.



24. Jokar Arsanjani, J., & Delavar, M. R. (2015). Machine learning techniques in geospatial data mining: A review. *ISPRS International Journal of Geo-Information*, 4(4), 2311-2336.
25. Liu, X., & Ma, J. (2019). A review of machine learning techniques for land use and land cover classification. *ISPRS International Journal of Geo-Information*, 8(12), 495.
26. Jebamalai, J. M., Marlein, K., Laverge, J., Vandeveld, L., & van den Broek, M. (2019). An automated GIS-based planning and design tool for district heating: Scenarios for a Dutch city. *Energy*, 183, 487-496.
27. Paz, D. H. F. D., Lafayette, K. P. V., & Sobral, M. D. C. (2018). GIS-based planning system for managing the flow of construction and demolition waste in Brazil. *Waste management & research*, 36(6), 541-549.
28. Ullah, I., Aslam, B., Shah, S. H. I. A., Tariq, A., Qin, S., Majeed, M., & Havenith, H. B. (2022). An integrated approach of machine learning, remote sensing, and GIS data for the landslide susceptibility mapping. *Land*, 11(8), 1265.
29. Ighile, E. H., Shirakawa, H., & Tanikawa, H. (2022). Application of GIS and Machine Learning to Predict Flood Areas in Nigeria. *Sustainability*, 14(9), 5039.
30. Effati, M., & Saheli, M. V. (2022). Examining the influence of rural land uses and accessibility-related factors to estimate pedestrian safety: The use of GIS and machine learning techniques. *International journal of transportation science and technology*, 11(1), 144-157.
31. Nguyen, H. D. (2022). GIS-based hybrid machine learning for flood susceptibility prediction in the Nhat Le–Kien Giang watershed, Vietnam. *Earth Science Informatics*, 1-18.

32. Nguyen, K. A., & Chen, W. (2021). DEM-and GIS-Based Analysis of Soil Erosion Depth Using Machine Learning. *ISPRS International Journal of Geo-Information*, 10(07), 452.
33. Adulaimi, A. A. A., Pradhan, B., Chakraborty, S., & Alamri, A. (2021). Traffic Noise Modelling Using Land Use Regression Model Based on Machine Learning, Statistical Regression and GIS. *Energies*, 14(16), 5095.
34. Zuo, R., Wang, J., & Yin, B. (2021). Visualization and interpretation of geochemical exploration data using GIS and machine learning methods. *Applied Geochemistry*, 134, 105111.
35. Motta, M., de Castro Neto, M., & Sarmiento, P. (2021). A mixed approach for urban flood prediction using Machine Learning and GIS. *International journal of disaster risk reduction*, 56, 102154.
36. Avand, M., & Moradi, H. (2021). Using machine learning models, remote sensing, and GIS to investigate the effects of changing climates and land uses on flood probability. *Journal of Hydrology*, 595, 125663.
37. Lee, S., Hyun, Y., Lee, S., & Lee, M. J. (2020). Groundwater potential mapping using remote sensing and GIS-based machine learning techniques. *Remote Sensing*, 12(7), 1200.
38. Fanos, A. M., Pradhan, B., Alamri, A., & Lee, C. W. (2020). Machine learning-based and 3d kinematic models for rockfall hazard assessment using LiDAR data and GIS. *Remote Sensing*, 12(11), 1755.
39. Kopeć, A., Trybała, P., Głębicki, D., Buczyńska, A., Owczarz, K., Bugajska, N., ... & Gattner, A. (2020). Application of remote sensing, gis and machine learning with geographically weighted regression in assessing the impact of hard coal mining on the natural environment. *Sustainability*, 12(22), 9338.

40. Costache, R., Bao Pham, Q., Corodescu-Roșca, E., Cîmpianu, C., Hong, H., Thi Thuy Linh, N., ... & Thai Pham, B. (2020). Using GIS, remote sensing, and machine learning to highlight the correlation between the land-use/land-cover changes and flash-flood potential. *Remote Sensing*, 12(9), 1422.
41. Sun, T., Chen, F., Zhong, L., Liu, W., & Wang, Y. (2019). GIS-based mineral prospectivity mapping using machine learning methods: A case study from Tongling ore district, eastern China. *Ore Geology Reviews*, 109, 26-49.
42. Al-Ruzouq, R., Shanableh, A., Yilmaz, A. G., Idris, A., Mukherjee, S., Khalil, M. A., & Gibril, M. B. A. (2019). Dam site suitability mapping and analysis using an integrated GIS and machine learning approach. *Water*, 11(9), 1880.
43. Costache, R., Hong, H., & Wang, Y. (2019). Identification of torrential valleys using GIS and a novel hybrid integration of artificial intelligence, machine learning and bivariate statistics. *Catena*, 183, 104179.
44. Lai, J. S., & Tsai, F. (2019). Improving GIS-based landslide susceptibility assessments with multi-temporal remote sensing and machine learning. *Sensors*, 19(17), 3717.
45. Bui, Q. T., Nguyen, Q. H., Pham, V. M., Pham, M. H., & Tran, A. T. (2019). Understanding spatial variations of malaria in Vietnam using remotely sensed data integrated into GIS and machine learning classifiers. *Geocarto International*, 34(12), 1300-1314.
46. Chen, W., Peng, J., Hong, H., Shahabi, H., Pradhan, B., Liu, J., ... & Duan, Z. (2018). Landslide susceptibility modelling using GIS-based machine learning techniques for Chongren County, Jiangxi Province, China. *Science of the total environment*, 626, 1121-1135.
47. Thach, N. N., Ngo, D. B. T., Xuan-Canh, P., Hong-Thi, N., Thi, B. H., Nhat-Duc, H., & Dieu, T. B. (2018). Spatial pattern assessment of tropical forest fire danger at Thuan

- Chau area (Vietnam) using GIS-based advanced machine learning algorithms: A comparative study. *Ecological informatics*, 46, 74-85.
48. Mollalo, A., Sadeghian, A., Israel, G. D., Rashidi, P., Sofizadeh, A., & Glass, G. E. (2018). Machine learning approaches in GIS-based ecological modeling of the sand fly *Phlebotomus papatasi*, a vector of zoonotic cutaneous leishmaniasis in Golestan province, Iran. *Acta tropica*, 188, 187-194.
49. Marjanović, M., Bajat, B., Abolmasov, B., & Kovačević, M. (2018). Machine learning and landslide assessment in a GIS environment. *GeoComputational Analysis and Modeling of Regional Systems*, 191-213.
50. Dalela, P. K., Bansal, P., Yadav, A., Majumdar, S., Yadav, A., & Tyagi, V. (2018, July). C4. 5 decision tree machine learning algorithm-based GIS route identification. In *2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN)* (pp. 213-218). IEEE.
51. Bui, D. T., Van Le, H., & Hoang, N. D. (2018). GIS-based spatial prediction of tropical forest fire danger using a new hybrid machine learning method. *Ecological Informatics*, 48, 104-116.
52. Mojaddadi, H., Pradhan, B., Nampak, H., Ahmad, N., & Ghazali, A. H. B. (2017). Ensemble machine-learning-based geospatial approach for flood risk assessment using multi-sensor remote-sensing data and GIS. *Geomatics, Natural Hazards, and Risk*, 8(2), 1080-1102.
53. Munawar, Z., Siswoyo, B., & Herman, N. S. (2017). Machine learning approach for analysis of social media. *ADRI Int. Journal. Information. Technol*, 1(1), 5-8.
54. Pham, B. T., Tien Bui, D., Prakash, I., Nguyen, L. H., & Dholakia, M. B. (2017). A comparative study of sequential minimal optimization-based support vector machines,

vote feature intervals, and logistic regression in landslide susceptibility assessment using GIS. *Environmental earth sciences*, 76, 1-15.

55. Arabameri, A., Pal, S. C., Rezaie, F., Nalivan, O. A., Chowdhuri, I., Saha, A., ... & Moayedi, H. (2021). Modeling groundwater potential using novel GIS-based machine-learning ensemble techniques. *Journal of Hydrology: Regional Studies*, 36, 100848.

# APPENDICES

## Appendix – 1. Code for land use classification in google earth engine.

```
Map.addLayer(Islamabad, {}, "Islamabad");
var Images=ee.ImageCollection("LANDSAT/LC08/C02/T1_TOA").
var s2_composite = Images.filterBounds(Islamabad)
  . filterDate('2021-01-01', '2022-06-30')
  . filter(ee.Filter.lte('CLOUD_COVER', 10))
  . median();
var comp=s2_composite. clip(Islamabad)
Map.addLayer(comp, {Bands:["B1","B2","B3"]}, 'SA satellite imagery');
print(comp).
var bands = ['B2', 'B3', 'B4', 'B5', 'B6', 'B7'].
var samples = Water.merge(Grassland.merge(Builtup.merge(Forest))).
print ('Sample', samples);
var points = comp. select(bands).sampleRegions({
  collection: samples,
  properties: ['Landcover'],
  scale: 30 }). randomColumn();
var training = points. filter(ee.Filter.lt('random', 0.7));
var validation = points. filter(ee.Filter.gte('random', 0.7));
var classifier =ee. Classifier.smileRandomForest(100).train({
  features: training,
  classProperty: 'Landcover',
  inputProperties: bands
}); var Classified_Islamabad = comp. select(bands).classify(classifier);
Map.addLayer(Classified_Islamabad.clip(Islamabad), {min: 0, max:4, palette: [ 'BLUE', 'Green','LightGreen',
'Brown','YELLOW']}, 'Classified Islamabad').
Export.image. toDrive({
  image: Classified_lahore ,
  description: '2022_Classified',
  region: Islamabad});
```

## Appendix – 2. Code of geospatial machine learning

```
# -*- coding: utf-8 -*-
```

```
"""
```

```
Created on Fri Dec 16 11:13:52 2022
```

```
@author: ghulam. hasnain
```

```
"""
```

```
import os.
```

```
os. mkdir("inputs")
```

```
os. mkdir("outputs")
```

```
import geopandas as gpd.
```

```
import shutil.
```

```
import glob.
```

```
# Checking the Sites data
```

```
for f in sorted(glob. glob('data/Sites*')):
```

```
    shutil. copy(f,'inputs/')
```

```
# or grab your data of choice and move to 'inputs/'
```

```
pa = gpd. GeoDataFrame.from_file("inputs/sites.shp")
```

```
pa.sample (5) # GeoDataFrame for the BTS data
```

```
print ("number of duplicates: ", pa.duplicated(subset='geometry', keep='first').sum())
```

```
print ("number of NA's: ", pa['geometry'].isna().sum())
```

```
print ("Coordinate reference system is: {}".format(pa.crs))
```

```
print ("{} observations with {} columns".format(*pa.shape))
```

```
pa [pa.CLASS == 1].plot(marker='*', color='green', markersize=8)
```

```
pa [pa.CLASS == 0].plot(marker='+', color='black', markersize=4)
```

```
# Data sorting
```

```
for f in sorted(glob. glob('data/factors/Factors*.asc')):
```

```
    shutil. copy(f,'inputs/')
```

```
raster_features = sorted (glob.glob(
```

```
    'inputs/Factors*.asc'))
```

```
# check number of features
```

```
print ('\nThere are', len(raster_features), 'raster features.')
```

```
from pyimpute import load_training_vector
```

```
from pyimpute import load_targets
```

```

train_xs, train_y = load_training_vector (pa, raster_features, response_field='CLASS')
target_xs, raster_info = load_targets(raster_features)
train_xs.shape, train_y.shape # check shape, does it match the size above of the observations?
# import machine learning classifiers
from sklearn.ensemble import RandomForestClassifier
from sklearn.svm import SVC
from xgboost import XGBClassifier
CLASS_MAP = {
    'rf': (RandomForestClassifier ()),
    'xgb': (XGBClassifier ()),
    'svm': (SVC ()),
}
from pyimpute import impute
from sklearn import model_selection
# model fitting and spatial range prediction
for name, (model) in CLASS_MAP.items():
    # cross validation for accuracy scores (displayed as a percentage)
    k = 10 # k-fold
    kf = model_selection. KFold(n_splits=k)
    accuracy_scores = model_selection. cross_val_score(model, train_xs, train_y, cv=kf, scoring='accuracy')
    print (name + " %d-fold Cross Validation Accuracy: %0.2f (+/- %0.2f)"
           % (k, accuracy_scores. mean() * 100, accuracy_scores.std() * 200))
    # spatial prediction
    model.fit (train_xs, train_y)
    os. mkdir('outputs/' + name + '-images')
    impute (target_xs, model, raster_info, outdir='outputs/' + name + '-images',
            class_prob=True, certainty=True)
    from pylab import plt
# define spatial plotter
def plotit (x, title, cmap="Blues"):
    plt. imshow(x, cmap=cmap, interpolation='nearest')
    plt. colorbar()

```



```
plt. title(title, fontweight = 'bold')

import rasterio.

distr_rf = rasterio. open("outputs/rf-images/probability_1.0.tif").read(1)
distr_svm = rasterio. open("outputs/svm-images/probability_1.0.tif").read(1)
distr_xgb = rasterio.open("outputs/xgb-images/probability_1.0.tif").read(1)
istr_averaged = (distr_xgb + distr_rf+distr_svm)/3

plotit (distr_averaged, "Tower Sites Suitability, averaged", cmap="Greens")

plotit (distr_averaged [100:150, 100:150], "BTS Tower Sites Suitability", cmap="Greens")
```