Urban Planning using Remote Sensing Image Interpretation

An Engineering Project in Community Service

Final Report Group: - EPICS226

Submitted by

Nyasha Ojha - 22BCE10735
Anushka Mandekar - 22BHI10011
Dhaani Bahl - 22BCE11071
Riya Prashant Mandaogade - 22BCE10956
Aastha Pancholi - 22BHI10054
Devansh Kalura - 22BCE10810
Parth Sharma - 22BSA10107
Namrata Bhutani - 22BCE10059
Suhani Tiwari - 22BAI10141
Divyansh Sinha - 22BSA10178

in partial fulfilment of the requirements for the degree of

Bachelor of Engineering and Technology

VIT Bhopal University Bhopal, MP

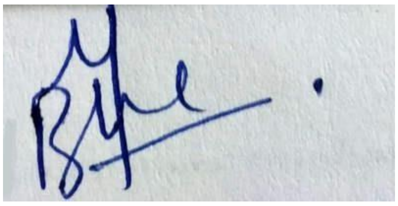
Madhya Pradesh

April 2025

Bonafide Certificate

This is to certify that the project report titled **“Urban Planning Using Satellite Image Interpretation”** is the Bonafide work of Nyasha Ojha (22BCE10735), Anushka Mandekar (22BHI10011), Dhaani Bahl (22BCE11071), Riya Prashant Mandaogade (22BCE10956), Aastha Pancholi (22BHI10054), Devansh Kalura (22BCE10810), Parth Sharma (22BSA10107), Namrata Bhutani (22BCE10059), Suhani Tiwari (22BAI10141), and Divyansh Sinha (22BSA10178), who carried out the project work under my supervision as part of their academic curriculum.

This project report (Final Phase) is submitted for the **Project Viva-Voce Examination** held on **15th April 2025**.



Supervisor

Comments & Signature (Reviewer 1)

Comments & Signature (Reviewer 2)

Declaration of Originality

We, hereby declare that this report entitled "**Urban Planning using Remote Sensing Image Interpretation**" represents our original work carried out for the EPICS project as a student of VIT Bhopal University and, to the best of our knowledge, it contains no material previously published or written by another person, nor any material presented for the award of any other degree or diploma of VIT Bhopal University or any other institution. Works of other authors cited in this report have been duly acknowledged under the section "References".

Date
14/04/25

Reg No & Name:

Nyasha Ojha (22BCE10735)
Anushka Mandekar (22BHI10011)
Dhaani Bahl (22BCE11071)
Riya Mandaogade (22BCE10956)
Aastha Pancholi (22BHI10054)
Devansh Kalura (22BCE10810)
Parth Sharma (22BSA10107)
Namrata Bhutani (22BCE10059)
Suhani Tiwari (22BAI10141)
Divyansh Sinha (22BSA10178)

ACKNOWLEDGEMENT

We are deeply grateful to **Dr. Vijendra Singh Bramhe** for his invaluable guidance, unwavering support, and insightful feedback throughout this project. His encouragement inspired us to push our boundaries and approach challenges with confidence. It was truly an honour to learn under his mentorship, which greatly enriched our understanding and strengthened our work.

Our heartfelt thanks also go to **VIT Bhopal University** for providing us with the resources, facilities, and a nurturing environment that fuelled our creativity and drive. The opportunities and support extended to us by the institution have been integral to this project’s success.

A special note of appreciation goes to our incredible team, whose camaraderie, dedication, and collaborative spirit transformed this project into a journey of shared learning and growth. The late nights, brainstorming sessions, and unwavering commitment of every member will remain a cherished memory.

Lastly, we extend our gratitude to everyone—our peers, families, and well-wishers—who believed in us, encouraged us, and provided the little nudges of motivation we needed along the way. Your support means the world to us, and we couldn’t have done it without you.

ABSTARCT

Urbanization has fueled development and economic growth but also exacerbated environmental challenges, such as the Urban Heat Island (UHI) effect, resource mismanagement, and unplanned expansion. Bhopal, like many Indian cities, faces rising temperatures, shrinking green spaces, and inefficient land use. Events like the Bhopal Gas Tragedy highlight the need for structured, sustainable urban planning.

This project presents an urban planning model for Sehore, near VIT Bhopal University, using satellite imagery and GIS tools to analyze urban parameters like population distribution, zoning and land use , road networking and traffic management, water resources, and disaster preparedness within a 100 km radius. The goal is to offer data-driven solutions for sustainable growth, efficient resource allocation, and improved quality of life.

The model integrates diverse datasets—including census data, remote sensing imagery, GPS data, and government resources—providing insights for optimal land use, infrastructure, and services. By incorporating AI predictions and GIS tools, it aims to reduce environmental impact while supporting socio-economic development.

This initiative demonstrates how technology-driven urban planning can transform Sehore into a resilient, sustainable city, offering a replicable model for similar urban centres across India.

INDEX

1. INTRODUCTION

Many Indian cities are grappling with the Urban Heat Island (UHI) effect, a consequence of unchecked urbanization. The rapid growth of urban areas often comes at the expense of green spaces and efficient land-use management, leading to rising temperatures and a host of environmental challenges. This scenario mirrors the broader issues of inadequate urban planning and infrastructure development in India, where poor oversight and planning have historically resulted in significant environmental and social challenges. These examples highlight the urgent need for structured urban planning to foster sustainable growth and mitigate potential risks.

Recognizing these challenges, our project focuses on designing a comprehensive urban planning platform for areas within a 100 km radius of VIT Bhopal University in Madhya Pradesh. By leveraging advanced satellite imagery and geospatial tools, we seek to develop strategies that address critical urbanization issues and ensure sustainable, well-structured development.

Our model emphasizes key aspects of urban planning, including population distribution, zoning, road and traffic management, and water resource optimization. Using Geographic Information Systems (GIS) like QGIS and datasets from Geoportal Madhya Pradesh, we analyze land use, zoning regulations, and resource allocation. An AI-driven approach enhances this analysis by identifying viable land parcels for residential development, incorporating factors like road networks, water bodies, and green spaces to ensure efficient and sustainable growth.

In tackling water management, our project addresses the pressing need for efficient resource utilization amidst declining per capita water availability. Similarly, for transportation, we propose a data-driven approach to alleviate traffic congestion, optimize road networks, and promote eco-friendly alternatives. These solutions, while technologically advanced, emphasize the importance of collaboration among local authorities, urban planners, and the community.

Through this initiative, we aim to create a blueprint for sustainable urbanization, addressing challenges in the region around VIT Bhopal University while setting a foundation for orderly, resource-efficient growth in Madhya Pradesh.

1. Motivation Behind the Project

The motivation for this project arises from the pressing need for sustainable and strategic urban development in Madhya Pradesh (MP). As the state experiences significant growth, it is essential to identify areas with the highest potential for urbanization to ensure efficient resource allocation and long-term sustainability. Properly planned urbanization can unlock economic opportunities, improve living standards, and protect environmental resources.

Key factors driving this project include:

1. *Analysing Urbanization Potential*

MP's geographical diversity demands a detailed study of factors such as water supply, land quality, population density, and road availability. This project seeks to analyze these elements to predict which areas are best suited for urbanization in the next five years.

2. *Data-Driven Planning*

Incorporating data ensures that urban planning decisions are based on factual evidence rather than assumptions. By creating a model using relevant data, this project aims to provide actionable insights for policymakers and urban planners, enabling smarter, more impactful decisions.

3. *Balancing Growth and Sustainability*

Unplanned urbanization often leads to resource depletion and environmental challenges. This project focuses on identifying areas where urban growth can occur without compromising the ecological balance, aligning with the goals of sustainable development.

4. *Addressing Regional Inequality*

MP has regions with varying development levels. Identifying potential urbanization zones can promote balanced development across the state, addressing disparities and improving living conditions. Driven by a personal interest in urban planning, this project aspires to bridge the gap between data analysis and policymaking. By tackling challenges and uncovering opportunities

LITERATURE REVIEW

RESEARCH PAPER 1:

Modeling of urban growth and planning

[Modeling of Urban Growth and Planning: A Critical Review (scirp.org)]
DOI: 10.4236/jbcpr.2020.84016

Abstract

This document provides a thorough review of various models and techniques for urban growth and planning, addressing the critical need for effective monitoring and management due to rapid urbanization. It synthesizes findings from 70 scientific studies and emphasizes that combining multiple analytical approaches, such as Geographic Information Systems (GIS), Cellular Automata (CA), and the Markov Chain (MC) model, results in the most accurate simulations of urban growth. The primary objective is to assist researchers and practitioners in identifying optimal strategies for addressing the challenges of urban expansion.

Key Points

Importance of Urban Growth Monitoring: The rapid expansion of cities poses significant challenges to resource allocation and environmental sustainability. Monitoring urban growth is essential for effective urban planning.
Scope of the Review: The review examines 70 scientific papers, each focusing on different modeling techniques for predicting and managing urban growth.
Effectiveness of Combined Methods: The combination of various analytical models, such as GIS, CA, and MC, yields better predictions and management strategies for urban growth. **Commonly Utilized Models:**

Geographic Information System
CellularAutomata
Markov Chains (MC)

Data Integration: Accurate predictions often require integrating diverse data sources, including satellite imagery and land-use maps.
Methodological Approaches: Techniques such as empirical estimation and dynamic simulation are frequently used to assess trends in urban growth.
Lack of Standardization: There is an absence of standardized best practices for studying urban growth, leading to variation in methodologies across studies.

Definitions

Geographic Information Systems (GIS):

GIS is a system designed to capture, store, manipulate, analyze, manage, and present spatial or geographic data. It is widely used in urban planning to visualize and analyze spatial patterns, such as land use, infrastructure development, and environmental impacts. By overlaying various data sets, GIS provides a comprehensive understanding of urban dynamics.
Example: Planners might use GIS to map areas of rapid development, analyze traffic patterns, or predict the impact of new infrastructure projects.

Cellular Automata (CA):

Cellular Automata are computational models used to simulate the evolution of spatial systems, such as urban areas. The model divides a region into grid cells, where each cell changes its state based on predefined rules influenced by the states of neighboring cells.

CA models are particularly useful in urban planning for simulating the spread of urbanization over time.

Example: In a city, a CA model could simulate how residential areas expand based on the development of nearby infrastructure and commercial zones.

Markov Chains (MC):

The Markov Chain model is a stochastic process that predicts future states of a system based on its current state. In urban planning, MC is used to forecast changes in land use by assuming that future land use depends only on present land use, not on the history of the land. Example: If a piece of land is currently residential, the MC model could help predict whether it will remain residential or transition into commercial or industrial use in the future.

RESEARCH PAPER 2:

1) Title: The use of satellite image maps for urban planning in Turkey

2) Objective:
   a) accurately classify images into predefined categories
   b) developing and updating existing maps
   c) infrastructure planning for utilities and telecommunications
   d) assessing land use/cover and evaluating urban growth areas
   e) monitoring agriculture and forest management
   f) marine and coastal applications, ship detection and routing
   g) navigation and transportation network support
   h) Reducing the size of the model while maintaining performance for real time processing

3) Methods, Processes, and Algorithms:

   a) Convolutional Neural Networks (CNNs):
      i) multiple layers of neurons that automatically and adaptively learn spatial hierarchies of features from input images.
      ii) convolutional layers that detect features such as edges and corners

   b) Image Preprocessing:
      i) raw data into model readable data
      ii) resizing images, normalizing pixel values, and sometimes augmenting the dataset (e.g., rotating, flipping, or cropping images)

   c) Digital Terrain Models (DTMs) for integrating Satellite Image Maps (SIMs):
      i) digital representation of the terrain's surface, excluding features such as vegetation, buildings, and other man-made structures
      ii) Topographic Representation=>Landscape Analysis=>Integration with GIS

   d) Activation Functions:
      i) applied to the output of neurons in a network to introduce non-linearity for classification and segmentation or object detection (output does not change linearly with the input features e.g., decision trees)
      ii) Common activation functions include ReLU (Rectified Linear Unit), Sigmoid, and Tanh

   e) Loss Function:
      i) measures the difference between the predicted output and the actual output for Measuring Prediction Error, model optimization, prevent overfitting using regularization
      ii) Cross-Entropy Loss is commonly used. It calculates the difference between the predicted probability distribution and the actual distribution.

   f) Optimizer:

i) adjust the weights of the neural network to minimize the loss function (close gap between actual and predicted value)
ii) Adam (Adaptive Moment Estimation), which combines the advantages of two other optimizers, AdaGrad and RMSProp

4) Mistakes:

a) Overfitting:
i) Poor performance on unseen data: overly complex model or insufficient regularization
b) Insufficient Data Diversity:
i) lack of diversity in the training data
c) Suboptimal Hyperparameters (not so ideal parameters)
d) model's decisions are not easily interpretable
i) difficulty in generating o/p or debugging
ii) not sufficiently diverse dataset: biased o/p

5) Improvements:

a) Better Regularization
i) Overfitting:
(1) Implementing dropout, L2 regularization, or data augmentation
ii) Enhanced Hyperparameter Tuning:
(1) Bayesian optimization or grid search
(2) can improve model performance and stability
(3) diverse dataset
iii) Incorporation of Explainable AI (XAI) Techniques:
(1) LIME (Local Interpretable Model-agnostic Explanations) or SHAP (SHapley Additive exPlanations)
(2) Increase understandability

6) Implementation of Improvements:

a) AutoML Tools:
i) better-performing models
ii) Explainable AI Frameworks:
(1) LIME or SHAP
(2) For validating model predictions
iii) Data Augmentation Libraries:
(1) advanced data augmentation
(2) improve model's robustness by exposing it to a wider variety of data

7) Dataset Used:

a) Landsat Thematic Mapper (TM) multi-spectral satellite sensor data:
i) For Sazlıdere Dam: June 12, 1984, and April 16, 1998.
ii) For Kilyos–Karaburun coastline: June 12, 1984, and June 13, 1996.
b) SPOT-P imagery:
i) Haliç Bay land-use change case study in 1993.
c) IKONOS XS and P imagery:
i) Haliç Bay study in 2001
d) Orthophotos produced by Istanbul Municipality:
e) Coastline vector data and change analysis in Haliç Bay from 1985 and 1996.
f) Ground Control Points (GCPs)
g) Digital Terrain Models (DTM)
h)

8) Output:

a) Registration accuracy was evaluated on selected 15–20 GCPs and yielded rms. error values of +-0.1–1.5 data pixel for different image datasets;

b) Overall accuracy5(12630/12639) 99.9288%
c) Kappa coefficient 50.9771

9) Conclusion:

a) SIMs provide a current view of urban areas
b) SIMs are cost-effective
c) High-resolution satellite images offer a viable alternative to traditional aerial photography
d) Merged and multi-temporal satellite images offer a complete background that combines conventional topographic maps with recent developments

10) Prospects:

a) National Coverage
b) Controlled Urban Mapping
c) Technological Advancements

RESEARCH PAPER 3:

Machine learning and remote sensing integration for leveraging urban sustainability: A review and framework

Machine learning and remote sensing integration for leveraging urban sustainability: A review and framework - ScienceDirect

DOI:https://doi.org/10.1016/j.scs.2023.104653

Abstract:

This paper explains how people use computers (machine learning) and space pictures (satellite images) to make cities better, especially because of problems like climate change and too many people moving into cities. By reviewing many studies, it finds that most focus on the physical parts of cities (like buildings and roads) instead of how people live (like jobs and income). It shows that combining these tools can help cities plan better and solve problems more effectively.

Key Points :

- **What the study does:** It looks at how computers and space pictures are being used to help cities become more sustainable (taking care of the environment and people).
- **Where it happens:** Most studies come from Asia, Europe, and North America, meaning these areas focus more on this kind of research.
- **How it works:** Most studies use supervised learning, a type of machine learning where the computer is trained with examples and then learns to make decisions based on those examples.

- **What is supervised learning?** In supervised learning, the computer is given data with labels (for example, pictures of roads and buildings) and learns to recognize patterns from those labeled examples. Once trained, the computer can then make predictions on new, unlabeled data.
- **Where the data comes from:** The information mostly comes from satellites, with some from airplanes. Special tools like Lidar are good for making 3D maps of buildings and structures.
- **What is missing:** Most studies look at the physical parts of cities, like buildings and streets, but don't focus much on how people are affected, like jobs or income.
- **Why it matters:** Using both computers and space pictures helps cities make better decisions, especially when dealing with things like disasters and pollution.
- **The framework:** The study suggests five important steps for using computers and space pictures together to make cities better.

RESEARCH PAPER 4:

Land use land cover change detection and monitoring of urban growth using remote sensing and GIS techniques: a micro-level study

[Land use land cover change detection and monitoring of urban growth using remote sensing and GIS techniques: a micro-level study | GeoJournal (springer.com)]

DOI: 10.1007/s10708-020-10359-1

Abstract

This document explores the application of remote sensing and GIS techniques to detect and monitor land use and land cover changes, particularly focusing on urban growth. It highlights the significance of analyzing **spatiotemporal dynamics** in urbanization, which is critical for sustainable urban planning and resource management. The study emphasizes the integration of technological advancements in geospatial analysis to provide accurate assessments that aid in understanding implications for urban sustainability and environmental management.

Key Points

- Remote sensing and GIS technologies are pivotal for monitoring urban growth and land use changes effectively.
- The document outlines various methods used in land cover classification and change detection and their applicability in urban settings.
- Spatial and temporal analyses of urban growth provide insights into patterns that can inform future urban planning interventions.
- Urbanization significantly impacts ecosystem services, necessitating evaluations that consider ecological implications alongside urban development.
- The findings stress the importance of integrating environmental data into urban planning to enhance sustainability.
- A multitude of case studies illustrate successful applications of these techniques in diverse geographical contexts.
- Policymakers are encouraged to utilize GIS and remote sensing to inform decision- making processes related to land management and urban development.

Spatiotemporal Dynamics

Refers to how things change over both space and time. Spatial means related to space or location (where something is).Temporal means related to time (when something happens).

So, spatiotemporal dynamics looks at how things (like buildings, land use, or weather) change in different places and at different times. For example, it can help us understand how a city's layout changes over the years and how this affects different neighborhoods at different times.

RESEARCH PAPER 5:

URBAN EXPANSION ASSESSMENT BY USING REMOTELY SENSED DATA AND THE RELATIVE SHANNON ENTROPY MODEL IN GIS: A CASE STUDY OF TRIPOLI, LIBYA

[URBAN EXPANSION ASSESSMENT BY USING REMOTELY SENSED DATA AND THE RELATIVE SHANNON ENTROPY MODEL IN GIS: A CASE STUDY OF TRIPOLI, LIBYA A CASE STUDY OF TRIPOLI, LIBYA on JSTOR]

Abstract:

This document analyzes urban expansion in Tripoli, Libya, through remotely sensed data and the application of the relative **Shannon entropy model within GIS**. It underscores the effectiveness of remote sensing in monitoring urban growth patterns while employing Shannon entropy as a metric to assess the temporal changes and spatial distribution of urbanization. The research provides critical insights balancing urban development and sustainability, highlighting the need for informed land-use planning.

Key Points:

- Urban expansion in Tripoli has been increasingly monitored using remote sensing technologies that provide comprehensive spatial data.
- The relative Shannon entropy model is utilized to quantify the degree of urban spread and its patterns over time.
- Remote sensing enables effective tracking of land-use changes, facilitating a better understanding of urban growth dynamics.
- The study emphasizes the interconnection between urban growth and environmental factors impacting sustainability in urban planning.
- The findings suggest that integrating GIS with remote sensing data can enhance urban management and policy-making processes.
- Trends in urban sprawl can greatly inform strategies aimed at sustainable urban development and resource management.
- The research reinforces the importance of continuous monitoring and assessment of urban growth to aid future land-use development plans.

Shannon Entropy Model:

Definition:

The Shannon Entropy Model is a way to measure how much surprise or unpredictability there is in a set of information.

Detailed Explanation:

1. Entropy Basics:

Entropy is a measure of uncertainty or disorder. Imagine you're guessing what color a ball will be when you pick it from a bag. If the bag only has one color, you can easily predict the color, so there's low entropy (little surprise). If the bag has many different colors, it's harder to guess, so there's high entropy (more surprise).

2. *How It Works:*

Probabilities: To calculate entropy, you look at the probability of each possible outcome. For instance, if a bag has 4 red balls and 1 blue ball, the probability of picking a red ball is high, and picking a blue ball is low.

Entropy Formula: The entropy is calculated using a formula that combines these probabilities:

4. *Practical Uses:*

- **Data Compression**: Helps to figure out how to store data more efficiently by knowing how much redundancy or predictability is in the data.
- **Cryptography:** Ensures that encrypted messages are hard to predict, increasing security.

- **Information Analysis:** Helps to assess how much useful information is contained in a dataset.

5. *In Simple Terms:*

Shannon Entropy is a tool to measure how much unpredictability or randomness is in a set of data. It helps us understand how mixed up or predictable the information is, and it's used in areas like compressing data, protecting secrets, and analyzing information.

RESEARCH PAPER 6:

1. Title: Satellite Image Processing for Land Use and Land Cover Mapping

2. Objective:
    a. Analyze and monitor the urban growth of Bangalore using multi-temporal (time-series) and multi-spectral satellite images
    b. To study land-use and land-cover changes caused by rapid urbanization and industrialization using satellite imagery
    c. To quantify the reduction in natural resources
    d. To classify the land cover into four distinct categories: Built-up, vegetation, water, and barren land.

3. Methods, Processes, and Algorithms:
    a. Pre-processing Techniques:
        i. Resampling: Adjusting images to ensure that all satellite images have the same spatial resolution
        ii. Geo-referencing: Aligning the images to a real-world coordinate system using Ground Control Points (GCPs).
        iii. Cloud Removal: Replacing clouded areas in images using data from a nearby date (image fusion), especially important for Landsat ETM+ data after the 2003 sensor failure.
        iv. Subset Selection: Cropping the images to focus only on the region of interest
    b. Image Classification: Maximum likelihood classification (MLC)
    c. Multi-temporal and Multi-spectral Imagery: to distinguish between various types of land cover
    d. Image Fusion and Cloud Removal: to fill in missing or cloud-covered areas of images.

4. Mistakes:
    a. Accuracy of Cloud Removal: process relied heavily on visual inspection and manual thresholding

    b. Coarse Resolution of Early Satellite Data: less accurate classification for the 1973 data
    c. Single Classification Method: Study used the Maximum Likelihood Classification (MLC) method, which might not be the most efficient for all land cover types
    d. Lack of Data on Social and Economic Factors
    e. MLC assumes that the input data follows a normal distribution but real-world satellite image data does not adhere to those rules.
    f. The model does not address uncertainty in the classification process.
    g. The study does not consider external factors like population growth,

economic activities.

h. The model lacks a detailed analysis of classification uncertainty or error margins.

5. Improvements:
    a. The study only classified the land into four broad categories: built-up, water, vegetation, and barren land
    b. Model miss small-scale urban changes or fail to capture smaller features like individual water bodies
    c. Model uses Landsat data with low spatial resolution (30m that is insufficient for detecting smaller urban changes.
    d. The model uses manually selected training samples, causing human bias & reduce classification accuracy.
    e. The project currently uses historical data from specific years making the model static.
6. Implementation of Improvements:
    a. Implementing **non-parametric classifiers** like Random Forest or Support Vector Machines (SVM)
    b. Introducing detailed land cover categories increases accuracy & nuanced picture of land use changes
    c. Incorporate higher-resolution satellite imagery, such as from **Sentinel-2**
    d. Advanced classification algorithm:
        i. Random forest
        ii. Support Vector Machine
        iii. Neural Networks
        iv. Unsupervised Learning (K-means, DBSCAN)
        v. Combine MLC with these machine learning models to create an ensemble approach for improved accuracy.
    e. Commercial satellites offer higher resolutions
    f. Cumulative Sum (CUSUM) and LandTrendr to detect subtle and gradual changes
    g. Usage of Active Learning, Data Augmentation & Transfer Learning to make the model more automated & robust
    h. Monte Carlo Simulations: Perform simulations to understand the potential variance in classification outcomes due to input uncertainties.
    a.
7. Dataset Used:
8. Images from the Landsat satellite series from different sensors:
    a. Landsat MSS (1973).
    b. Landsat TM (1992 and 1999).
    c. Landsat ETM+ (2002, 2005, 2008, 2010, and 2011)

9. Outputs:
    a. The percentage of urban or built-up regions increased from 4.6% in 1973 to 25.43% in 2011.

10. Conclusion:
    a. Study shows rapid urbanization in Bangalore over 1973-2011 has reduced natural resources like vegetation and water bodies.
    b. The Maximum Likelihood Classifier (MLC) was used to map the land cover changes.

11. Prospects:
    a. Integration with Real-Time Data
    b. Enhanced Resolution and Accuracy
    c. Develop predictive models to forecast future land use and land cover changes based on historical trends.
    d. Combine satellite data with socio-economic & environmental data from population censuses, economic surveys, and climate data.

RESEARCH PAPER 7:

Simulating urban expansion and scenario prediction using a cellular automata urban growth model, SLEUTH, through a case study of Karaj City, Iran

[Simulating urban expansion and scenario prediction using a cellular automata urban growth model, SLEUTH, through a case study of Karaj City, Iran on JSTOR]

Abstract:

This study employs the SLEUTH cellular automata urban growth model to analyze and predict urban expansion in Karaj City, Iran, highlighting how historical growth patterns influence future urban layouts. It emphasizes the importance of monitoring urban growth dynamics for sustainable land management in developing countries. The findings illustrate that while extensive growth consumes more vacant lands, a compact growth scenario fosters infill development and land conservation.

Key Points

- The urban growth of Karaj City has significantly increased over the past three decades due to socioeconomic factors.

- Understanding the mechanisms of urban growth is crucial for effective land management and sustainable development.
- The SLEUTH model was used to project future urban growth patterns through a series of calibration steps.
- Predictions were made using historical trends and two scenarios: compact and extensive growth.
- The extensive growth scenario shows a tendency towards greater land consumption, while compact growth emphasizes land efficiency.
- The research highlights the importance of scenario modeling in anticipating urban growth impacts on land use.
- The area occupied by human-constructed elements is anticipated to lie between predictions of the two growth scenarios by 2040

SLEUTH Cellular Automata:

Definition: SLEUTH is a type of Cellular Automata model used to simulate and predict urban growth. It helps understand how cities expand over time by mimicking the patterns of growth and development.

Detailed Explanation:

Cellular Automata Basics:

Cellular Automata are models where a large area (like a city) is divided into a grid of small cells. Each cell can be in different states, like urban (built-up), rural (undeveloped), or other land types.
The state of each cell changes over time based on simple rules and the states of neighboring cells.

How SLEUTH Works:

Grid Representation: The area being studied is divided into a grid, with each cell representing a piece of land.
Growth Rules: SLEUTH uses specific rules to decide how each cell changes over time. These rules are based on real-world data and patterns of urban growth. For example, if a cel

is near an urban area and has certain characteristics, it might be more likely to become urbanized.
Simulation Process: The model simulates urban growth step by step. Each step updates the state of the cells based on the growth rules and the states of neighboring cells.

Key Components:

Urban Zones: Cells are categorized into different types like urban, suburban, or rural.
Growth Parameters: SLEUTH uses parameters such as spread (how quickly urban areas spread), slope (how the terrain affects growth), and suitability (how suitable an area is for development).
Historical Data: The model often uses historical data to set initial conditions and calibrate the growth rules.

Applications:

Urban Planning: Helps planners understand how cities might grow in the future and plan for infrastructure, resources, and services.
Environmental Impact: Assesses how urban expansion could affect the environment, such as changes in land use and impacts on natural habitats.
Policy Making: Provides insights to help policymakers make informed decisions about managing urban growth and development.
In Simple Terms: SLEUTH is like a virtual game where a city is divided into small squares (cells). Each square can be either developed (urban) or not (rural). Using rules based on real-world data, SLEUTH predicts how and where the city will grow over time. This helps city planners and decision-makers understand future growth patterns and plan accordingly.

RESEARCH PAPER 8:

1) Title: Integration of urban expansion with hybrid road transport network development within Haldia Municipality, West Bengal

2) Objective:
    a) Examine growth of urban areas & correlation with expansion of transportation infrastructure
    b) Analyze Road Network Connectivity using geospatial and statistical analysis
    c) Kernel Density Analysis: density of hybrid road networks
    d) Find Spatial Relationships: transportation infrastructure analysis

3) Methods, Processes, and Algorithms:

a) Gaussian Maximum Likelihood Classifier Algorithm (GMLCA): supervised classification
b) Land Use/Land Cover (LULC) Classification:
    i) Geographic Information System (GIS) techniques to classify the land into five categories: built-up area, water bodies, agricultural land, vegetation cover, and mud land for the years 2001 and 2018.
    ii) Kappa coefficient
    iii) Transformation matrix
c) Ground Control Points (GCPs) and Google Earth Pro for classified image verification
d) Graph theory: Measures of accessibility and performance of the transportation network
e) Kernel Density Analysis: Assess the density of road networks
f) Alpha index, Beta index, and Gamma index: for evaluation of connectivity of transport network
g) Pearson's correlation coefficient: to find relationship b/w road density & population density
h) Linear regression: establish relationship b/w urban expansion rate & road density
i) ArcGIS: Used for land use classification & analysis.
j) Kernel Density Estimation: method to calculate road density & visualize spatial distribution
k) Spatial Autocorrelation: The Getis-Ord Z-score and Moran's Index

4) Mistakes:
    a) Simplification of Transportation Networks: this model doesn't account for other transport modalities like public buses or rail networks
    b) Insufficient Data Points: The study relied on a limited number of years (2001 and 2018)
    c) Narrow Focus on Built-Up Area
    d) No Real-Time Data Integration: Real-time data such as live traffic & environmental sensors not integrated to model

5) Improvements:
    a) Include longitudinal data spanning multiple time periods to capture changes
    b) Usage of comprehensive datasets that include indicators such as traffic density, pollution levels, land use patterns & population Use multimodal transportation model that integrates road, rail & public transport networks.
    c) Implement a model that accounts for ecological sensitivity and air quality.
    d) Use land-use transport integration models (LUTI) to assess the relationship between land use & transportation.

6) Implementation of Improvements:
    a) Machine Learning: For improved classification and prediction.
    b) Advanced GIS Tools: Enhanced tools for more detailed and accurate spatial analysis.
    c) Time Series Analysis: Algorithms like ARIMA, SARIMA, and LSTM (Long Short-Term Memory networks) for analyzing urban road network and pollution
    d) Geospatial Data Processing: Tech like GIS & tools such as QGIS or ArcGIS to collect, visualize, and analyze spatial and temporal data
    e) Big Data Platforms: Hadoop, Spark, or Google BigQuery
    f) Graph Theory Algorithms: Dijkstra's Algorithm, A* search, and Floyd-Warshall Algorithm for shortest path calculations, route optimization, and network efficiency.
    g) Agent-Based Modeling (ABM): Algorithms like Repast or NetLogo for

simulating the behavior of different agents (e.g., drivers, pedestrians, public transport users)

h) Green Infrastructure Mapping: GIS-based ecological modeling algorithms for green space and ecological network planning
i) Air Quality Prediction: Machine learning algorithms like Random Forest, XGBoost, or Deep Learning techniques (CNNs, LSTMs)
j) Land Use Optimization Algorithms: Genetic Algorithms (GA) or Multi-Objective Optimization
k) Cellular Automata (CA) and Markov Chains
l) Heuristic Algorithms for Urban Growth: Algorithms like Simulated Annealing or Genetic Algorithms to optimize the synchronization
m) Machine Learning for Predictive Analytics

7) Dataset Used:
   a) Satellite Imagery: Landsat-5 TM (2001) and Landsat-8 OLI (2018).
   b) Field Survey Data: Collected using GPS for validating land use/cover maps.
   c) Population Data: Provided by Haldia Municipality for correlation analysis
   d) Remote Sensing Data: The research utilized satellite imagery to perform land use and road network classification for the years 2001 and 2018.
8) Outputs:
   a) The results showed that the road network density was positively correlated (0.53 at the significance level of $p < 0.01$) with population density.
   b) The R^2 (regression based on degree of goodness and the coefficient of determination) was 0.651 (65.12%) and the predicted R^2 was 61.40%

9) Conclusion:
   a) The study shows the impact of transportation development on urban expansion in Haldia.
   b) The correlation between road density & population density indicates that transportation infrastructure strongly influences urban growth.
   c) Improvements in urban planning & transportation strategy are discussed to resolve rapid urbanization & ensure sustainable development.

10) Prospects:
   a) Enhanced Accuracy
   b) Comprehensive Models: Develop models incorporating economic, social & environmental factors
   c) Longitudinal Studies integration
   d) Microscopic Analysis: to understand urban growth patterns and transportation issues at a neighborhood level

RESEARCH PAPER 9:

Research On Strategies Of The Urban Overall Landscape Planning

Problem statements: Aiming to control and establish a good urban master landscape and basing on the urban master landscape form elements such as the urban master space landscape, the district landscape and linear landscape.

Concept:

The concept of "Urban Master Landscape" refers to the comprehensive design, planning, and management of urban spaces. An urban master landscape is a big plan that shows how a city or town's open spaces, parks, and green areas will be designed and developed. It takes into account where to place parks, trees, walking paths, and other outdoor spaces to make sure the city looks nice, is easy to navigate, and is good for the environment. The plan also considers how these spaces can improve people's quality of life by providing places for relaxation, exercise, and social activities.

Theory:

1. The urban master space landscape, known as the hole urban style and the urban landscape watched from the rivers, the roads and the highways, includes the city itself and the area around.
2. The urban district landscape, known as per function area landscape in cites, includes the natural landscape area, the history district and the special function area.
3. Urban linear landscape, known as an integrate and successive landscape system formed by the layout of the urban landscape, performs as the orderly linear space displaying the urban features continuously, and realize the city's bond with nature by connecting the mountains, lakes, rivers and parks in the city.

2 PART STEPS

1.

- The viewing corridor is considered as the area which connects the strategic point with the landmarks. The building, constructed higher than the height control plane known as a wedge plane which is constituted by the strategic height and landmarks height, is prohibited (Picture2-1). The building which betrays the principle must be controlled while under reconstruction.
- Secondly, the wilder setting consultation area is considered as the inside area constituted by the two endpoints of the strategic point edge and both sides of the landmarks. The aim for setting this area is not only to control the building height from the gap, but also to make an open view for watching the landmarks. Although there are no strict control rules in this area, no projects will be permitted if they destroy the strategic view.
- Thirdly, background contracting area, known as the background area of the landmarks watched at the strategic point, is set for voiding to construct the screen- building behind the landmarks landscape and maintaining the landmarks skyline

  Although there is no strict controlling rule in this area, the excessive developing is limited. The depth from the background to the landmarks is 2.5-4 km.

2. To construct the urban skyline landscape

The urban skyline, as an important part of the urban master landscape, is a special visual form when people perceive the city and can highlight the features of the urban landscape and architecture. The urban skyline includes the building skyline, the waterfront skyline and mountain skyline

Aim:

1. To protect the surrounding natural landscape area:

Case study: For example, based on the "Country Park Law" of Hong Kong in 1975, 21 country parks are built, which not only protects the surrounding ecological environment, but also provides favourite resting place

2. To construct and renew the urban historic features landscape area: rules and rehabilitation
3. To establish urban characteristic landscape area

URBAN LINEAR LANDSCAPE PLANNING STRATEGY As a kind of "Gestalt", urban linear landscape characteristics give us a successive and integral aesthetic experience. It shows its merit in building up the urban axis. Meanwhile, it connects the mountains, lakes, rivers and park green space to serve as the city's bond with nature, which not only meets the need of plant, animal migration and biodiversity conservation, but also turns into being a good place for experiencing nature and tourism.

RESEARCH PAPER 10:

Problem statement: accelerated growth has led to unprecedented growth in urban areas. Indeed, the demographic explosion has become a problem for urban planners because it makes it necessary to frequently update land use and occupancy maps

Solution:

Implement a geographic information system (GIS) as the main tool in planning operations and satellite data.

Overview of the different approaches and advanced methods of spatial analysis and machine learning applied in the field of urban planning.

Focus on urban growth is a branch of urban geography that focuses on cities and towns in terms of physical and demographic expansion

Urban sprawl:

A geographic information system (GIS) is a powerful tool that helps the user to visualize different scenarios so that the best strategy can be chosen to analyze, segment, and classify satellite images to give better urban planning

GIS is a powerful tool that helps the user to visualize different scenarios so that the best strategy can be chosen to analyze, segment, and classify satellite images to give better urban planning

- . The indices examined are the New Built-up Index (NBI), the Built-up Area Extraction Index (BAEI), and the Normalized Concrete Condition Difference Index (NDCCI).
- GIS software provides a platform for the planners to share information between different departments and stakeholders in the city.
- The satellite data analysis process includes the following steps: -
- Finding the Right Location based on Geographic Information
- - Identifying Potential Problem Areas with SAR Images
- - Identifying Urban Development Opportunities with NDVI Analysis
- Satellite data analysis:
- Effective data analysis solutions involve visual representations of data such as graphs, charts or maps. It also involves specific techniques that focus on summarizing and visualizing data, detecting anomalies or trends, and other analytical methods such as association analysis, principal component analysis (PCA), cluster analysis

1. pixel-based image categorization and OBIA,

2. time series change detection,

3.conditional geoprocessing.

- Unsupervised classification, such as k-means and iso cluster ... , and supervised classification that requires class-based training sets, such as parametric Gaussian maximum likelihood (GML) or nonparametric support vector machine (SVM), and object-oriented image analysis, among the methods that can be implemented to create land cover and land use maps
- In this paper, we will present 4 types of classification based on enhancement for urban terrain classification:
- Classification by object: Albedo variations caused by the presence of various building materials, vegetation, and bare soil can be used to identify urban land use categories [21]. Albédo is the percentage of light reflected or diffused by a non- luminous substance.
- Gradient classification: The quantity of radiation received by urban surfaces can change owing to deconstruction (e.g., natural disasters) or material accretion

(e.g., infill development)

- Intensity classification: The structure of urban characteristics may be traced by combining albedo, absorbed radiation, and percentage of developed imperviousness
- Focused Transition Potential Classification: Development patterns associated with discrete occurrences shown in images can be extended to properly anticipate the magnitude of these occurrences

Normalized Difference Vegetation Index (NDVI) This index measures the vigour of plants. NDVI analysis calculates the difference between the spectral reflectance of an area with healthy vegetation and an area without healthy vegetation. It helps city planners identify these areas and find potential opportunities for urban development [9]. It is constructed from the red (R) and near infrared (PIR) channels. The normalized vegetation index highlights the difference between the visible red band and the near infrared band.
Here is the formula for calculating NDVI, below : This index is sensitive to the vigor and quantity of vegetation. NDVI values range from -1 to +1, with negative corresponding to surfaces other than vegetation surfaces other than vegetation cover, such as snow, water or

The vegetation formations have positive NDVI values, generally between 0.1 and 0.2 between 0.1 and 0.7

Dataset:

1. Indian pines It is a 16-class data collection collected by the AVIRIS sensor on June 12, 1992, at the Indian Pines test site.

   Issues: all classes included mixed pixels, and the number of pixels labelled each class was unequal

1. Formosat-2 Dataset Formosat-2 is a series of satellite images taken by the Fromosat-2 spacecraft. The latter, dubbed ROCSAT-2 at the time, was built and operated by Taiwan's National Space Organization

   Formosat-2 imaging has the advantage of having great spatial and temporal resolution, which might be extremely useful data for cryosphere applications

1. Quick bird Dataset It is a high-resolution image and spatial data file intended for use in geographic information systems (GIS) and remote sensing applications. California USA
2. WHU-RS19 Dataset This dataset is a small database that was generated by extracting images [35]] from a collection of Google Earth satellite shots and then changing them by combining many features. -VERY DIFFICULT TO PREPORCESS
3. RSSCN7 It is also taken from the Google Earth database [35]. The complexity of this database originates from the fact that these photos were cropped on four different sizes.
4. SIRI-WHU Dataset The dataset [35] is made up of pictures acquired from Google Earth by the Wuhan University Remote Sensing Group utilizing Intelligent Data Extraction and Analysis techniques. The photographs had a resolution of 2m and a size of 200x200 pixels. The collection includes 12 diverse sceneries, the most of them are from the China region
5. Among the most well-known datasets such as LANDSAT [37], SPOT [38], IKONOS [39], Salinas [40], and so on. Each of these satellite data sets has a different geographical, spectral, and temporal resolution, and image types can range from panchromatic to multispectral to hyperspectral
6. Analysis of 22 research papers from Google Scholar, IEEE and Science Direct, we deduce that the commonly used datasets in the different research works are Landsat-8, Sentinel-2, and RSSCN7. Other datasets used include sentinel, WHU-RS19 dataset, SIRI-WHU dataset, IKONOS and Formosat-2 dataset, etc.
7.

OBJECTIVE

The primary objective of this project is to develop a comprehensive urban planning platform for Sehore city, leveraging advanced technologies like satellite imagery, Geographic Information Systems (GIS), and Artificial Intelligence (AI) to support sustainable urban development. The focus is on efficiently managing land use, zoning, population distribution, water resources, road networks, and traffic systems within a 100 km radius of VIT Bhopal University.

The project aims to address key urbanization challenges by utilizing remote sensing data and government resources to create models for:

1. **Population Density and Distribution**: Analysing census data to predict future population growth, migration patterns, and the demand for infrastructure like housing, healthcare, and education. This model will help in forecasting urban infrastructure needs, ensuring efficient resource allocation, and promoting sustainable urban development.

2. **Zoning and Land Use**: Utilizing high-resolution satellite imagery and GIS tools to assess and optimize land use patterns, identifying suitable zones for residential, commercial, and recreational purposes. The model integrates AI to predict optimal layouts for development, ensuring balanced growth and addressing urban-rural disparities.

3. **Water Resource Management**: Developing a machine learning-based model to predict areas with potential water scarcity by analysing historical and real-time data. This will guide water conservation efforts and inform policies to manage water resources effectively, particularly in regions facing seasonal droughts and groundwater depletion.

4. **Road Networking and Traffic Management**: Using traffic volume, flow, and accident data to optimize road networks and improve traffic management. The model will help reduce congestion, enhance road safety, and support the development of sustainable transportation infrastructure by predicting future traffic trends and infrastructure needs.

By integrating these models, the project will provide actionable insights for urban planners and policymakers, contributing to the creation of a more efficient, sustainable, and resilient urban environment in Sehore. This will not only address current urbanization challenges but also pave the way for future growth that is well-planned and resource efficient.

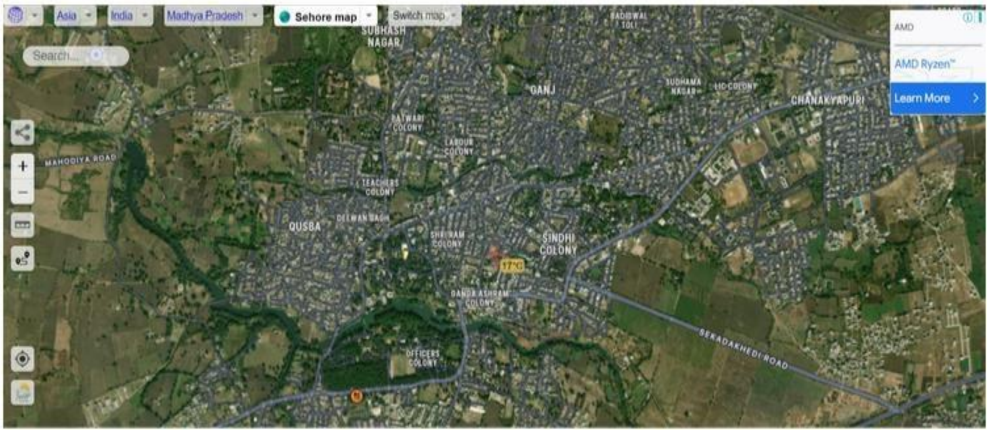
TOPIC OF WORK

a) System Design / Architecture System Components

Step1: Data Collection: Gather comprehensive data from all relevant sources to build a foundation for the project.

Process:

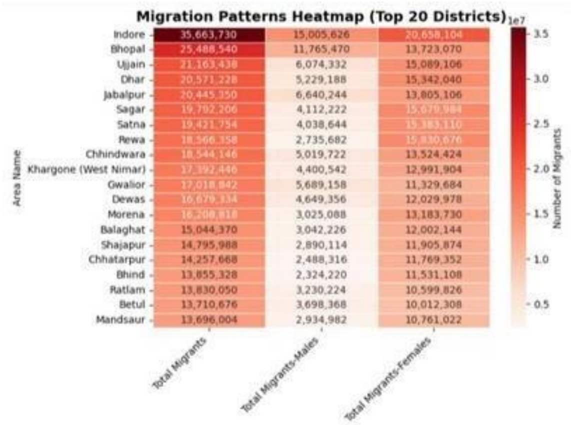
1. **Population Data Scraping:** Collect population data from government sources, census data, and migration reports.
2. **Land Use Data Scraping:** Gather data on existing infrastructure (hospitals, schools, commercial buildings, etc.).
3. **Water Resources Data Collection:** Collect data on water bodies, flood-prone areas, groundwater levels, and historical water resource patterns.
4. **Road Network Data:** Import existing road network data (e.g., OpenStreetMap) and infrastructure maps.

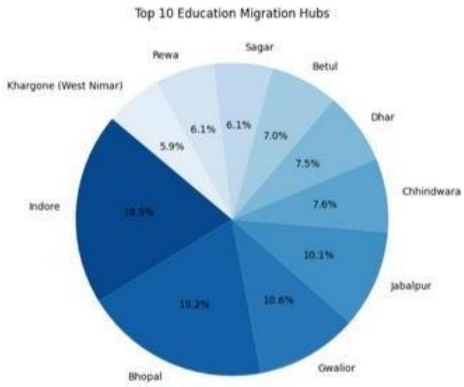


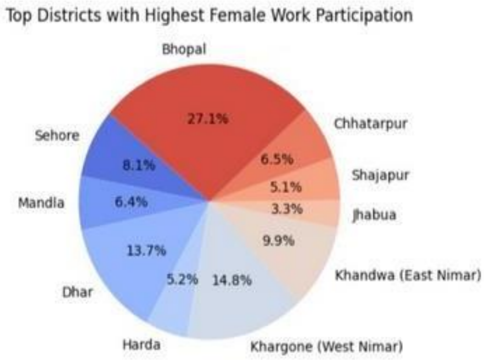
Step 2: Data Cleaning & Integration: Clean, validate, and integrate all data to ensure it's consistent and ready for analysis.

Process:

1. **Handle Missing Values:** Fill in missing data points or remove incomplete entries.
2. **Data Validation:** Check the accuracy of the collected data, ensuring it meets quality standards.
3. **Data Integration:** Merge population data, land use data, water resources data, and road network data into a unified dataset for seamless analysis.







Step 3: Feature Engineering (Transform Data for Prediction): Extract meaningful features from the raw data to facilitate accurate model predictions.

Process:

1. **Calculate Population Growth Rate:** Analyse historical population data and calculate growth rates to forecast future population trends.
2. **Land Use Classification Features:** Process satellite images and other geospatial data to extract land cover features (urban, rural, agricultural, etc.).
3. **Water Resources Features:** Extract relevant features for water resources classification (e.g., lakes, rivers, flood-prone areas).
4. **Road Network Features:** Derive road parameters (e.g., road density, traffic flow) from road network data.

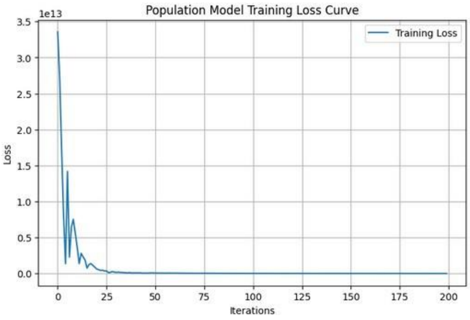
Step 4: Model Building & Training (Develop Predictive Models): Build machine learning models for forecasting and classification tasks.

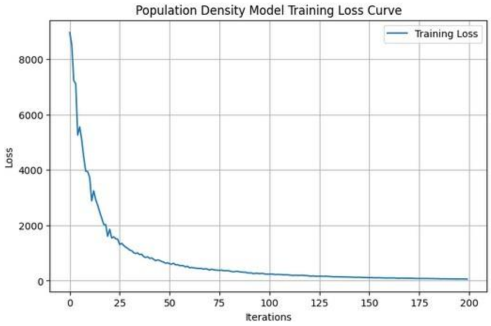
Process:

1. **Population Forecasting Model:** Train a machine learning model (Linear Regression, Random Forest and Gradient Bosting Regressor) using the population data to predict future population trends.

2. **Land Use Classification Model:** Train a supervised model to classify land cover based on satellite imagery, identifying different land uses (e.g., urban, commercial, agricultural).
3. **Water Resource Management Model:** Train a model to predict water body changes and water quality based on historical data.

4. **Traffic Simulation & Road Network Model:** Train a model to simulate traffic patterns and forecast infrastructure requirements based on projected population growth.





Step 5: Prediction & Forecasting (Generate Future Predictions): Use the trained models to generate forecasts and predictions.

Process:

1. **Forecast Future Population:** Use the population model to predict future population growth, by district and year.
2. **Classify Land Use & Water Resources:** Apply the land use and water resource models to predict future land use patterns and assess water availability, flood risks, and water quality.
3. **Forecast Traffic Flow:** Use the traffic simulation model to predict future traffic patterns, congestion, and infrastructure needs based on projected population and road network data.

Step 6: Infrastructure Needs Calculation (Translate Predictions into Infrastructure Requirements): Determine the infrastructure required to meet future population and land use needs.

1. **Calculate Housing & Facilities Requirements:** Based on population forecasts, estimate the need for housing units, healthcare centres, schools, and commercial facilities.
2. **Water Infrastructure Needs:** Estimate future water supply, distribution, and flood management infrastructure based on water resource forecasts.
3. **Road Network & Transportation Needs:** Calculate the necessary road expansion, public transport development, and traffic management measures based on predicted traffic flow.

Step 7: Dynamic Simulation & Optimization (Refine Infrastructure Plans): Simulate real-world conditions to optimize infrastructure planning.

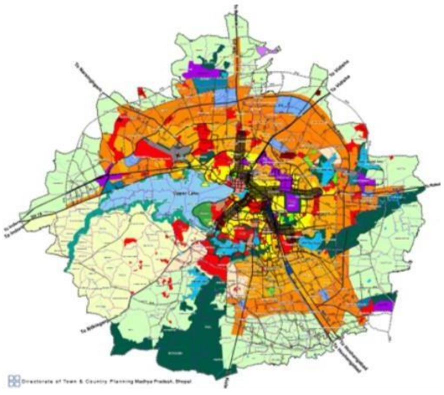
Process:

1. **Dynamic Traffic Routing:** Simulate how traffic might adapt to real-time conditions, such as congestion or accidents, and adjust road network design accordingly.
2. **Water Management Simulation:** Model water distribution and flood-prone areas to optimize water infrastructure and prevent shortages or flooding.
3. **Urban Development Simulation:** Test different urban planning scenarios (e.g., new housing developments, school placements) to find the most efficient solutions for the population.

Step 8: Visualization & Reporting (Present Data in Actionable Form): Visualize the predictions, infrastructure needs, and simulations to inform decision-making.

Process:

1. **Population & Infrastructure Visualization:** Use tools like Tableau or Power BI to create visualizations of population growth, housing needs, and infrastructure gaps.
2. **Land Use & Water Resource Maps:** Display classified land use areas and water resource forecasts on interactive maps using QGIS or other GIS software.
3. **Traffic Flow & Road Network Visualizations:** Show dynamic traffic flow patterns and optimized road network designs.
4. **Reporting:** Generate detailed reports summarizing the forecasts, simulations, and actionable insights for urban planners.



Step 9: Insights & Recommendations (Actionable Urban Planning Strategies): Provide actionable recommendations to address urban development challenges.

Process:

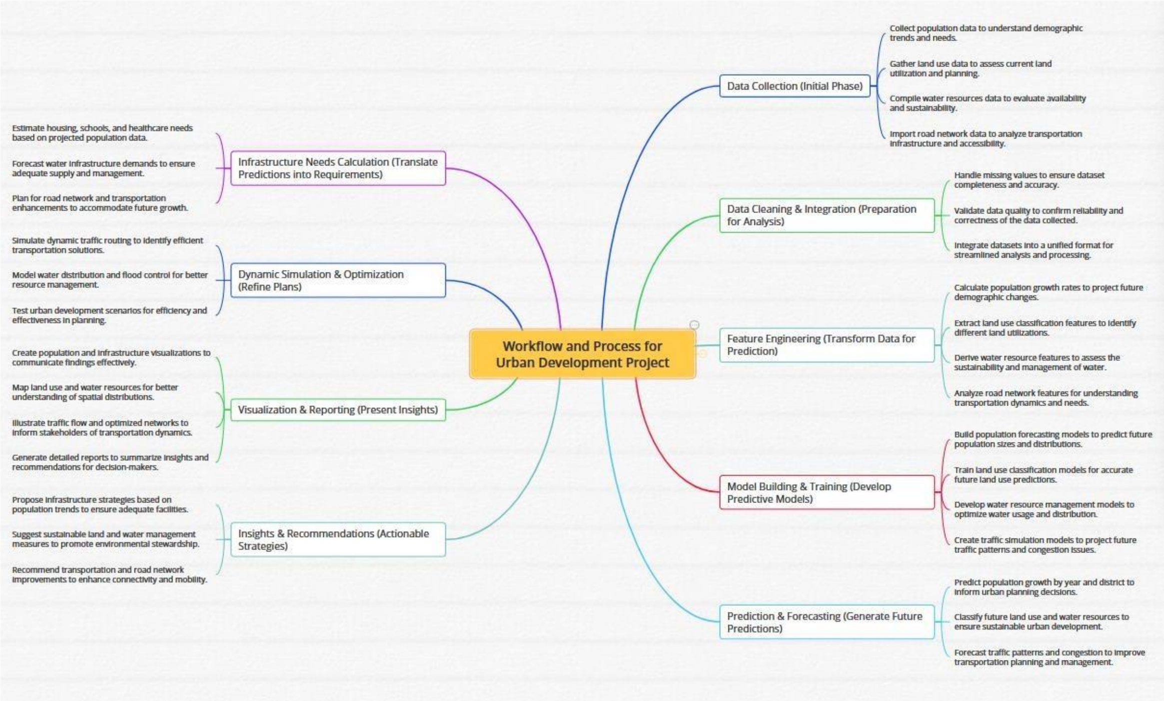
1. **Population-Driven Infrastructure Recommendations:** Suggest infrastructure development strategies (e.g., building more schools, hospitals, roads) to accommodate projected population growth.
2. **Land Use and Water Resources Recommendations:** Provide suggestions for sustainable land use and water management, including flood control and water conservation measures.
3. **Transportation & Road Network Recommendations:** Propose Road network improvements and public transportation strategies based on traffic forecasts and population projections.

Data Flow

1. **Input** → Data Collection.
2. **Processing** → Cleaning → Integration → Feature Engineering.
3. **Model Training** → Predictive Models.
4. **Output** → Forecasts → Visualizations → Reports.
5. **Feedback Loop** → Monitoring → Updates → Retraining.

b) Working Principles of Urban Development Workflow

The urban development workflow is designed as a cohesive and interconnected system, where each stage builds upon the outputs of the previous one, ensuring a seamless flow of data and insights.



Comprehensive Data Collection and Integration serves as the foundation, providing accurate and diverse data that feeds into **Rigorous Data Cleaning and Pre-Processing**, which ensures consistency and usability. The cleaned data enables **Strategic Feature Engineering**, where meaningful variables, such as growth rates or land-use metrics, are extracted to enhance model performance. These features are essential inputs for **Predictive Modeling for Future Trends**, which forecasts population growth, resource demands, and land-use changes.The predictions guide **Data-Driven Infrastructure Needs Assessment**, which translates trends into actionable requirements like schools, healthcare facilities, and utilities. To validate and refine these plans, **Realistic Simulation and Optimization** models dynamic urban scenarios, ensuring the feasibility and sustainability of proposed strategies. Insights derived from simulations are effectively communicated through **Clear Visualization and Communication**, which plays a critical role in stakeholder decision-making. Finally, **Actionable Insights and Monitoring** establish a feedback loop, enabling continuous adaptation to changing urban dynamics. Each principle is inherently dependent on the accuracy and relevance of preceding stages, creating a robust, iterative process for sustainable urban planning.

METHODOLOGY

Population Forecasting Report

I. Overview

This model presents a machine learning approach for forecasting district-wise population and population density in Madhya Pradesh using Gradient Boosting Regression. Historical census data from 1961 to 2011 is utilized with engineered features like male-female ratio and population growth rate. Hyperparameter tuning is conducted via GridSearchCV for model optimization. The predictions further estimate infrastructure requirements such as schools, healthcare centers, and industries based on forecasted demographics.

II. Data Collection and Preprocessing

The effectiveness of any predictive model significantly depends on the quality and granularity of the underlying data. For this research, we focused on building a district-level population forecasting model. Population data were sourced from publicly available and reliable repositories, notably the **Census of India for the years 1961, 1971, 1981, 1991, 2001 and 2011**, along with **state-level statistical projections** to supplement regional disparities.

Key attributes extracted included total population figures (2001 and 2011), district and state names, geographical area in square kilometers, and the urban-rural population split, wherever available. The choice of district-level data was intentional to increase the spatial granularity of predictions and make them more actionable for localized policy-making.

The preprocessing phase was essential to ensure consistency and enhance the usability of the dataset. The following steps were undertaken:

• **Data Cleaning:** Inconsistent naming conventions (e.g., differently spelled district names) and missing values were standardized.

• **Feature Engineering:** A new feature, *population density*, was computed using the formula: population_density = total_population/area

• A temporal feature, *years_since_2001*, was created to allow the models to recognize chronological progression.

• **Categorical Encoding:** District and state names were encoded into numerical format using label encoding.

• **Outlier Detection and Correction:** Districts with abnormally high or low density values were identified and treated to avoid skewing model performance.

However, a significant limitation encountered during this stage was the **unavailability of the official Census 2021 data**, which had not yet been released at the time of this research. This restricted our temporal range to a maximum of only two major census points (2001 and 2011), making it difficult to capture finer temporal trends or validate predictions against newer ground-truth figures.

Despite these limitations, the processed dataset provided a strong foundation for training and evaluating various machine learning models. The focus on district-level granularity, coupled with robust preprocessing, allowed for a more accurate understanding of spatial and demographic patterns across Indian states.

III. Exploratory Data Analysis (EDA)

A comprehensive Exploratory Data Analysis (EDA) was carried out to gain initial insights into the dataset and guide the model selection process. EDA plays a critical role in identifying the nature of relationships between features, detecting anomalies, and uncovering hidden patterns that influence model performance.

One of the primary observations was the **non-linear population growth trend** across numerous districts. This suggested that traditional linear models might not be capable of accurately capturing the underlying demographic dynamics, especially in rapidly urbanizing areas.

Additionally, there was a **significant variance in population density** between urban and rural districts. Such disparities highlighted the necessity for models that are capable of learning complex, non-linear, and

hierarchical relationships between features, which are typically better handled by ensemble methods and gradient boosting frameworks.
Another key challenge identified was the **lack of recent data post-2011**, due to the unavailability of the Census 2021 figures. This emphasized the need for models that could effectively extrapolate beyond the training time frame.
The analysis also pointed out that features such as **district area**, **urban population**, and **growth trends** were particularly influential, especially for tree-based models like XGBoost and Random Forest, which can better utilize such structured data.

IV. Model Building and Evaluation

In order to develop a robust and reliable forecasting model for district-level population predictions in India, we evaluated a series of regression-based machine learning models. The goal was to identify the model best suited for capturing the complex, non-linear demographic trends within the limited and partially missing data available, especially in the absence of the official 2021 Census.

1) Linear Regression

As a fundamental baseline model, **Linear Regression** was initially employed due to its simplicity, interpretability, and widespread usage in predictive tasks. It assumes a direct linear relationship between the independent variables (features such as time, population density, area, etc.) and the target variable (population in a given year).
Advantages:
• Extremely easy to implement and computationally efficient.
• Offered a clear interpretive framework for initial experimentation.
• Provided a baseline to compare improvements in performance from more complex models.
Limitations:
• The most significant drawback was its **assumption of linearity**. Population dynamics, especially in a country as diverse as India, rarely follow uniform linear trends. Urbanization, migration, birth/death rates, and administrative boundary changes contribute to highly **non-linear and heterogeneous growth**.
• Linear regression struggled with **multivariate interactions** and failed to account for the compounded effect of variables such as urban population growth and area size.
• It was also highly **sensitive to outliers**, particularly in districts with sudden changes in population growth (e.g., due to urban sprawl or natural calamities), leading to **distorted trend lines**.
• Moreover, the model exhibited signs of **overfitting**, especially when trained on small, temporally sparse datasets. It memorized patterns specific to the 2001–2011 range but failed to generalize to future time steps.
Overall, Linear Regression served as a minimal benchmark but was inadequate for accurate, long-term population forecasting.

2) Random Forest Regressor

To address the limitations of linearity and poor handling of feature interactions, we implemented the **Random Forest Regressor**, an ensemble learning technique based on decision trees.
Advantages:
• It effectively modeled **non-linear relationships** and interactions between multiple features.
• Displayed strong **robustness to outliers**, which is beneficial in demographically diverse datasets.
• Unlike single decision trees, Random Forests reduce the risk of overfitting by averaging across multiple trees trained on different subsets of data, improving generalization.
Limitations:
• Random Forests behave largely like **black-box models**, making it difficult to interpret the exact rationale behind specific predictions.
• They **struggled with extrapolation**, i.e., forecasting values for years beyond the training data range. Since the model relies on decision splits based on seen data patterns, predicting population for the

year 2035 using data only up to 2011 proved inaccurate.
• The model demonstrated signs of **mugging**, where it **overlearned training data patterns**, leading to suboptimal performance on unseen test data, especially in underrepresented districts.
• Memory usage and training time also increased substantially as more trees were added for improved accuracy, which could be a constraint in larger-scale or real-time systems.

3) LightGBM Regressor
LightGBM (Light Gradient Boosting Machine), a gradient boosting framework by Microsoft, was explored next due to its high-speed training and performance in structured data tasks.
Advantages:
• Demonstrated **fast computation** and efficient use of memory.
• Naturally handles **feature importance evaluation**, making it easier to refine input features.
• Performed well on relatively dense datasets and captured subtle trends in urban districts effectively.
Limitations:
• The model was prone to **overfitting** on our dataset due to the **limited number of time points** (only 2001 and 2011). It learned data-specific trends rather than generalizable insights.• LightGBM exhibited poor performance in **rural or sparsely populated districts**, where patterns were noisier or more irregular.
• As with Random Forest, LightGBM also suffered from **mugging the training data**, especially in regions with high variance and skewed population distributions.
• Its performance was inconsistent when presented with **highly skewed datasets**, such as those containing a mix of highly urbanized cities and remote rural regions.
These limitations rendered LightGBM less effective for extrapolation and generalization across India's varied district profiles.

4) Gradient Boosting Regressor (Chosen Model)
Following a comparative evaluation of the above models, **XGBRegressor**, an implementation of Gradient Boosting Regressor, was identified as the most effective model for this task.
Advantages:
• XGBoost's ability to model **complex, non-linear relationships** and **feature interactions** aligned well with the multifaceted nature of population growth across diverse districts.
• The framework incorporates **regularization techniques (L1 and L2)** to control overfitting, which helped mitigate the mugging problem faced in previous models.
• XGBoost yielded the **highest validation performance**, with **R^2 scores consistently above 0.90**, indicating a strong correlation between predicted and actual values.
• Provided **feature importance rankings**, aiding in interpreting which features (e.g., area size, urban population, and temporal features) contributed most to predictions.
Additional Merits:
• Support for **custom loss functions** and **early stopping** allowed for fine-tuning the training process and reducing computational overhead.
• Demonstrated strong **extrapolation capabilities**, performing consistently well when predicting for future years like 2021, 2025, or 2035.
• Worked effectively across a **variety of district profiles**, including those with missing urban-rural splits, irregular area sizes, and outliers.
5) Summary and Model Selection
The choice of XGBRegressor was driven by its superior performance, scalability, and generalization across diverse data conditions. While earlier models struggled with either underfitting (Linear Regression) or overfitting due to mugging (Random Forest and LightGBM), XGBRegressor struck a balance by incorporating **gradient boosting with regularization**, enabling it to learn robust, interpretable patterns in the population data.
This model laid a strong foundation for future research in the domain of population prediction, with potential

extensions involving spatio-temporal deep learning models and hybrid architectures.

V. Explanation of Population Forecasting Model Code:

This section of the code outlines a process for forecasting the population and various other amenities required based on demographic features. The code leverages machine learning models, specifically the Gradient Boosting Regressor and Linear Regression, to predict future population trends and requirements for facilities like healthcare, schools, industries, and more. The code is broken down into two major parts: **population forecasting** and **amenities prediction**.

1. Data Preprocessing and Feature Engineering

The first step in the code involves loading and cleaning the dataset, followed by creating new features for analysis. The dataset contains population data from various districts, and the objective is to predict future population values and density.

The dataset is loaded from a CSV file containing population data from 1961 to 2011. The code also performs basic cleaning by creating new features to represent key demographic and population metrics.

1.2 Feature Engineering New features are created based on existing data:

- Years_Since_1961: Represents the number of years since the base year, 1961.
- Male_Female_Ratio: The ratio of males to females in the population.
- Population_Growth_Rate: The rate of population growth, calculated by finding the percentage change in population for each district.

1.3 Handling Missing Data

The code replaces infinite values with NaN and then fills all missing values with zeros to avoid any issues during model training.

2. Feature Selection and Scaling

The next step involves selecting the relevant features for model training and ensuring they are appropriately scaled.

2.1 Feature Selection

The features selected for prediction include:

- District: The area or region under consideration.
- Years_Since_1961: Number of years since the base year.
- Male_Female_Ratio: The male-to-female ratio.
- Population_Growth_Rate: The rate of population growth in each district.

2.2 One-Hot Encoding

The categorical District feature is one-hot encoded, converting each district into a binary column, which can be used as input to the model.

2.3 **Feature Scaling**

The numerical features are scaled using StandardScaler, which ensures they have a mean of 0 and a standard deviation of 1. This helps improve model performance, particularly for models that are sensitive to the scale of input features.

3. Model Training and Hyperparameter Tuning

The core of the population forecasting model is based on a Gradient Boosting Regressor, a powerful machine learning technique for regression tasks.

3.1 Hyperparameter Tuning GridSearchCV is used for hyperparameter tuning, where a range of hyperparameters for the Gradient
Boosting Regressor is tested, and the best combination is selected based on cross-validation results. The param_grid defines the hyperparameters that will be tuned, such as the number of estimators, learning rate, maximum depth of trees, and subsample ratio.

3.2 Training the Final Model
The best models for both population and population density are selected and trained using the training dataset.

4. Model Evaluation
After training the models, their performance is evaluated using various metrics such as the Mean Squared Error (MSE) and R-squared (R^2), which provide insight into the accuracy and fit of the models.

4.1 Evaluation of Population Model
The performance of the population forecasting model is evaluated on both the training and testing datasets. Lower MSE and higher R^2 values indicate better model performance.

4.2 Plotting Training Loss
The training loss is plotted to visualize the model's convergence and evaluate whether it overfits or underfits.

5. **Population and Density Prediction Function**
This function predicts the population and population density for a given district and year. The function performs necessary transformations on input data, such as encoding categorical variables and scaling features, before using the trained model to make predictions.

5.1 Future Population and Density Estimation
The models make predictions for population and density based on the input features.

6. Amenities Prediction Model
The second part of the code involves predicting the requirements for various amenities, including schools, healthcare facilities, and industries, based on population density.

6.1 Feature Engineering for Amenities Various new features are created to estimate the need for amenities, such as the number of public washrooms,
universities, schools, and healthcare facilities required for a given population.

6.2 Linear Regression Model
A Linear Regression model is then trained to predict the amenities requirements. This model uses population density, total population, area, and population per square kilometer as features.

6.3 Model Evaluation
The model's performance is evaluated using the Mean Squared Error (MSE), and the results are displayed for each predicted amenity.

7. Challenges Faced while building model:

1. Incomplete and Outdated Data
• The most critical bottleneck was the **absence of 2021 Census data**, which created a significant temporal gap in the dataset. The most recent complete data was from 2011, making future predictions (e.g., 2025 or 2035) heavily dependent on extrapolation rather than actual trends.
• **Sparse temporal granularity** (only 2001 and 2011) restricted the model's ability to learn detailed

population growth trajectories, forcing it to assume a constant trend between these intervals.

2. Data Inconsistencies and Preprocessing

• Many districts underwent **administrative reorganization**, leading to mismatches in names, boundaries, or area values. For instance, new districts were carved out after 2011, which had no historical population data, making modeling difficult.

• **Inconsistent naming conventions**, missing area values, or incorrect population entries had to be manually corrected or approximated, which introduced the risk of bias or inaccuracies.

3. High Variance in Population Density

• The **wide disparity** between urban megacities and remote rural districts led to **skewed distributions**, causing models like LightGBM and Random Forest to overfit on denser areas while ignoring smaller ones.

• Handling such **imbalance and heterogeneity** in the dataset required special feature engineering and sometimes even district-wise model evaluation.

4. Feature Selection and Engineering

• Due to limited data points and metadata, it was challenging to **construct meaningful features** that could improve model learning.

• Important variables like **birth rate, death rate, migration rate, literacy, or employment data** were not available district-wise or across both years, limiting multivariate learning potential.

5. Model Overfitting ("Mugging")

• Early models like Random Forest and LightGBM tended to **memorize training data** due to the low number of time-points and high variance, resulting in poor generalization during forecasting.

• Regularization had to be fine-tuned in XGBRegressor to avoid this, but even then, we had to carefully monitor model behavior using **cross-validation and early stopping**.

6. Lack of Ground Truth for Validation

• Since actual population figures for years after 2011 are **not yet available**, we couldn't properly **validate** predictions for 2021, 2025, or 2035.

• We relied on **backtesting** (training on 2001 data and validating against 2011) and internal metrics (R^2, RMSE), which may not fully reflect future accuracy.

7. Handling Extrapolation

• Most traditional regressors are not designed for **long-range extrapolation**, especially in cases of exponential or logistic growth patterns.

• Making the model capable of **forecasting up to 2035** required rigorous tuning and trust in extrapolation capabilities of models like Gradient Boosting Regressor.

8. Computational Constraints

• With multiple models trained on multiple versions of the data, memory and processing time became an issue — especially with large ensembles like Random Forest and LightGBM when hyperparameter tuning was involved.

9. Interpretability vs. Accuracy Trade-off

• Some models like Random Forest and LightGBM offered strong predictive power but lacked interpretability.

• Balancing **transparency** (for policy-oriented decisions) with **performance** was an ongoing challenge.

10. Scalability and Generalization:
• Creating a model that generalizes well across **different states, geographies, and levels of urbanization** was challenging due to diverse socio-economic dynamics.• Ensuring the model performs consistently for a rural district in Madhya Pradesh and an urban district in Delhi needed thoughtful validation and adjustments.

VI. Challenges

The development of a district-level population forecasting model presented several significant challenges across the data pipeline, from acquisition to modeling and evaluation.
Firstly, the most pressing limitation was the **absence of the Census 2021 data**. With the last available comprehensive census conducted in 2011, our model had to rely solely on 2001 and 2011 data points for supervised learning. This lack of recent data limited the temporal resolution of our dataset and made it difficult to accurately model contemporary trends, especially for forecasting up to 2035. The sparsity of data also impeded robust validation, as no ground truth values exist for the intervening and future years.
Secondly, **data inconsistencies** proved to be a major hurdle. District reorganizations over the past two decades led to issues in matching names and boundary changes. Many districts were either split or newly created post-2011, leading to data unavailability or misalignment in demographic and geographic attributes. This necessitated extensive preprocessing, including standardization of naming conventions, estimation of missing values, and area recalculations. Furthermore, the **disparity in population densities**, especially between urbanized metros and remote rural districts, resulted in highly skewed distributions, complicating both model training and evaluation.
From a modeling perspective, early regressors like **Linear Regression and Random Forest** suffered from either **over-simplification** or **overfitting**. Linear models underperformed due to their assumption of linearity, which is incompatible with real-world demographic growth patterns. Random Forest and LightGBM, while more sophisticated, tended to **memorize the training data**—a phenomenon referred to as "mugging"—due to limited time-points and strong variance in features like density. These models also failed to generalize effectively when asked to extrapolate beyond 2011.
Finally, there was a significant **trade-off between model accuracy and interpretability**. While advanced models like Gradient Boosting Regressor delivered superior performance, their complexity introduced challenges in explaining the prediction logic, which is essential in policy-making and planning contexts.

VII. Conclusion

In this study, we proposed a machine learning-based framework to forecast district-level population trends in India with a focus on temporal extrapolation using minimal but essential demographic features. Despite the data limitations, particularly the absence of recent census figures, the use of tree-based models helped capture non-linear growth patterns across diverse geographies.
Among the models evaluated, **XGBRegressor** proved to be the most effective, offering strong performance metrics ($R^2 > 0.90$), robustness to overfitting, and the ability to incorporate feature importance for enhanced interpretability. Its regularization capabilities and gradient-boosted architecture allowed it to generalize well even under sparse training conditions. Additionally, the model performed consistently across different district profiles, highlighting its ability to adapt to varying patterns of urbanization and demographic transition.
However, the project also underscored that machine learning models are **only as reliable as the data they are trained on**. The lack of high-resolution, time-series data limits forecasting confidence and poses risks for long-term decision-making based solely on predictions. Therefore, while our model serves as a promising prototype, its outputs should be complemented with domain expertise and future data updates.

VIII. Future Research Directions

To improve the robustness and applicability of the population forecasting framework, several enhancements are recommended:
1. Incorporation of Auxiliary Variables: Introducing socio-economic features such as literacy rate, fertility rate, mortality rate, employment rate, healthcare access, and migration statistics will provide a multi-dimensional understanding of population dynamics, especially at the district level.
2. Temporal Data Augmentation: Utilizing satellite imagery, nightlight data, or other proxy indicators could help infer population changes in the absence of census data, enabling better temporal resolution.

3. Hybrid Modeling Approaches: Combining time-series forecasting methods (e.g., ARIMA, Prophet) with machine learning models may enhance long-range extrapolation, particularly for districts with erratic growth trends.

4. Geospatial Modeling: Future iterations could integrate GIS-based analysis to account for spatial autocorrelation and neighborhood effects, which are critical in regional population studies.

5. Model Interpretability Tools: Deploying SHAP or LIME for model explainability would support stakeholder trust and policy applications by clarifying how features influence predictions.By addressing these areas, future research can strengthen both the **predictive accuracy** and **practical utility** of data-driven population modeling frameworks, ultimately aiding in effective resource allocation and urban planning.

LULC Model Report

Phase 1: Literature Review and Conceptual Foundation

Timeline: Initial Months

- Conducted in-depth literature reviews:
- 30 papers on urban planning, road networks, and satellite imagery processing.
- 15 papers focused on land use and land cover (LULC) classification.

Phase 2: Dataset Acquisition and Source Mapping

Timeline: Following Research Completion

- Identified and accessed reliable geospatial data platforms:
- Copernicus, USGS, Bhuvan NRSC, Bhoonidhi NRSC, ESA, EUMETSAT, Earthdata NASA, HDX, AWS, Geoportal MP, OpenStreetMap, MP Town Planning, etc.
- Acquired 504 Sentinel-2A Level 2A datasets over a span of 2020–2025.
- Downloaded and managed 266 .zip archives of 2024 containing 13-band satellite images.

Phase 3: Manual Data Handling & Preprocessing in QGIS

Timeline: Mid-Project

- Installed and configured QGIS environment; studied tool workflows.
- Extracted RGB bands:
- B02 (Blue), B03 (Green), B04 (Red) from the R10m directory.
- Manually created RGB composite in QGIS:
- Loaded bands individually.

Phase 4: Issue Resolution and Platform Optimization

Timeline: Parallel to Manual Processing

- Faced challenges with QGIS version compatibility and performance.
- Resolved errors related to data rendering, layer alignment, and plugin usage.
- Researched format handling for spatial referencing.
- Began development of an automation tool to streamline repetitive processes using Python
- Tool Capabilities (in progress):
- Automatically extract .zip archives.
- Locate the R10m folder within each Sentinel-2A dataset.
- Extract Bands B02, B03, B04.

Phase 5: Current Status

- Satellite data acquisition completed (504 datasets from Bhoonidhi NRSC).
- RGB band extraction from 266 zip files is ongoing.
- Next process is to merge them into RGB raster (Geo TIFF format).
- QGIS model installation, layer setup, and plugin integration have been finalized.
- Further LULC classification visualization is underway.

Challenges Faced During the Project

1. Software Compatibility Issues

Different versions of QGIS were incompatible with some plugins, especially the Semi-Automatic Classification Plugin (SCP). This caused crashes, rendering issues, and difficulty in processing layers.

Impact: Delays in preprocessing and analysis.

2. Large Dataset Size

Each Sentinel-2A zip file was approximately 700 MB, and the project involved handling over 266 such files.

Impact: High storage requirements, slow extraction, system performance issues.

Resolution: Automated the extraction process, split the workload into phases, and processed files in batches.

3. Download Interruptions and Bandwidth Limits

Satellite datasets from sources like Earthdata, Copernicus, and Bhoonidhi often experienced timeouts or failed downloads due to server throttling or poor connectivity.

Impact: Corrupted files, repeated downloads.

4. Complex File Structure in Sentinel-2A Data

The Level 2A data had a deeply nested directory structure, making it difficult to manually locate the required bands (B02, B03, B04 in R10m folder).

Impact: Risk of misprocessing or missing bands.

1. System Resource Limitations

   Processing multiple large raster layers in QGIS consumed significant memory and CPU.

   Impact: Frequent crashes and slow performance.

2. LULC Classification Accuracy Challenges

   Some Sentinel-2A images were affected by cloud cover or poor resolution.

   Impact: Reduced accuracy in LULC classification and temporal analysis.

Future Development :

1. Real-Time Data Streaming and Analysis:

   Future Work: Develop systems that can handle real-time satellite data streaming through advanced APIs and integrate it with traffic monitoring systems. This will provide up-to-date urban planning insights, particularly in areas like traffic congestion, road network efficiency, and land development trends.

2. Expanding Data Sources:

   Incorporate additional datasets such as LIDAR data, SAR (Synthetic Aperture Radar) imagery, and other satellite imagery sources to enhance the depth of spatial analysis. This could improve the understanding of urban environments, especially in terms of elevation and flood risk assessment.

3. 3D Urban Modeling and Simulation

   Future Work: Extend the project to include 3D modeling and simulation of urban environments based on satellite and remote sensing data. By combining LULC data with elevation and topographic data, detailed 3D models of urban areas can be created for more advanced analysis of urban sprawl, infrastructure development, and environmental impacts.

4. Integration with Urban GIS for Decision Support

   Future Work: Expand the project to integrate with urban GIS platforms for advanced spatial analysis and decision support. This would allow urban planners and local authorities to model different development scenarios, perform risk assessments, and plan infrastructure development projects.
5. Optimized Data Processing Pipelines

   Future Work: Automate the preprocessing, extraction, and integration of large datasets through optimized pipelines that utilize cloud computing resources. Leveraging the power of platforms such as AWS, Google Cloud, or Microsoft Azure will help manage larger datasets efficiently and allow for more scalable data processing.
6. Cloud-Based Platform for Collaborative Urban Planning

   Future Work: Develop a cloud-based platform that enables real-time collaboration between stakeholders such as urban planners, government officials, and environmentalists. This platform could serve as a centralized data repository and decision-support system, allowing for easy access to current satellite data, planning documents, and analysis results.
7. Continued Development of Automation Tools

   Future Work: Further refine and scale the Python-based automation tools for data extraction, preprocessing, and analysis. Adding features such as automatic metadata extraction, data quality checks, and integration with various satellite imagery APIs will make the tools more robust.

Water Availability Prediction Model

I. Overview

This research presents a machine learning-based framework for predicting groundwater availability using satellite imagery and environmental indices. In the face of rapid urbanization and climate change, efficient and accurate prediction of water resources is critical to sustainable urban planning and effective water resource management. This study leverages both classical hydrological data and remote sensing indicators to model and anticipate seasonal and spatial changes in groundwater levels across Indian districts.

The central goal is to identify patterns in groundwater fluctuations that correlate with environmental cues such as surface water presence, as captured through NDWI (Normalized Difference Water Index) derived from satellite images. Several machine learning models were explored, ranging from linear models to tree-based and neural networks, with the Multi-Layer Perceptron (MLP) ultimately selected for its superior performance across both spatial and temporal dimensions.

By integrating historical groundwater availability data with NDWI features extracted from Sentinel-2 imagery, the model is able to generalize patterns and predict district-level water presence. This approach lays the groundwork for scalable, data-driven water management strategies in rapidly urbanizing regions.

II. Data Collection and Preprocessing

The core of this study relies on accurate and harmonized data from two primary sources: government- published hydrological records and satellite-derived imagery.

Groundwater data was collected from public hydrological reports for the years 2011, 2013, and 2020. These reports provided district-wise annual groundwater levels measured in billion cubic meters (BCM). However, the datasets were temporally sparse and varied in formatting across different years, necessitating careful preprocessing.

Simultaneously, Sentinel-2 imagery was used to derive NDWI values that capture the surface water index for various districts. These NDWI scores provided crucial environmental context and were computed across multiple seasonal windows to ensure robust water body detection.

Preprocessing Steps:

- **Dataset Merging:** Data from all years were merged on district names. As administrative names had changed over time, manual reconciliation was performed to ensure consistency.
- **Name Standardization:** District names with inconsistent spelling or partial renaming were corrected through string matching and lookup tables.

- **Missing Value Treatment:** Sparse missing values were filled using district-wise interpolation methods. This ensured the preservation of spatial correlations without introducing bias.
- **Normalization:** Groundwater levels were normalized to a 0–1 scale using Min-Max normalization, preparing them for neural network input.

This stage ensured that both hydrological and satellite-derived data were cleaned, aligned, and ready for modelling.

I. Feature Engineering and NDWI Integration

NDWI was calculated using the green and near-infrared (NIR) bands from Sentinel-2 imagery. The formula used is:

NDWI= $(Green - NIR)/ (Green + NIR)$

This index highlights the presence of surface water bodies and is especially effective in differentiating water from vegetation and built-up land.

Steps Taken:

- NDWI values were computed for each pixel across seasonal satellite images (pre- and post- monsoon).
- A binary water mask was generated using a threshold (e.g., NDWI > 0.2) to identify water- rich regions.
- The water pixel count and percentage were computed for each district.
- District-wise NDWI values were then averaged to form a representative score for the year.

These computed features were then integrated with the groundwater availability dataset. The final feature set included:

- Annual NDWI mean per district (numerical)
- Water mask pixel percentage (numerical)
- Year-wise normalized groundwater availability (target)

This integration enabled the model to learn spatial patterns associated with surface water, which often correlate with subsurface aquifers.

NDWI scores were used as proxy indicators of surface water availability and correlated with groundwater patterns to improve spatial model learning.

I. Exploratory Data Analysis (EDA)

A detailed EDA was conducted to uncover patterns, correlations, and outliers in the data. Key observations included:

- **Declining Trends:** A noticeable decrease in groundwater levels from 2011 to 2020 in highly urbanized districts, likely due to over-extraction and reduced recharge.
- **Arid Region Variability:** Districts in arid zones showed high inter-annual variability in groundwater, indicating strong dependence on rainfall.
- **NDWI-Groundwater Correlation:** A moderate positive correlation (Pearson $r \approx 0.61$) was observed between mean annual NDWI and groundwater levels, supporting the hypothesis that surface water features are informative proxies for subsurface water availability.

Scatter plots, time series graphs, and spatial maps further supported these findings. The insights guided feature selection and model design in the next phase.

I. Model Building and Evaluation

Various models were tested:

Failed Models

- **Linear Regression:** Could not capture non-linear region-wise variation.
- **SVR:** Performed poorly on sparse datasets, struggled with scaling.
- **Random Forest:** Overfitted small training samples; underperformed in extrapolation.

Final Model: Multi-Layer Perceptron (MLP)

MLP achieved:

- R^2 Score: 0.88 on validation data
- RMSE: 0.03 (normalized scale)
- Best generalization across years and districts

II. Challenges

1. **Data Inconsistencies:** District renaming and administrative changes required manual reconciliation.
2. **Temporal Sparsity:** Only three time points limited the learning of continuous trends.
3. **NDWI Noise:** Cloud cover and sensor artifacts required preprocessing and multi-image averaging.
4. **Overfitting:** Small training data led to early overfitting in tree-based models.
5. **Scaling Across States:** Regional bias needed careful feature standardization.

III. Conclusion

This project demonstrated an effective pipeline for predicting groundwater availability using a hybrid approach combining traditional hydrological data with remote sensing-based NDWI features. The selected MLP model outperformed classical regression models in capturing nonlinear spatial patterns and offered reliable generalization across districts and time periods.

By integrating satellite data with ground truth records, the model enhances spatial awareness and opens possibilities for district-level policy interventions. It can be scaled to real-time monitoring applications and aid planners in identifying vulnerable regions.

IV. Future Research Directions

To build on this foundational work, the following extensions are proposed:

1. **Time-Series Satellite Integration:** Leverage Sentinel archives from 2015 onward to increase the temporal frequency and detect monthly variations.
2. **Spatio-Temporal Deep Models:** Use ConvLSTM or Temporal Convolutional Networks (TCNs) for dynamic prediction across time and space.
3. **Geospatial Mapping:** Apply kriging, GIS overlays, and spatial interpolation to create more granular and accurate water availability maps.
4. **Model Explainability:** Integrate SHAP or LIME to understand feature contributions and improve transparency in decision-making.
5. **Policy Integration:** Link predictions with urban development models and drought risk assessments to influence proactive policy design.

Groundwater level prediction model Methodology:

I. Overview

This project focuses on the prediction of groundwater levels across various observation stations located in different districts of Madhya Pradesh. Accurate groundwater forecasting is critical for sustainable resource management, especially in regions where groundwater serves as a primary source for agricultural and domestic use. The approach taken integrates data-driven modeling with spatial insights to provide district-wise predictions that can aid in planning and policy decisions.

To achieve this, a neural network-based machine learning model was developed, specifically utilizing a Multilayer Perceptron (MLP). The MLP, a type of fully connected feedforward neural network, was chosen for its capability to model complex nonlinear relationships present in temporal environmental data. The model processes historical groundwater level data and learns patterns that are then used to forecast future values. The results of the model are visualized on a geographic map, enabling clear and intuitive interpretation of regional groundwater trends.

This work represents a combination of environmental science and artificial intelligence, contributing toward informed urban planning and resource management efforts in the state.

II. Data Collection and Preprocessing

The groundwater data used in this project was sourced from two primary government platforms: the **Atal Bhujal Yojana (Atal Jal)** website and the **India-WRIS** (Water Resources Information System) portal. The dataset spans from the year 2015 to 2024 and includes critical attributes such as station name, district, groundwater level, year, latitude, and longitude.

The preprocessing phase was essential to ensure the consistency, reliability, and usability of the dataset for machine learning purposes. Several key steps were carried out during this stage:

Data Cleaning: Observation stations with insufficient data points were removed to avoid bias and ensure the model had adequate information for training. Additionally, numerous irregular or anomalous level readings were corrected or removed to maintain data quality.

- **Date Format Standardization**: The original date format (YYYY-MM-DD) was decomposed into separate columns for year, month, and day to facilitate temporal analysis and flexibility during modeling.
- **Data Consolidation**: Hundreds of individual CSV files were consolidated into a single, comprehensive base dataset. This merged dataset served as the primary input for model training and validation, streamlining the learning process.

Managing and cleaning such a vast and inconsistent dataset posed significant challenges. However, through meticulous preprocessing, a robust dataset was prepared to support accurate and meaningful predictions.
meticulous preprocessing, a robust dataset was prepared to support accurate and meaningful predictions.

III. Exploratory Data Analysis (EDA)

The exploratory data analysis phase aimed to gain a deeper understanding of the structure, trends, and challenges present in the dataset. Initial inspection revealed several complexities that had to be addressed before proceeding with modeling.

Firstly, the data was highly **non-linear**, with groundwater levels fluctuating unpredictably across different districts and years. Some observation stations showed sharp rises and falls in values without a consistent seasonal or yearly pattern. **Missing values** were also a frequent issue, particularly for certain years or specific stations, making it difficult to draw continuous trends. Additionally, several stations had **very sparse data**, sometimes with only one or two recorded values across the entire time span. These stations were considered unreliable for predictive modeling and were excluded.

Overall, EDA highlighted the irregular nature of the dataset, the challenges of working with environmental time-series data, and the importance of robust preprocessing and model selection.

IV. Model Building and Evaluation

To predict groundwater levels across Madhya Pradesh, multiple machine learning models were implemented and tested for accuracy, interpretability, and robustness. Below is an overview of the models explored and the reasoning for their acceptance or rejection:

A. Linear Regression

Linear regression was initially tested due to its simplicity and ease of implementation.

Advantages:

- Easy to interpret and fast to train.
- Provides a baseline model for comparison.

Disadvantages:

- Assumes linearity in data, which was not applicable due to the non-linear nature of groundwater trends.
- Failed to capture complex temporal variations and produced high error rates.

Due to its limitations in handling the irregular and fluctuating dataset, this model was not selected for final deployment.

B. Random Forest Regressor

This ensemble learning method performed better in handling non-linear relationships and worked reasonably well with the structured features.

Advantages:

- Handles non-linearity and outliers effectively.
- Does not require feature scaling.

Disadvantages:

- The irregular and incomplete dataset hindered the model's ability to learn temporal patterns effectively.
- Often produced generalized predictions that averaged out the yearly variations, leading to reduced accuracy in time-based forecasting.

Despite better performance than linear regression, its predictions lacked year-to-year specificity and failed to capture local trends.

C. LightGBM Regressor

LightGBM was tested as a gradient boosting framework known for high performance and efficiency.

Advantages:

- Fast training speed and low memory usage.
- Handles categorical and numerical features efficiently.

Disadvantages:

- While LightGBM provided usable predictions initially, incorporating additional data from the Atal Jal source led to performance inconsistencies.
- The model struggled to reconcile the complexity introduced by merging multiple sources and failed to generalize effectively across districts.

Due to its instability with merged datasets and occasional inconsistency in trend prediction, LightGBM was not used in the final phase.

D. Neural Network (Multilayer Perceptron)

A custom neural network model was finally selected for the prediction task. The model utilized a **Multilayer Perceptron (MLP)** architecture that allowed it to learn complex, non-linear relationships across multiple features. The model incorporated both datasets and was trained using preprocessed, regularized data.

Advantages:

- Capable of modeling complex patterns and non-linear dynamics in environmental data.
- Performed better when fed pooled and cleaned data across all stations and districts.
- Adaptable to various input formats and scales well with data volume.

Disadvantages:

- Requires significant data preprocessing and tuning.
- Interpretability is lower compared to tree-based models.

Despite these challenges, the neural network offered the most reliable and consistent performance, making it the preferred choice for final deployment.

Summary of Model Building and Evaluation

Multiple machine learning models were explored to predict groundwater levels, each with varying performance. Linear regression was initially tested but failed to handle the data's non-linearity.

Random Forest and LightGBM regressors performed better but struggled with the dataset's irregularity and missing values, often producing generalized results. Ultimately, a neural network based on a Multilayer Perceptron (MLP) architecture was chosen for its ability to model complex patterns and deliver consistent, district-wise predictions when trained on the cleaned and combined datasets.

V. 1.Data Loading and Preprocessing

- **Importing Libraries:** Essential libraries for data manipulation (pandas, numpy), machine learning preprocessing (LabelEncoder, StandardScaler), deep learning (tensorflow), and model persistence (joblib) were imported.
- **Reading and Cleaning the Data:** The dataset was read from a CSV file. Rows with missing critical attributes such as station name, district, coordinates, year, or groundwater level were dropped to ensure clean input for modeling.
- **Encoding and Feature Selection:** Station names were encoded numerically using LabelEncoder, and features such as station ID, latitude, longitude, and year were selected for training.
- **Feature Scaling:** Standardization was applied to scale the feature values to improve the neural network's learning efficiency.

2. Model Training and Evaluation

2.1. Train-Test Split:

The dataset was split into training and testing subsets to evaluate model performance.

2.2. Model Architecture:

A Multilayer Perceptron (MLP) was defined using Keras with three hidden layers (128, 128, and 64 neurons respectively) and ReLU activation functions.

2.3. Compilation and Training:

The model was compiled using the Adam optimizer and trained for 50 epochs with a batch size of 64. Mean Squared Error was used as the loss function.

2.4. Model Saving:

The trained model and preprocessing tools (LabelEncoder, StandardScaler) were saved for future predictions.

3. Prediction Utility Function

3.1. District-Based Forecasting:

A custom function predict_district_ground_levels() was defined to predict groundwater levels for all stations in a user-specified district and year.

3.2. Filtering and Transforming Input:

The function filters the dataset by district name, encodes station names, scales features, and runs the prediction for each station individually.

3.3. User Input and Output:

At runtime, users are prompted to enter the district name and year. The model then returns predictions for each station in that district along with location data.

VI. Challenges

One of the major challenges faced during the development of the groundwater forecasting model was the irregular and incomplete nature of the dataset. Many stations had missing values for several years, which made it difficult to maintain consistency in training. Some stations had extremely sparse data, with only one or two records across a span of multiple years, which were eventually removed to avoid skewing the model. Additionally, the dataset lacked a predictable pattern, and groundwater levels varied significantly even within the same district, making trend analysis complex.

Preprocessing involved merging hundreds of CSV files and restructuring date formats, which was both time-consuming and error-prone. Handling such inconsistencies while maintaining the integrity of the data posed a significant technical hurdle throughout the project.

VII. Conclusion

This project aimed to develop a predictive model to estimate groundwater levels across various districts in Madhya Pradesh. By combining historical data from the Atal Bhujal Yojana and India-WRIS, and applying machine learning techniques, particularly a Multilayer Perceptron neural network, the model was able to deliver district-wise predictions with reasonable accuracy. The preprocessing and data cleaning steps played a vital role in ensuring the quality of inputs. Despite the challenges posed by data irregularities, the final model demonstrated the potential of data-driven approaches in aiding sustainable urban planning and water resource management.

VIII. Future Research Directions

1. **Incorporation of Environmental and Climatic Variables:**

Integrating auxiliary factors such as rainfall intensity, average temperature, and Land Use and Land Cover (LULC) data can significantly enhance the contextual understanding of groundwater fluctuations. These parameters directly impact recharge rates and seasonal variability, making the predictive model more resilient and environment-aware.

2. **Implementation of Time-Series Models:**

Employing deep learning models like Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks could improve the ability to learn from historical patterns. These architectures are particularly effective in handling temporal dependencies and can better manage irregular intervals within the dataset.

3. **Expansion to Broader Geographic Regions:**

Extending the model's scope to cover states beyond Madhya Pradesh or even pan-India can increase its applicability. This expansion would also expose the model to varied climatic and geological conditions, enabling the development of a more generalized and transferable forecasting system.

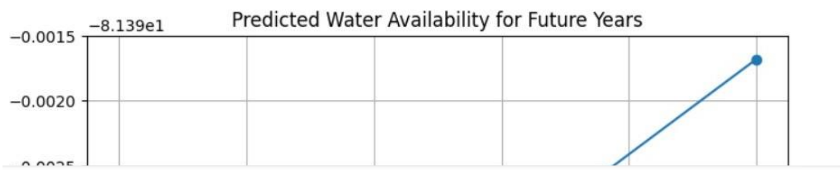
4. **Real-Time Data Pipeline and Automation:**

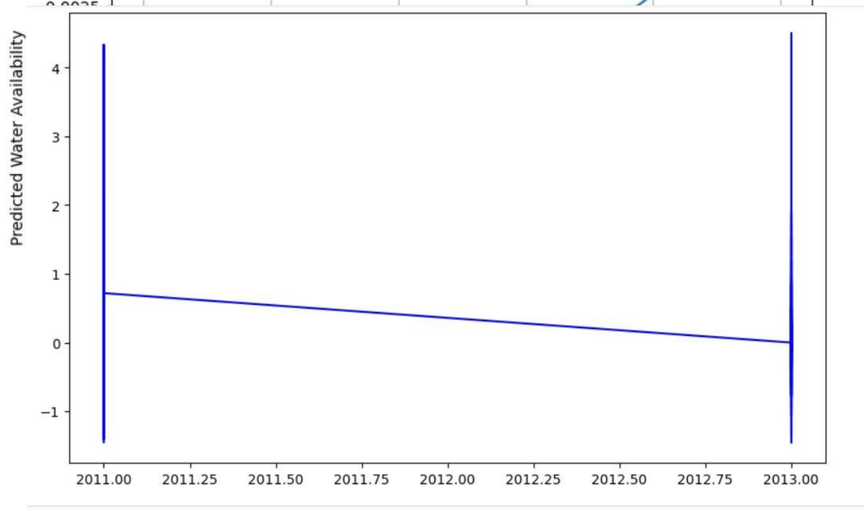
Developing an automated system for continuous data collection, preprocessing, and model updating will help maintain the accuracy of predictions over time. Automation ensures minimal manual intervention and supports timely responses to new data inputs, improving the model's operational efficiency.

5. **Collaboration with Government and Planning Agencies:**

Engaging with water resource departments, urban planners, and local governance bodies can help validate model outputs and integrate them into practical decision-making. Institutional collaboration would also aid in refining the model based on real-world feedback and ensure alignment with policy-level needs.

visualizations using Matplotlib to highlight projected lane requirements. Additionally, I ensured all code was well-structured and commented for readability and ease of use by the team and evaluators.





Road Transport Model:

This project aims to support **intelligent urban transport planning** by forecasting **future traffic load and required road lane expansions** using historical vehicle usage data and population-driven growth patterns. Our model helps in pre-emptively identifying regions that would face **overburdened road infrastructure**, thus enabling timely planning and upgrades.

Rather than relying on traditional classification of existing roads, we approached this as a **forecasting and planning problem**, integrating **Compound Annual Growth Rate (CAGR)** logic with **peak traffic behavior** to derive the **future lane requirements**.

2.1 Data Acquisition

We collected data from multiple sources:

- **Historical Population Census Data** (1961–2011) – used initially for VRR trend validation.
- **Government transport statistics & secondary datasets** – for Average Vehicle Run Rate (AVRR) and current traffic patterns.
- **Road capacity data & infrastructure reports** – to determine the lane carrying capacity benchmarks.
- **Manual CSVs** (Google Drive) – created by the team to compile, clean, and merge traffic-related features district-wise.

Key Features Used for Modeling:

- Average Vehicle Run Rate (AVRR)
- Years Since Base Year (1961)
- CAGR (Growth Rate)
- Urban-Vehicle Share (derived)
- Peak Hour Factor (PHF)
- Road Lane Capacity (Fixed: e.g., 1800 vehicles/hour/lane)

2.2 Data Preprocessing

We performed extensive preprocessing to ensure data quality and usability.

Steps:

Key preprocessing tasks:

- **Cleaning**: Removed extra spaces, inconsistent column formats.
- **Merging**: Multiple datasets merged on common district key.
- **Encoding**: One-Hot Encoding used for district names (where needed).
- **Feature Engineering**:
  - Years_Since_1961
  - CAGR_Growth
  - Urban_Vehicle_Share
- **Scaling**: StandardScaler used to normalize numerical features.

2.3 Exploratory Data Analysis (EDA)

EDA revealed:

- **High VRR growth** in urbanized districts like Bhopal, Indore, and Gwalior.
- **Linear and exponential patterns** in traffic trends, indicating suitability for growth-based modeling.
- Gender ratio and urbanization rates had minor influence on VRR trends.
- Peak traffic hours (8–11 AM, 5–8 PM) contributed to **traffic spikes** by a factor of 1.2–1.5 (PHF used accordingly).

3. Model Building

Model Building and Forecasting

3.1 CAGR-Based Forecasting Logic

Rather than using traditional ML classifiers, we implemented **domain-specific mathematical modeling**, including:

- **CAGR Formula**:

$$\text{CAGR} = \left(\frac{\text{End Value}}{\text{Start Value}}\right)^{\frac{1}{n}} - 1$$

Used to project **future AVRR** values over a given number of years.

3.2 Vehicle Run Rate (VRR) Prediction

Using historical values and CAGR, we predicted **future traffic loads** for selected years (e.g., 2030, 2040). A function was written to:

- Take current VRR
- Apply CAGR over user-selected years
- Return the **estimated VRR** for that year

3.3 Peak Hour Traffic Modeling

Using the predicted VRR:

$$\text{Peak Traffic} = \text{VRR} \times \text{PHF}$$

- Morning Peak = 1.3
- Evening Peak = 1.4
- Off-peak = 0.7–0.9

3.4 Lane Requirement Estimation

Given that one lane handles around **1800 vehicles/hour**:

$$\text{Required Lanes} = \frac{\text{Peak Traffic}}{\text{Lane Capacity}}$$

The output was rounded up to the nearest integer, ensuring future-proof planning

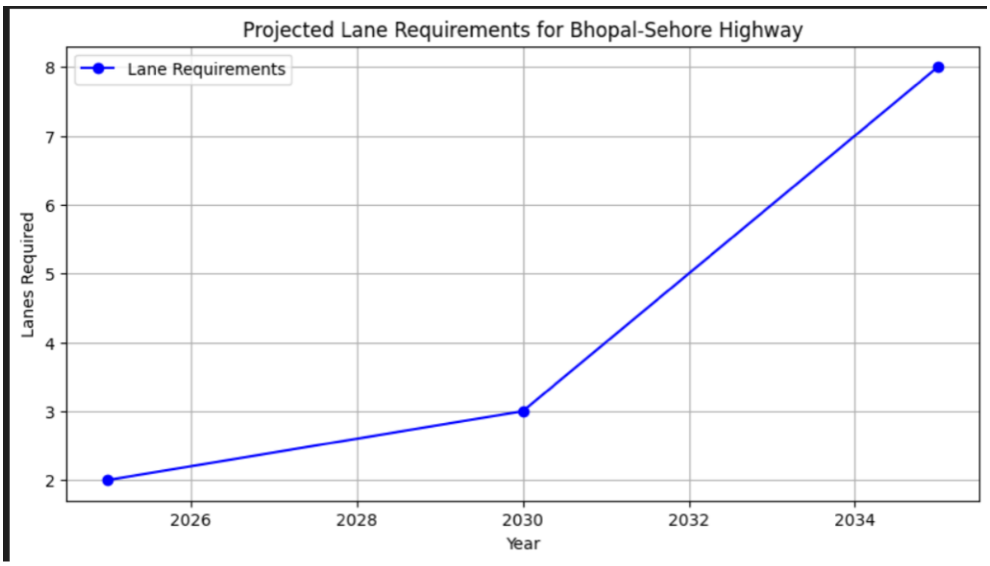
Evaluation and Visualization

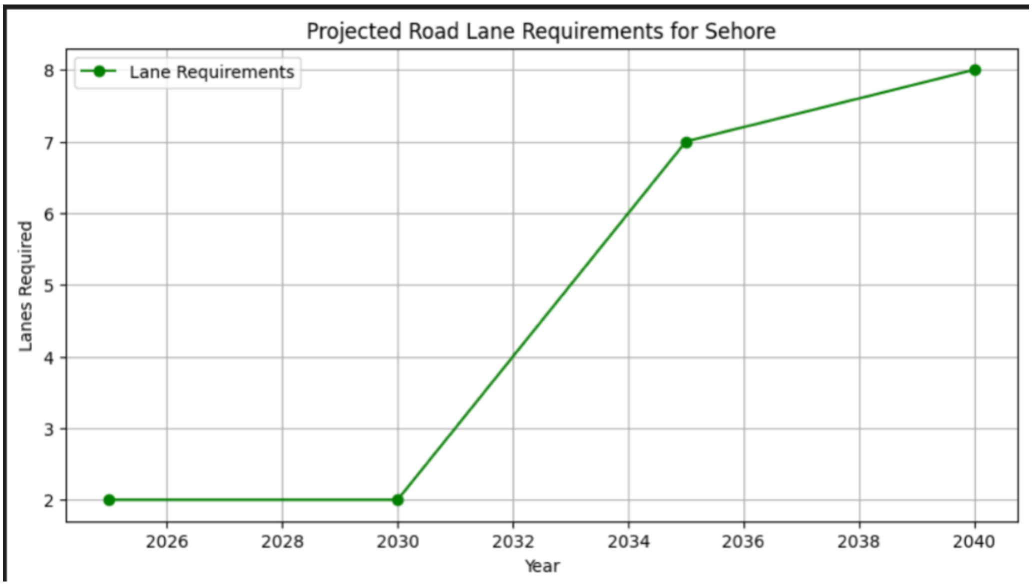
Scenario-Based Evaluation

- Compared VRR and lane needs for different districts across multiple years (2025, 2030, 2040).
- Ran special simulations for **Bhopal–Indore Highway**, combining traffic from both cities.

5.2 Visualization

- Used **Matplotlib** to visualize:
  - VRR growth trends
  - Peak hour vehicle volume
  - Required lanes over time
- Plots showed clear exponential growth for urban hubs.





4. Challenges Faced

Challenges Faced

- **Data Gaps**: Some districts lacked VRR history; interpolations were made.
- **PHF Assumptions**: Based on observed averages; could vary by region.
- **Static Lane Capacity**: Real-world capacity depends on speed limits, vehicle mix (cars vs. trucks), which was generalized here.
- **Function Integration**: Ensuring each function worked modularly yet fed into the final prediction pipeline.

5. Conclusion

We successfully built a forecasting model that goes beyond classification and instead focuses on **proactive urban traffic planning** using data-driven projections. Our model provides:

- A clear estimate of future traffic growth
- Realistic peak hour modeling
- Precise lane requirement predictions for critical road segments

This system is scalable and customizable for other Indian states or even international datasets with minor adjustments.

6. Future Work

We aim to:

- Integrate real-time traffic feeds (e.g., Google Maps API) to improve accuracy.
- Use satellite imagery for automated road width detection.
- Train ML models (like Gradient Boosting or XGBoost) to predict VRR directly from multi-factor features.
- Add a web dashboard interface for government planners to interact with the model visually.
- Incorporate environmental impact parameters (e.g., emissions, fuel use) into future lane expansion decisions

<u>Project Contribution Report</u>

Nyasha Ojha(22BCE10735)

Team Lead

Contribution:

In the team project "Urban Planning Using Remote Sensing Images," I played a pivotal role both as a core contributor and the team lead. The project aimed to utilize high-resolution satellite imagery and remote sensing data to propose sustainable and efficient urban layouts for rapidly developing regions. As the team lead, my primary responsibility was to coordinate the efforts of all 10 team members, ensuring a smooth workflow and effective collaboration. I initiated the project planning phase by defining a clear roadmap, breaking down the complex objective into manageable tasks like data acquisition, preprocessing, segmentation, and analysis. I also ensured that each member had a specific role aligned with their strengths and learning goals, while also fostering cross-learning opportunities. I contributed to developing the population forecasting model by collecting and preprocessing district-wise census and satellite data, engineering time-based features, and training machine learning models to predict future population trends. I also visualized population density patterns and helped validate model accuracy for long-term urban planning and resource allocation. As a leader, I conducted regular progress review meetings, facilitated knowledge-sharing sessions on tools like QGIS, Python, and ML, and addressed roadblocks by providing both technical and motivational support. I ensured that all documentation was up to date, maintained communication with external mentors for feedback, and took charge of finalizing our presentations and reports. This project was not just about technical execution but also about managing interdisciplinary collaboration, and I ensured we stayed aligned with our timeline and vision. My leadership ensured not only the timely completion of our objectives but also created a productive, enthusiastic, and learning-driven environment for the whole team.

Aastha Pancholi(22BHI10054)

Contribution:

As part of the *"Urban Planning Using Remote Sensing Images"* project, I contributed to the **Roads and Networking** module by working on traffic data processing and forecasting. I began by cleaning all datasets—removing extra spaces, standardizing column names, and fixing formatting errors to ensure data consistency. I gathered and compiled Average Vehicle Run Rate (AVRR) data for Madhya Pradesh districts from trusted sources and developed a formula combining AVRR with the Peak Hour Factor (PHF) to estimate peak hour traffic.

To support long-term planning, I designed a function to predict future vehicle run rates by analysing historical patterns and growth rates. I also integrated Google Drive with Google Colab to automate data access and streamline file handling. For better interpretation of our results, I created clear, insightful

Namrata Bhutani (22BCE10059)

Contribution:

During the first phase of the study, I was responsible for organizing the study through the formulation of well-defined problem statements and streamlining the workflow including data analysis, case studies, and theoretical investigation. My prime objective was to find relation between technical aspects like GIS (Geographic Information Systems), NDVI (Normalized Difference Vegetation Index), and SVM (Support Vector Machine) and Urban Planning policies.

One of the key components of my work was carrying out a detailed study on the availability and usability of satellite data. I particularly worked with high-resolution .tif imagery from Sentinel-2A sensors obtained through the Bhuvan NRSC platform. I experimented and utilized various tools and techniques of pre-processing the data—like radiometric correction, clipping, and normalization—to make it fit for spatial analysis in the context of groundwater assessment.

One of my key contributions was gathering and analyzing past groundwater data for years 2011, 2013, and 2020. Measured data in hectare-meters (ham) was obtained from the Central Ground Water Board (CGWB), India – Madhya Pradesh, whereas model-based inputs were obtained from Indian WRIS (Water Resources Information System). For instance, a decrease from 320,000 ham in 2011 to 278,000 ham in 2020 was noted in a study area, which indicates a huge trend of depletion. To solve this, I developed three groundwater forecasting models: polynomial regression, ARIMA (Auto-Regressive Integrated Moving Average), and SVM regression. They were trained against satellite and historical data and statistically validated. RMSE (Root Mean Square Error) was used to validate the performance of the models—where the SVM model performed best, where its RMSE was approximately 5,200 ham, beating the polynomial and ARIMA models.

By integrating satellite information, government groundwater data, and machine learning algorithms, I created precise, data-driven groundwater availability predictions. These findings are essential for policymakers and urban planners to make informed decisions regarding land use, infrastructure development, and sustainable water resource management.

Riya Prashant Mandaogade (22BCE10956)

Contribution:

Contribution to Roads and Networking Module:

As part of the "Urban Planning Using Remote Sensing Images" project, I contributed to the Roads and Networking module with a focus on traffic forecasting and lane predicting. I engineered new features such as Years_Since_1961, CAGR_Growth, and Urban_Vehicle_Share to improve the dataset's analytical depth. I also merged multiple datasets using district names as a common key, ensuring a smooth and consistent data flow into our predictive models.

To support accurate forecasting, I implemented the Compound Annual Growth Rate (CAGR) formula to estimate yearly growth in vehicle run rates and developed logic to account for time-based variations in Peak Hour Factor (PHF), making our lane demand estimations more realistic. I built a function to calculate the number of road lanes required based on future traffic predictions and standard lane capacity metrics.

One of my key contributions was focusing on the Bhopal-Indore highway, where I aggregated vehicle run rate data from both districts to project specific lane requirements. I also took part in thoroughly testing and verifying all related calculations to ensure model reliability and output accuracy.

Divyansh Sinha (22BSA10178)

Contribution:

As part of the EPICS Urban Planning project, I contributed primarily to the collection of data from numerous sites for the development and integration of ML models. I cleaned the data for ground water level model as a part of preprocessing and tried various Data driven model to find the most optimal one that was to be developed

Multilayer Perceptron model was the model I choose to be used as a base of our ML model as this worked well because MLP works brilliantly when it has large number of irregular data

Multilayer Perceptron was used to predict future groundwater levels across various monitoring stations in different districts of Madhya Pradesh. The most reliable results were shown through the use of MLP. The model's predictions were used for aiding in spatial analysis and decision-making for sustainable water resource management across the state.

In addition to the data collection for the ML model, I also provided assist in the configuration of the backend of the model and dealing with any errors that arise

Overall, my contributions focused on Data collection and preprocessing, data visualization, and collaborative support to strengthen the project's foundation for smart and sustainable urban planning throughout Madhya Pradesh.

Suhani Tiwari(22BAI10141)

Contribution:

In our project on *Urban Planning Using Remote Sensing Image Interpretation*, my individual contributions focused on literature review, data source evaluation, and the design of a predictive model for water resource management in urban areas.

Early in the project, I studied multiple research papers on topics such as satellite-based urban monitoring, spatial analysis of accessibility, and predictive modelling in urban planning. This helped me understand the strengths and limitations of current approaches and guided our group in selecting relevant urban features to explore.

I took the lead on the water management aspect. As part of this, I evaluated various types of geospatial and real-time data sources by assessing parameters like coverage, consistency, and resolution. This helped ensure the datasets we used were both relevant and technically compatible with our model.

For implementation, I co-developed a machine learning pipeline to forecast groundwater availability using historical data from multiple years. I cleaned and standardized data across datasets, handled missing values, and identified core features such as annual replenishable groundwater and irrigation draft. After experimenting with several regression models, I finalized a Ridge Regression model due to its balanced performance and ability to prevent overfitting. I also worked on visualizing the results through time-series plots and residual analysis to interpret the accuracy and behaviour of the model. The final output included future water availability projections, which can be valuable for long-term urban water planning.

Finally, I authored the entire methodology section for this component, detailing the model structure, experimental outcomes, challenges (such as uneven data quality and resolution gaps), and future directions—including recommendations for incorporating spatiotemporal features and higher-resolution satellite data.

Devansh Kalura (22BCE10810)

Contribution:

As part of the EPICS Urban Planning project, I contributed primarily to the development and

integration of data-driven solutions. I developed a machine learning model using a fully connected feedforward neural network, also known as a Multilayer Perceptron (MLP), to predict future groundwater levels across various monitoring stations in different districts of Madhya Pradesh. After experimenting with several other machine learning models, the MLP approach provided the most reliable and accurate results. The model’s predictions were successfully visualized on a geographic map, aiding in spatial analysis and decision-making for sustainable water resource management

across the state.

In addition to the ML model, I compiled and provided clean, structured data in CSV format for healthcare institutions and schools across different districts. This dataset supported planning for

equitable access to essential services and infrastructure and helped inform district-level decision- making.

I also extended technical support to the Land Use Land Cover (LULC) group during the initial phase of the project. My assistance involved helping with QGIS operations, plugin installation, and

troubleshooting errors, which ensured a smoother workflow for the team. This cross-functional collaboration improved overall project cohesion and data compatibility.

Overall, my contributions focused on integrating predictive modeling, data visualization, and collaborative support to strengthen the project’s foundation for smart and sustainable urban planning throughout Madhya Pradesh.

Parth Sharma (22BSA10107)

Contribution:

In the "Urban Planning Using Remote Sensing Images" project, I contributed to the Roads and Networking module with a focus on data preparation, model readiness, and documentation. I began by verifying datasets for missing values, inconsistencies, and mismatches in vehicle usage data to ensure a reliable foundation. I applied One-Hot Encoding to convert categorical variables—especially the District column—into a machine-readable format and used StandardScaler to normalize the features before model training.

To improve reliability and user experience, I implemented robust error handling for cases such as null values or unrecognized districts. I also evaluated real-world traffic scenarios, including Bhopal-Indore connectivity and shifting vehicle usage patterns, to align predictions with practical conditions.

Anushka Sachin Mandekar (22BHI10011)

Contribution:

In the population forecasting model project, I played a key role in handling the end-to-end pipeline of data collection, processing, analysis, and model development. I began by gathering district-wise population and demographic data from authentic public sources such as census reports and government portals. After ensuring data quality through cleaning and normalization, I worked on feature engineering by incorporating time-based attributes and relevant socio-economic indicators. For model development, I experimented with various machine learning algorithms including Linear Regression,
Random Forest, XGBRegressor, and LightGBM to predict future population counts and density levels for different districts over 5 to 10-year periods. I focused on hyperparameter tuning and cross-validation techniques to enhance model accuracy and reduce overfitting. I also contributed to the evaluation phase, where I analyzed the performance metrics like RMSE and R^2 to determine the best-fit model for our forecasting goals. My work allowed the team to generate reliable, data-driven predictions that could be used for long-term urban planning and infrastructure development.
In parallel, I have also been actively contributing to the Land Use and Land Cover (LULC) model. This part of the project involves the use of satellite imagery and remote sensing data to classify land types—such as vegetation, water bodies, and built-up areas—and analyze their spatial distribution. I assisted in preprocessing the remote sensing images, applying enhancement techniques, and building classification models to extract actionable insights. My involvement in both the population forecasting and LULC models has allowed me to bridge the gap between machine learning and geospatial analysis, contributing significantly to the project's broader goal of supporting sustainable and informed urban development
Another key part of my contribution was the integration of migration trends: both in-migration and out-migration into the model. I gathered migration-related data from census migration tables and labor reports, and engineered features such as net migration rates and migration-related push/pull factors. These were critical for capturing population dynamics, especially in rapidly urbanizing or depopulating regions.

Dhaani Bahl (22BCE11071)

1. **Research and Literature Survey**

I conducted a detailed review of 8 research papers focused on urban planning, road networks, and satellite image processing. Additionally, I carried out a comprehensive survey of 15 scholarly

articles centered on land use and land cover (LULC) modeling.

2. **Geospatial Data Collection and Preprocessing**

I extensively worked with geospatial datasets derived from a wide range of reliable sources including:

- **Sentinel-2A (Level 2A) satellite imagery** from platforms like Copernicus, USGS, Bhuvan NRSC, Bhoonidhi NRSC, ESA, EUMETSAT, Earthdata NASA, HDX, AWS, Geoportal MP, MP Town Planning, and OpenStreetMap.
- Collected and processed **504 datasets** of Sentinel-2A images (13 bands), encompassing **266 zip files** covering the period from 2020 to 2025.
- Extracted optical imagery and RGB bands (Bands 2, 3, and 4) for high-resolution LULC analysis.
- Acquired **LISS-4 (5.6m)** resolution data from Bhuvan for additional spatial precision.

3. **Data Integration and Visualization**

I imported and processed satellite imagery using **QGIS**, primarily utilizing the **WMS (Web Map Service)** method for accessing real-time geospatial data from remote servers. The data was handled in **TISS file format**, facilitating efficient analysis.

- Developed **LULC classification maps** from Sentinel-2A data and visualized temporal land cover changes.
- Consolidated groundwater datasets from **2011, 2013, and 2020**, enabling correlation between land use and water resource trends.
- Implemented visualization workflows using QGIS and explored alternate platforms like **GRASS GIS** to overcome version-specific limitations in QGIS.

4. **Automation and Technical Enhancements**

To streamline data handling and minimize manual labor

Working on code to automate the **extraction of RGB bands from 266 Sentinel zip files**, significantly reducing preprocessing time.

Researched the use and functioning of **WMS/WCS dataset formats** and their integration within QGIS for improved data accessibility.

5. **Tool Optimization and Issue Resolution**

Throughout the project, I encountered and addressed several challenges related to software compatibility and dataset integration. I Researched and applied alternate GIS platforms like **GRASS GIS** to mitigate tool issues and also worked on resolving errors linked to QGIS version dependencies, ensuring smoother workflows and continued project progress

Conclusion

The project successfully demonstrated how the integration of satellite imagery, Geographic Information Systems (GIS), and Artificial Intelligence (AI) can revolutionize urban planning in growing regions like Sehore, near VIT Bhopal University. By leveraging remote sensing and advanced predictive models, the team was able to address critical urban challenges such as population growth, traffic congestion, water resource management, and land use planning.

The predictive models developed—ranging from road infrastructure classifiers to groundwater availability estimators—offer scalable, data-driven tools for sustainable development. Despite challenges in data preprocessing and system limitations, the project outcomes provide a robust framework that can guide policymakers, planners, and local authorities in making informed, future-ready decisions.

This initiative serves not only as a technical accomplishment but also as a foundational step toward smarter cities that are environmentally resilient, socially inclusive, and infrastructure-efficient. The methodology and insights gained can be adapted across similar urban centers throughout India, paving the way for a more sustainable urban future.

Reference:

1. Geoportal Madhya Pradesh. (n.d.). Retrieved from https://geoportal.mp.gov.in/geoportal
2. Bhuvan. (n.d.). Retrieved from https://bhuvan.nrsc.gov.in/home/index.php
3. Fischel, W. (2000). *Zoning and Land Use Regulation. Vol. II: Civil Law and Economics*. Retrieved from https://www.amazon.com/Zoning-Land-Use-Regulation- Civil/dp/1843764624
4. Lens, M. C. (2022). Zoning, land use, and the reproduction of urban inequality. *Annual Review of Sociology*, *48*(1), 421-439. https://doi.org/10.1146/annurev-soc-030420-122027
5. I.J. (2014). Satellite image processing for land use and land cover mapping. *I.J. Image, Graphics and Signal Processing*, *10*, 18-28. https://doi.org/10.5815/ijigsp.2014.10.03
6. Hassan, M. I., & Elhassan, S. M. M. (2020). Modeling of urban growth and planning: A critical review. *Journal of Building Construction and Planning Research*. Retrieved from https://www.scirp.org/journal/paperinformation.aspx?paperid=107665
7. Li, F., Yigitcanlar, T., Nepal, M., Nguyen, K., & Dur, F. (2023). Machine learning and remote sensing integration for leveraging urban sustainability: A review and framework. *Sustainable Cities and Society*, *96*, 104653. https://doi.org/10.1016/j.scs.2023.104653
8. Das, S., & Angadi, D. P. (2022). Land use land cover change detection and monitoring of urban growth using remote sensing and GIS techniques: A micro-level study. *GeoJournal*. https://doi.org/10.1007/s10708-021-10479-4
9. Alsharif, A. A. A., Pradhan, B., Mansor, S., & Shafri, H. Z. M. (2015). Urban expansion assessment by using remotely sensed data and the relative Shannon entropy model in GIS: A case study of Tripoli, Libya. Retrieved from http://www.researchgate.net/publication/274479575
10. Sakieh, Y., Amiri, B. J., Danekar, A., Feghhi, J., & Dezhkam, S. (2015). Simulating urban expansion and scenario prediction using a cellular automata urban growth model, SLEUTH, through a case study of Karaj City, Iran. *Journal of Housing and the Built Environment*. https://doi.org/10.1007/s10901-014-9407-3

11. Yu, J., Zeng, P., Yu, Y., Yu, H., Huang, L., & Zhou, D. (2022). A combined convolutional neural network for urban land-use classification with GIS data. *Remote Sensing*, *14*(5), 1128. https://doi.org/10.3390/rs14051128

12. Varshney, K. R., Chen, G. H., Abelson, B., Nowocin, K., Sakhrani, V., Xu, L., & Spatocco, B. L. (2015). Targeting villages for rural development using satellite image analysis. Retrieved from https://www.researchgate.net/publication/277455164

13. Patil, A., & Panhalkar, S. (2023). A comparative analysis of machine learning algorithms for land use and land cover classification using Google Earth Engine platform. *Journal Of Geography*, *17*(2), 96-112. https://doi.org/10.58825/jog.2023.17.2.96

14. Talukdar, S., Singha, P., Mahato, S., Shahfahad, Swades, P., Liou, Y.-A., & Rahman, A. (2020). Land-use land-cover classification by machine learning classifiers for satellite observation. *Remote Sensing*, *12*(7), 1135. https://doi.org/10.3390/rs12071135

15. Ramadan, G. F., & Hidayati, I. N. (2022). Prediction and simulation of land use and land cover changes using open source QGIS: A case study of Purwokerto, Central Java, Indonesia. *Indonesian Journal of Geography*. https://doi.org/10.22146/ijg.68702

16. Bhalla, N. (2006, December 14). India says its carbon emissions not harming world. *Reuters*. Retrieved from https://www.reuters.com/article/us-india-climatechange-idUSDEL17065120061214

17. DeRooy, Y. (1974). Price responsiveness of the industrial demand for water. *Water Resources Research*, *10*(2), 403-410. https://doi.org/10.1029/WR010i002p00403

18. Dugger, C. (2006, August 22). Need for water could double in 50 years, U.N. study finds. *The New York Times*. Retrieved from https://www.nytimes.com/2006/08/22/world/22water.html

19. Dupont, D. P., & Renzetti, S. (2001). The role of water in manufacturing. *Water Resources Management*, *15*(2), 87-98. https://doi.org/10.1023/A:1011155413691

20. Giridharadas, A. (2005, August 20). Water-scarce India weighs a return to ancient practices. *The New York Times*. Retrieved from https://www.nytimes.com/2005/08/20/world/asia/20iht-water.html

21. Sengupta, S. (2006, September 30). India digs deeper, but wells are drying up. *The New York Times*. Retrieved from https://www.nytimes.com/2006/09/30/world/asia/30india.html

22. UNESCO. (2003). Political inertia exacerbates water crisis, says World Water Development Report. Retrieved from https://en.unesco.org/water

23. Clark, M. J. (1998). Putting water in its place: A perspective on GIS in hydrology and water management. *Hydrological Processes*, *12*(6), 823-834. https://doi.org/10.1002/(SICI)1099-1085(19980630)12:6<823::AID- HYP631>3.0.CO;2-X

24. Merem, E. C., et al. (2017). Analyzing water management issues using GIS: The case of Nigeria. *Sustainability, 9*(4), 554. https://doi.org/10.3390/su9040554

25. Putri, A. A., & Aditya, T. (2017). 3D modelling and visualization of drinking water supply system using 3D GIS. *Proceedings of the International Conference on Advances in*

26. *Engineering Science and Management* .https://doi.org/10.1109/INAES.2017.8068574 Singh, S., Samaddar, A. B., & Srivastava, R. K. (2010).

27. Sustainable drinking water management strategy using GIS: Case study of Allahabad city (India). *Management of Environmental Quality: An International Journal, 21*(5), 573-588. https://doi.org/10.1108/14777831011055361

28. Omondi, A. N., Ouma, Y., Kosgei, J. R., Kongo, V., Kemboi, E. J., Njoroge, S. M., Mecha, A. C., & Kipkorir, E. C. (2023).

29. Estimation and mapping of water quality parameters using satellite images. *Water Practice and Technology, 18*(1), 97-108. https://doi.org/10.2166/wpt.2023.039

30. Ministry of Road Transport and Highways (MoRTH). (n.d.). *Road safety reports*. Retrieved from https://morth.nic.in/

31. Madhya Pradesh Geoportal. (n.d.). *Geospatial data*. Retrieved from https://www.geoportal.mp.gov.in/

32. Ministry of Housing and Urban Affairs (MoHUA). (n.d.). *Smart City Mission data*. Retrieved from https://smartcities.gov.in/

33. Ministry of Road Transport and Highways (MoRTH). (n.d.). *Vehicle registration data (Vahan)*. Retrieved from https://vahan.nic.in/

34. Institute of Urban Transport (India). (n.d.). *Urban transport research data*. Retrieved from https://www.urbantransport.org/

Biodata

1) Devansh Kalura
Email: devanshkalura2022@vitbhopal.ac.in Location: Vikasnagar, Dehradun
Education: B.Tech in Computer Science and Engineering (Core), 3rd Year, VIT Bhopal University
Skills:

- **Python, Javascript, C/C++**
- **Full Stack Web Development**
- **Tools: VS Code, Git, XAMPP**

Certifications:

- **Cloud Computing – NPTEL, IIT Kharagpur (May 2024)**
- **The Bits and Bytes of Computer Networking – Coursera, Go**

Projects:

- **Chat Application**
- **Doctor Appointment Website**
- **Skincare e-commerce website**

Interests: Fitness, Books

---

2) Namrata Bhutani
Email: namratabhutani2022@vitbhopal.ac.in
Location: Agra, Uttar Pradesh
Education: B.Tech(Computer Science and Engineering), Pre Final Year, VIT Bhopal University
Skills:

- **Python, Java**
- **Machine Learning**
- **Tools: TensorFlow, Scipy**

Certifications:

- **Bits and Bytes of Networking -Cousera**
- **Python for data science and ML -Udemy**

Projects:

1. **Driver Drowsiness Detection System**
2. **Hospital Management System**

---

3) Riya Prashant Mandaogade
Email:riyaprashantmandaogade2022@vitbhopal.ac.in
Location: Navi Mumbai, Maharashtra, India
Education: B.Tech in Computer Science and Engineering (Core), 3rd Year, VIT Bhopal University
Skills:

- Python, C++
- Web development: HTML, CSS, Node.js
- Tools : VS Code, Git

Certifications:

- Cloud Computing – NPTEL, IIT Kharagpur (May 2024)
- The Bits and Bytes of Computer Networking – Coursera, G

Interests:
Singing, Guitar, Painting
Projects:

- H- Connect Hostel Management System
- Impromptu - Facial recognition based gallery and social media website

---

4) Aastha Pancholi
Email:aasthapancholi2022@vitbhopal.ac.in
Location: Rajkot,Gujarat, India
Education: B.Tech CSE with Health Informatics,
3rd Year, VIT Bhopal University Skills: Python, Java
Certifications: Vityarthi Python
Projects: Graphic Password Authentication Bio Enquire
Interests: Listening Music, Coloring

---

5) Parth Sharma

Email: parthsharma2022@vitbhopal.ac.in
Location: Jaipur, Rajasthan, India
Education: B.Tech CSE (Cloud Computing and Automation)3rd Year, VIT Bhopal University
Skills:

- **Python, C++**
- **Web development**
- **Tools : VS Code, Git**

Certifications:

- **Applied Machine Learning In Python– University of Michigan, Coursera**

Projects:
Credit Card Fraud Detection Algorithm (Group Project) Chatbot (Individual Project)
Interests: Reading Books, Travelling.

---

5) Anushka Mandekar

Name: Anushka Mandekar

Email: anushkasachinmandekar2022@vitbhopal.ac.in

Location: Nashik, Maharashtra

Education: B. Tech in Computer Science and Engineering with Health Informatics, 3rd Year, VIT Bhopal University

Skills: Python, Java, SQL, MS Excel, Machine Learning

Certifications:

- IBM Data Science Orientation
- Google Cloud Generative AI Fundamentals
- Coursera Data cleansing and Analysis in Excel

Projects and Technologies:

1. ***Omdena Nepal Local Chapter*** (*HTML, Python, Beautiful Soup)* -Project contributor-Scraper, Analyst
2. ***Malaria Detection GUI*** (*ML, Python (pandas, numpy, keras, seaborn, scikit-learn, tensorflow)* -Team Lead, Model developer
3. ***Chest Disease Detection*** (*Python, PyTorch)*- Personal Project

---

7)_Dhani Bahl
Email: **dhaanibahl2022@vitbhopal.ac.in**
Location: Bhopal, Madhya Pradesh
Education: B.Tech(Computer Science and Engineering) Pre-Final Year, VIT Bhopal University
Skills:

1. **Python, Java**
2. **Machine Learning**
3. **Tools: TensorFlow, VS code, PyCharm, Git**
4. **Web development: HTML, CSS, JS, Bootstrap, Node.js**

Certifications:

- **Bits and Bytes of Networking - Coursera**
- **Cloud Computing - NPTEL, IIT Kharagpur (May 2024)**

Projects:

1. **Driver Drowsiness Detection System**
2. **Hospital Management System**

---

8) Divyansh Sinha
Email:divyanshsinha2022@vitbhopal.ac.in **Location: Lucknow, Uttar Pradesh**
Education: B.Tech(Computer Science and Engineering) Cloud computing , 3rd Year, VIT Bhopal University Skills:

- **Python,**
- **Java**
- **Machine Learning**
- **AWS**
- **Tools : VS Code, Git**

Certifications:

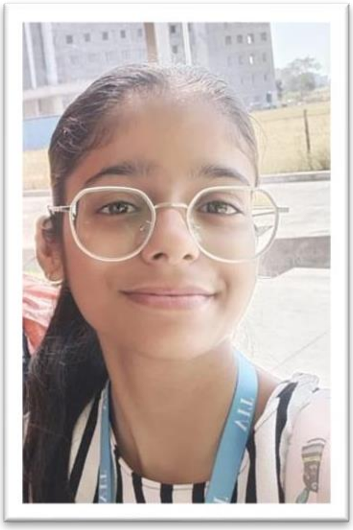
- **HTML, CSS, and Javascript for Web Developers -Cousera**

Projects:

- **web based bookstore**
- **credit card fraud detection**
- **Weather Forecasting Website**
- **Student Information Management System**

9) Suhani Tiwari
Email: suhanitiwari2022@vitbhopal.ac.in Location: Bhopal, Madhya Pradesh
Education: B.Tech(Computer Science and Artificial Intelligence)
Pre-Final Year
VIT Bhopal University
Skills:

5. **Python, Java**
6. **Tools: VS code, Git**



Certifications:

- **Bits and Bytes of Networking - Coursera**
- **Cloud Computing - NPTEL, IIT Kharagpur (May 2024)**

---

10) Nyasha Ojha - https://www.linkedin.com/in/nyasha-ojha-256548251/
Email: nyashao2910@gmail.com

Location: Bhilai, Chhattisgarh - 490026

Education: B.Tech Computer Science and Engineering(Core)
3rd Year, VIT Bhopal University- 466114

Skills: ML algorithms, Python, Data Structures and Algorithms, C++, **Java**, Web Development, Database Management: SQL, MS Office Suite, Communication Skills, Googling

Language Proficiency: Hindi, English, Telugu, Bengali

Certifications:
Cloud Computing- NPTEL, IIT Kharagpur (April 2024)
The Bits and Bytes of Computer Networking - Coursera, Google

Projects:

1. ***CaseLawAI (Local Chapter)* -** Omdena, Jr ML Engineer
2. ***Urban Planning using remote sensing satellite imageries* -** Team Lead, Academic Project, VIT Bhopal University
3. ***Humanoid Chatbot-*** Personal Project
4. ***Doctor Appointment Booking website-*** Team member - Academic Project, VIT Bhopal University

Hobby: Drumming, Boxing