SSRG International Journal of Civil Engineering
ISSN: 2348-8352/ https://doi.org/10.14445/23488352/IJCE-V11I2P108

Volume 11 Issue 2, 80-92, February 2024

Original Article

Comparative Study of Satellite Imageries for the Vegetation Analysis with Geospatial Artificial Intelligence: Using Python and Scikit-Learn

A.S. Vickram[1], S. Vidhya Lakshmi[2*], Anand Raju[2], V.P. Veeraraghavan[3]

[1]*Department of Biosciences, Saveetha School of Engineering, SIMATS, Tamilnadu, India.*
[2]*Department of Civil Engineering, Saveetha School of Engineering, SIMATS, Tamilnadu, India.*
[3]*Centre of Molecular Medicine and Diagnostics, Saveetha Dental College and Hospitals, SIMATS, Saveetha University, Tamilnadu, India.*

[*]*Corresponding Author : vidhyalakshmis.sse@saveetha.com*

Received: 12 December 2023 Revised: 11 January 2024 Accepted: 12 February 2024 Published: 29 February 2024

Abstract *- The datasets were collected for the urban area of Salem, which is located in India. As part of the investigation, four different datasets were gathered. A machine learning process was applied to the satellite imagery, with seventy percent of the area designated as the training set data and the remaining thirty percent utilized as test data. Using the K-means Clustering method, the research primarily concentrated on evaluating the first stage of vegetation in Salem City. A visual representation of the results obtained can be found in pictures 1, 2, 3, and 4. The statistical analysis of the research region reveals that areas with limited vegetation are experiencing consistent annual growth, with an exceptionally substantial rise recorded between February 2019 and February 2024.*

Keywords *- Artificial Intelligence, Scikit-learn, Types of vegetation, Remote sensing, Python.*

1. Introduction

By exploiting location-based data obtained from Geographical Information Systems (GIS), spatial analytics examines the geographical properties of geospatial datasets. This is accomplished through the utilization of location data. Many of today's human endeavours occur in the actual and virtual worlds.

The research gap identified as a part of this study is that most of the research articles reviewed were focused on the supervised classification from the vector datasets that were analyzed using the machine learning approach, as in this study, the major focus is emphasized on the unsupervised classification of the raster imagery and found that it is a challenging task as there were only few studies exist till today in the aspects of handling Geospatial raster imagery.

The changes have brought about a significant revolution in Geographic Information Science (GIScience) since the digital realm offers new perspectives on the functions of temporal and spatial elements, such as the challenges of losing one's sense of place and the potential to do away with time constraints. The referred article explores the potential and threats at the nexus of cyberspace and physical space, focusing on data visualisation and analytics. Artificial intelligence, machine learning, and virtual reality could further improve these capabilities. To promote sustainability and address complex challenges associated with geospatial applications and other technological advancements in environmental and urban sciences, the method is proposed to integrate cyber and geographical data processing and analysis as a synergistic partnership (Chen et al., 2023).

Dependability is essential for Landslide Susceptibility Mapping (LSM) to prevent and mitigate catastrophes (Wei et al., 2022). The author used machine-learning models to categorise unobserved fishing behaviours and provided set and trip-level descriptions (Suter et al., 2022). Researchers have employed numerous methodologies to examine the temporal and spatial distribution of Non-Polluting Petroleum (NPP) in the open ocean. Satellite data and the Vertically Generalized Production Model (VGPM) are examples.

However, Estuaries and coastal waterways are often unsuitable for these algorithms (Xu et al., 2022). The data was used to create a leishmaniasis prediction map, build a model, and assess the outcomes using 70:30 ratios and the holdout approach, respectively (Shabanpour et al., 2022). A total of 4345 agricultural subsoil samples and sixteen environmental parameters were used to create three machine learning models:

Random Forest (RF), eXtreme Gradient Boosting (XGBoost), and Random Forest + Residuals Kriging (RFRK) (Zhang et al., 2022). Beyond the era of operational remote sensing and digital cartography, the authors showed how new approaches to quantitative analysis of long-term urbanization processes and landscape changes are made possible by integrating historical map series information with contemporary geospatial data (Uhl et al., 2022). The limited distribution of weather stations poses a challenge and obstacle, notwithstanding their capacity to deliver precise and temporally resolved surveillance of air temperature (Ta).

Conversely, satellite data can furnish Land Surface Temperature (LST) measurements that are extraordinarily disseminated globally and exhibit a strong correlation with Ta; thus, they serve as an exceptional instrument for enhancing Ta estimation. A monthly average Ta dataset covering Taiwan from 2003 to 2020 with an accuracy of 1 km by using auxiliary and satellite-derived data. Three distinct Machine Learning (ML) approaches and seven datasets were used. The Land Surface Temperature (LST) derived from the MODIS was one of the twelve explanatory variables included in these datasets. The study aimed to identify the best dataset and method combination for Taiwan temperature (Ta) estimation (Tran and Liou, 2024).

Recently, scientific investigations have benefited from the increasing accessibility of huge surveillance databases to increase our understanding of human movement patterns. This has been done to understand better how people move around. Meanwhile, the absence of standardization in data processing pipelines for the varied data-gathering methods restricts the transferability, comparability, and repeatability of findings and approaches in quantitative human mobility analysis. This work proposes Trackintel, an open-source Python toolkit for investigating human movement (Martin et al., 2023). SNCF Réseau must efficiently manage the vegetation on the network to ensure the consistency and security of train service and the condition of the railway infrastructure.

To accomplish this, the administration of the French railway infrastructure must be aware of the diverse array of vegetation that grows adjacent to the network. A processing chain comprising satellite images with an exceptionally high spatial resolution has been constructed to compile a vegetation inventory. The vegetation inventory along the French railway network was industrialised by employing a machine learning methodology for supervised classification, streamlining the processing pipeline, and repurposing previously trained models.

A significant temporal frequency was attained through this (Onody et al., 2023). Deterioration models are utilised whenever possible to assist in prioritising and maintaining bridges. Two broad classifications may be applied to these models: stochastic and deterministic. There is a possibility that both physical models and mechanical models produced by Artificial Intelligence (AI) could display stochastic or deterministic properties (Srikanth and Arockiasamy, 2020). It is possible to ascribe the extensive use of hyperspectral imaging in intelligence and surveillance applications to the abundant spectral content that it possesses (Yadav et al., 2019).

The development of a specialized programme known as "Artificial Intelligence for Digital Forest (AID-FOREST)" is required to implement this concept. By utilizing point clouds obtained through Mobile Terrestrial Laser Scanning (MTLS), the technology generates a diverse range of precise and practical dendrometric and forest stand attributes (López Serrano et al., 2022). The Multivariate Adaptive Regression Spline (MARS), a novel data-driven non-linear approach, is presented as a forecasting tool that successfully clarifies the geochemical complexity present in regolith (Majeed et al., 2022).

2. Data and Digital Image Processing

The research may use either raster or vector data, depending on the immediately available data. The method will be modified acceptably. The nature of vector data, which consists of borders, locations, and lines, makes it far simpler to manipulate than raster data. On the other hand, raster data is measured in pixels per square inch, and resolution is an extremely important factor.

Furthermore, a high level of experience in data formats is required, as the visual data utilised in geospatial research and analysis must be in the Tagged Image File format (TIF) to guarantee the most efficient analysis possible. The primary goal of digital image processing was to convert the provided raster picture into numerical data. This would allow for the semi-automated or fully automatic extraction of features and the detection of edges. The purpose of this method was to disentangle the borders of the particular areas of research interest. By utilising sophisticated algorithms, it can expedite the workflow and accomplish the processing of such data. The vast majority of programmers make use of it as a foundation for conducting grid-based or pixel-based analysis on specific datasets.

3. Information Retrieval and Spatial Statistics

Programming expertise in Python makes it possible for academics worldwide to acquire information using a series of Python modules that are freely accessible to the public. These modules include Principal Component Analysis (PCA), TfidfVectorizer, CountVectorizer, NearestNeighbors, and others. When attempting to discover clustered data, spatial pattern analysis is frequently employed. This is accomplished by analyzing groups of pixel values. Utilizing spatial interpolation allows for the determination of values that were previously ignored at the locations that have been provided.

4. Terrain Analysis in Python

Paleoclimate scientists use spectrum analysis methods extensively to look into cyclical events that might have affected past climate fluctuations. Climate time series data has become more widely available, and advances in computational technology have made it easier to consolidate this methodology. However, visual representations of spectral analysis results have been slow to emerge. Using two-dimensional colour graphs, time-frequency analysis can detect periodic signals that change over time by plotting spectral bands that might be seen in the image background.

The paleoclimate literature contains numerous examples of these depictions, including the evolutionary Fast Fourier Transform (FFT) spectrogram, continuous wavelet analysis, and the recently introduced synchrosqueezing transform. Our methodology is based on a stack of spectral analysis findings from thousands of fixed interval time series. These time series were derived from a lengthier paleoclimate time series with irregular sampling. We assume that the behaviour of these time series is stationary and non-evolutive. The time series of interest is obtained from the LR04 Global Pliocene-Pleistocene Benthic 18O stack and summer insolation data for the previous 5.3 million years at 65 N. For this, the Lomb-Scargle periodogram technique is utilised. The visualisation is improved and made more transparent by integrating state-of-the-art terrain analysis tools such as hillshading, slope, and colour mapping.

To accomplish a seamless transition, Python code is also employed to combine the various images. A more thorough and exact interpretation of the cyclical patterns is made possible, among other things, by the analysis's output-a graphic depiction that is both visually accurate and complete. In addition, the algorithm can hide pixels with a value below a given threshold and uses confidence levels determined from the spectral approach (Sánchez-Morales, Pardo-Igúzquiza, and Rodríguez-Tovar 2023).

5. Overview of Artificial Intelligence and Spatial Analytics with Case Studies

Weakly-supervised learning has recently been popular for classification tasks, where the true classifications are frequently murky or unreliable. Nevertheless, this learning environment has not yet been thoroughly studied for regression difficulties despite its prevalence in macroecology. It also creates a new computational paradigm for structurally incomplete and chaotic target labels. This type of setup could be necessary for multi-output regression work that requires all outputs to add up to one.

The author suggests that an algorithmic approach can be used to improve predictions and reduce noise in the target labels. To put this assertion to the test, a case study from global vegetation modelling was studied. This modelling technique involves constructing a model that uses global remote sensing data to forecast probable changes in plant cover distribution as a result of different climates. Many imperfect target baselines are utilized to determine how well the suggested method performs.

Based on the findings, the proposed partial imputation technique has the potential to reduce the number of errors that occur in the targets. In place of training with complete observations alone, it has been discovered that addressing structural incompleteness in the target labels improves the capacity to grasp global links between flora and climate. This contrasts training with full observations alone (Beigaitė, Read and Žliobaitė, 2022). Two approaches quickly rising to the top of environmental research and management are citizen science and Machine Learning (ML). Computer science and machine learning can potentially improve public engagement, which benefits other governance players. However, validity and other quality assurance considerations must be considered when using these technologies, especially in managerial situations.

By demonstrating how machine learning can advance computer science by assuring compliance with California stormwater programme laws for quality assurance, this study investigates the prevalent problem of urban trash. This investigation aims to demonstrate how machine learning may enhance computer science. To examine the predictions that five machine learning models made regarding a multiclass "Litter Index" score, a crucial regulatory metric typically evaluated exclusively by qualified professionals and site-specific, the use of quantitative data obtained from computing systems to train the models was studied.

Regarding accuracy, precision, recall, and F-1 scores, XGBoost achieved the highest possible score of 0.98, proving that it achieved the most advantageous outcomes. The earlier persuasive findings demonstrate that machine learning can be efficiently integrated into computer science evaluations and enhanced quality assurance inside a controlled setting. To this day, computer science and machine learning continue to significantly contribute to implementing waste management strategies. These two domains can potentially discover significant synergies, which may affect other aspects of environmental management administration. In environmental research and management, two approaches quickly rising to the top are citizen science and Machine Learning (ML). Public engagement, local governments, and other players in governance can reap the benefits of AI and computer science.

Considering validity and quality assurance concerns is necessary when applying these innovations to specific management situations. This study employs the pervasive problem of urban refuse to illustrate how ML could assist CS via quality assurance within the regulatory framework of California's stormwater programme. The authors evaluated

the predictions of five ML models regarding a qualitative, multiclass, site-specific "Litter Index" score using quantitative data obtained from CS. Generally evaluated exclusively by professionals, this metric is critical for regulatory purposes. With scores of 0.98 for F-1, recall, accuracy, and precision, XGBoost demonstrated superior performance. These encouraging results indicate that ML can potentially improve the dependability of CS evaluations and regulatory quality assurance. It is discovered that the integration of ML and CS can yield significant synergies that lead to innovative applications in various domains of environmental management.

This extends the existing contributions of each domain to the field of waste management (Yang et al., 2023). Since accurate crop predictions are critical to fostering social cohesion, guaranteeing food security, and attaining long-term sustainability, they are of utmost importance to farmers, researchers, governments, etc. In the past, data collection and analysis methods for yield estimation were typically costly, time-consuming, site-specific, and filled with many mistakes and uncertainties. This reviewed article presents a novel machine-learning approach that integrates topographical and environmental data with high-resolution satellite imagery to predict the variability of wheat production on a farm level.

The approach is intended to enhance the accuracy of wheat production forecasts (Singh Boori et al., 2023). When landscapes are classified, it is much simpler to consider planning for the analysis. Current unsupervised clustering algorithms for landscape classification rely on categorical input data for pattern quantification and consider only a small selection of pattern metrics. This is despite landscape patterns being significant differentiators between different types of landscapes.

Using a unique unsupervised deep learning technique called Deep Convolutional Embedded Clustering (DCEC), created a landscape typology for Switzerland to utilise continuous spatial data, including remote sensing photos, to its fullest potential. DCEC divides the input images into separate clusters while encoding lower-dimensional representations of the images in a hidden layer. Topographical, ecological, demographic, and visual modules generated from the satellite photography were also subjected to DCEC implementation. DCEC successfully separated 45 different landscape types from the endless stream of input data. In conclusion, DCEC shows promise as a fresh approach to landscape and land-system study (van Strien and Grêt-Regamey, 2022).

6. Materials and Methods

The management of geographic data is an essential component for a wide variety of sectors and applications, including urban development and transportation planning, to name a few. When handling geographic data, the first things businesses should be concerned about are the data's accuracy, consistency, and interoperability. By adopting open data initiatives, one can enhance the dissemination of information and develop a unified comprehension of spatial patterns and trends. In essence, using analytical data provided by an adept geographic data system empowers decision-makers to optimize the allocation of resources and formulate strategic plans.

Table 1. Data set information collected and utilized for the study

Data Description	Acquisition Date	Data ID	Machine Learning Tool
LandSAT8 Data	17.01.2019	LC08_L1TP_143052_20190117_2012	Python with Scikit-Learn
LandSAT8 Data	06.03.2019	LC08_L1TP_143052_20190306_20200829	
LandSAT8 Data	22.03.2019	LC08_L1TP_143052_20190322_20	
LandSAT9 Data	08.02.2024	LC09_L1TP_143052_20240208_20240208	

In particular, this study uses four datasets, explicitly focusing on Landsat8 Band 5 data spanning January to December of 2019 and 2024. Three datasets originating from 2019 are present, along with one dataset originating from 2024. The datasets were collected for the Salem study area in India. The research uses four datasets in total, applying machine learning to satellite imagery by reserving 70% of the region for training and 30% for testing.

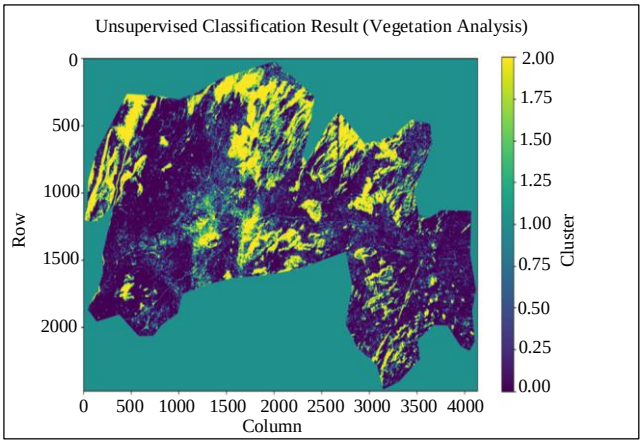

Fig. 1 Vegetation analysis using Machine Learning for the landsat imagery acquired on 17.01.2019

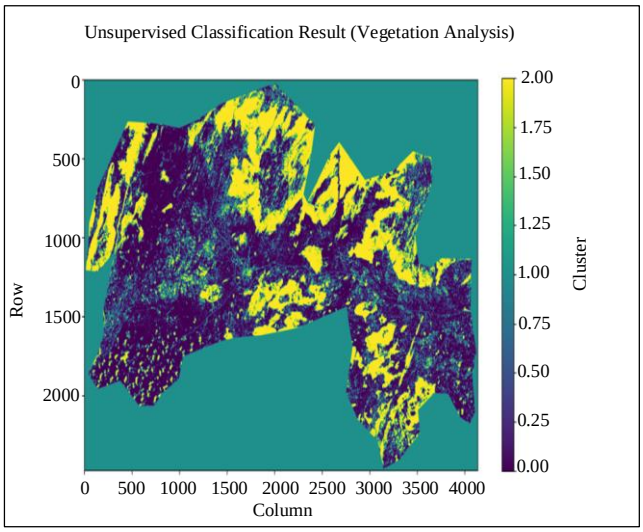

Fig. 2 Vegetation analysis using Machine Learning for the landsat imagery acquired on 06.03.2019

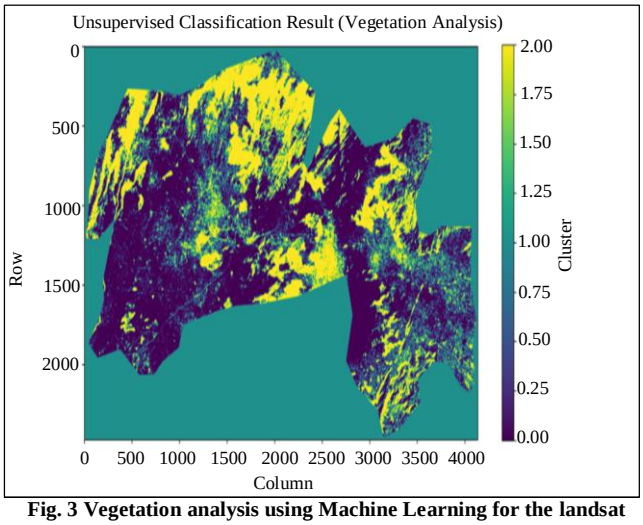

Fig. 3 Vegetation analysis using Machine Learning for the landsat imagery acquired on 22.03.2019

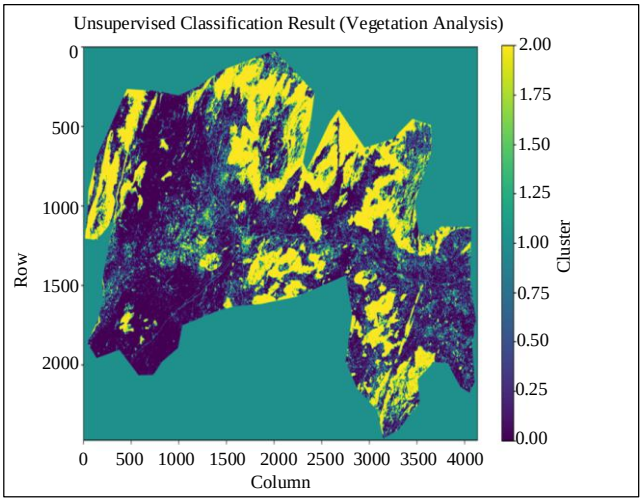

Fig. 4 Vegetation analysis using Machine Learning for the landsat imagery acquired on 08.02.2024

Figures 1, 2 and 3 make it clear that there is a decrease in the overall growth of vegetation in 2019. On the other hand, beginning in the year 2024, the imagery demonstrates a significant increase in the category of vegetation that includes both medium and low-level vegetation, such as grass, shrubs, mixed vegetation cover, and taller shrubs. Based on the study's findings that compared the data collected at various times throughout the same year (2019), it was determined that the forest fire on February 25th, 2019, caused the vegetation in the forest areas of the Yercuad region to be low.

According to the statistical plots given in Figures 5-12, it is abundantly obvious that the cluster 3 symbolises the scant vegetation. This is the case despite the fact that the vegetation cover has remained low since 2019, and it has improved based on the statistical plot for 2024. On the other hand, cluster 2 is characterised by almost the same kind of vegetation, which is known as mixed vegetation. Additionally, some shrubs are of a greater height and are spread out across the area.

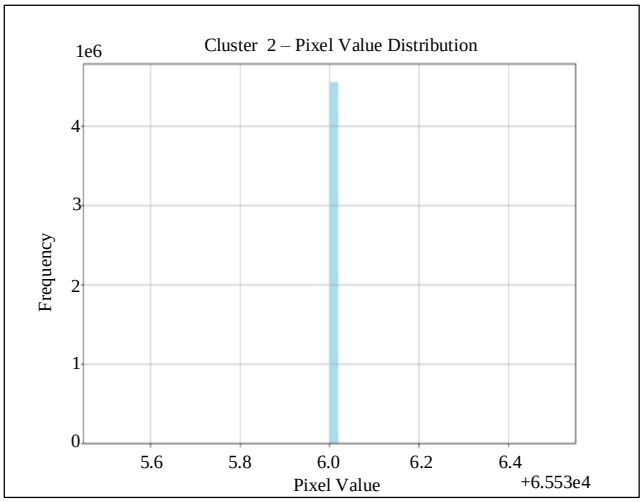

Fig. 5 Vegetation type of cluster - 2 representing moderate vegetation density on 17.01.2019

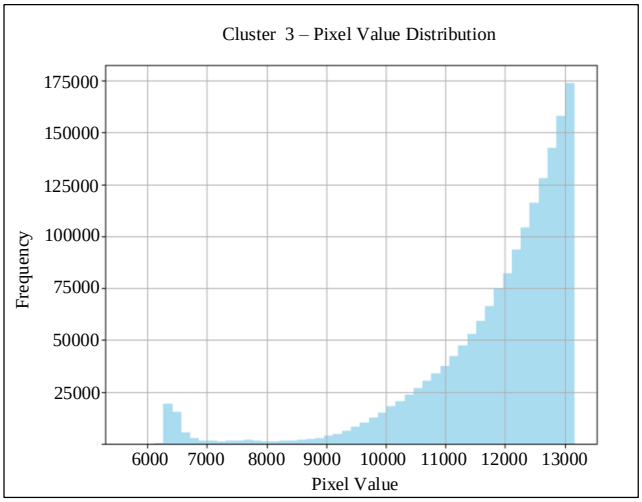

Fig. 6 Vegetation type of cluster - 3 representing high vegetation density on 17.01.2019

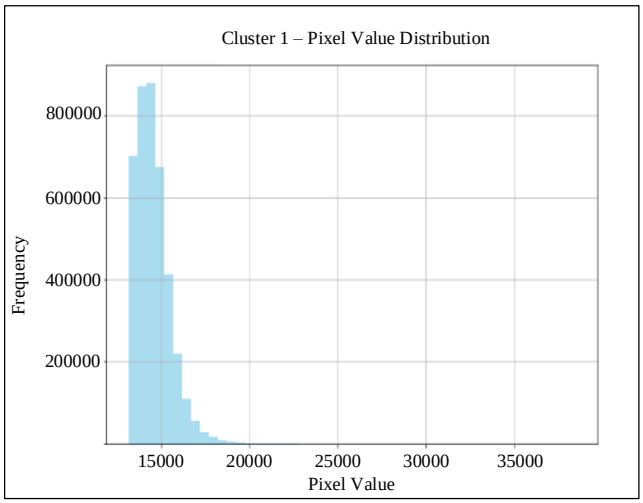

Fig. 7 Vegetation type of cluster - 1 representing low vegetation density on 17.01.2019

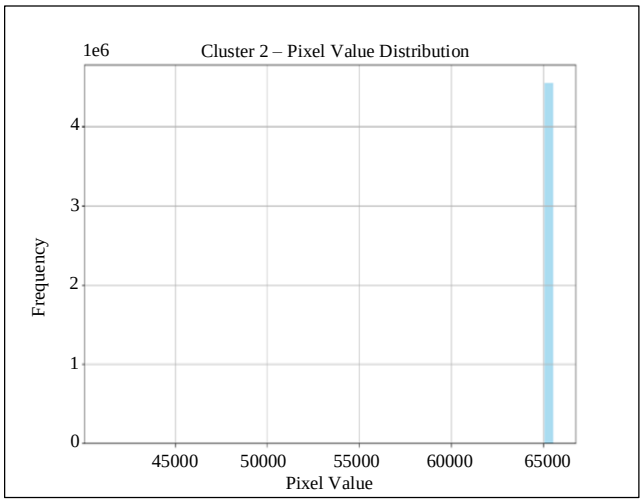

Fig. 8 Vegetation type of cluster - 2 representing moderate vegetation density on 06.03.2019

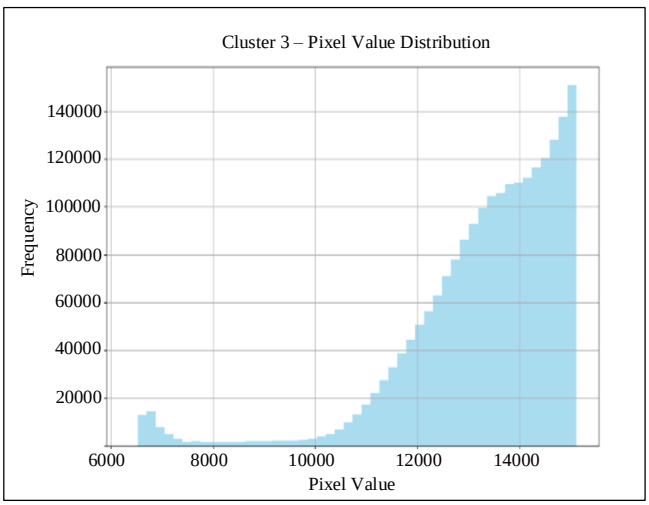

Fig. 9 Vegetation type of cluster - 3 representing high vegetation density on 06.03.2019

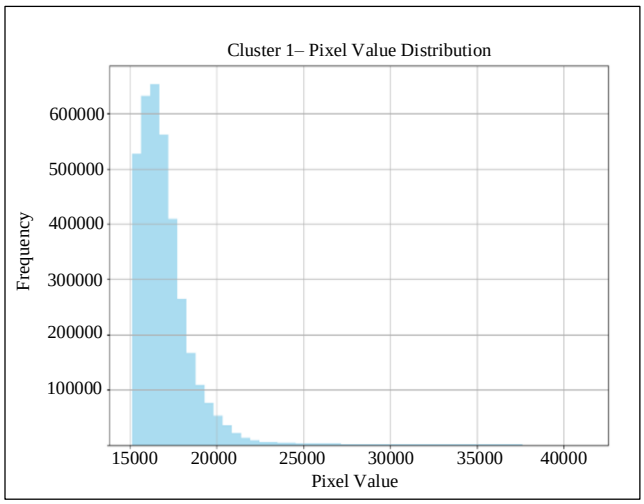

Fig. 10 Vegetation type of cluster - 1 representing low vegetation density on 06.03.2019

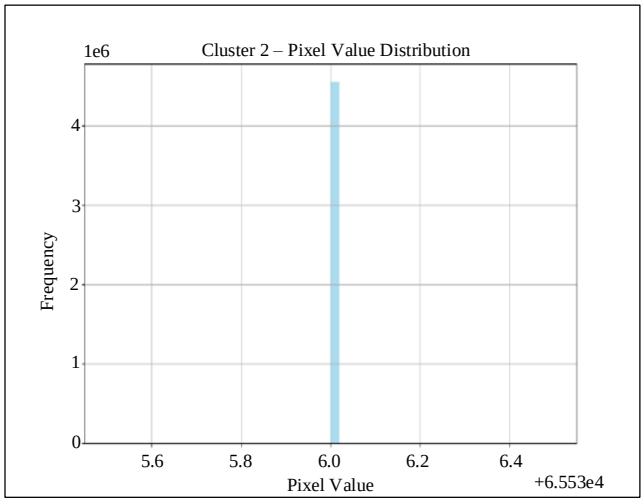

Fig. 11 Vegetation type of cluster - 2 representing moderate vegetation density on 08.02.2024

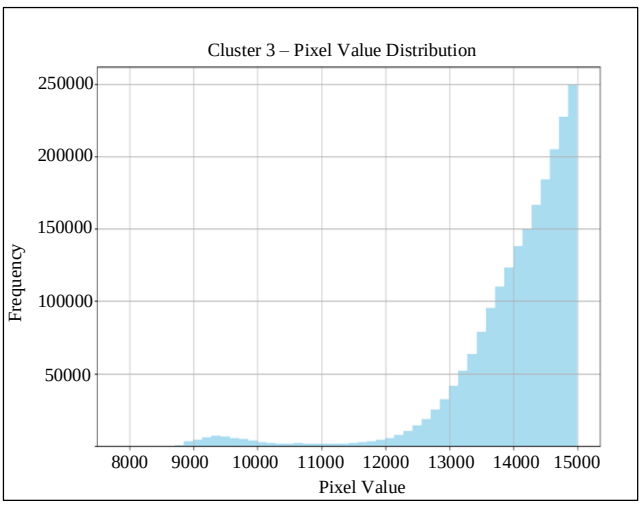

Fig. 12 Vegetation type of cluster - 3 representing high vegetation density on 08.02.2024

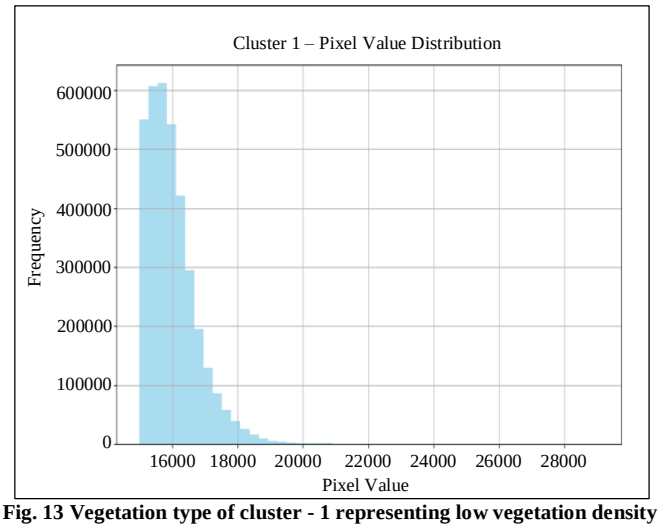

Fig. 13 Vegetation type of cluster - 1 representing low vegetation density on 08.02.2024

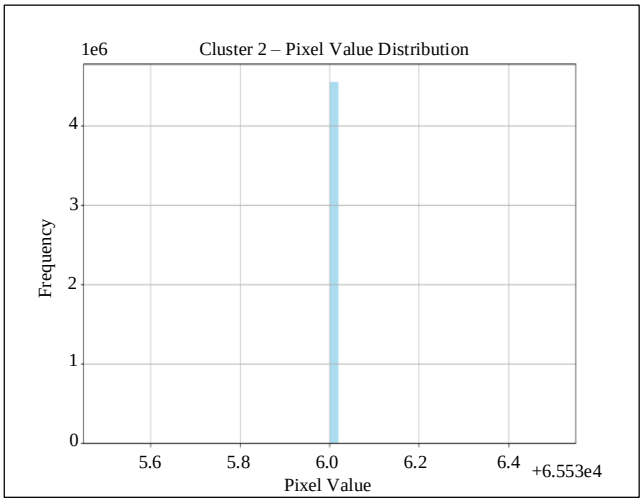

Fig. 14 Vegetation type of cluster - 2 representing moderate vegetation density on 22.03.2019

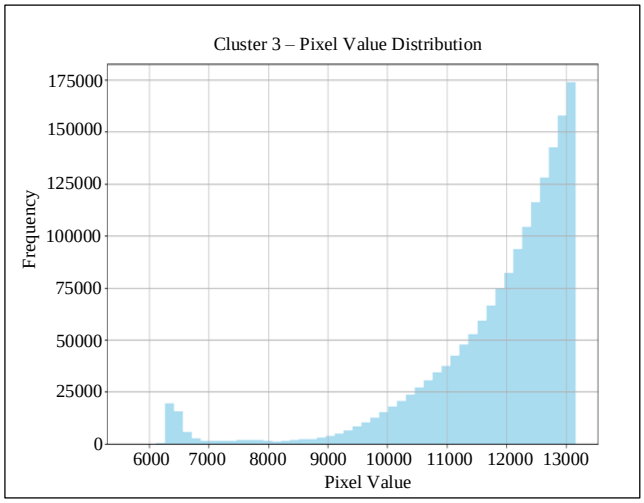

Fig. 15 Vegetation type of cluster - 3 representing high vegetation density on 22.03.2019

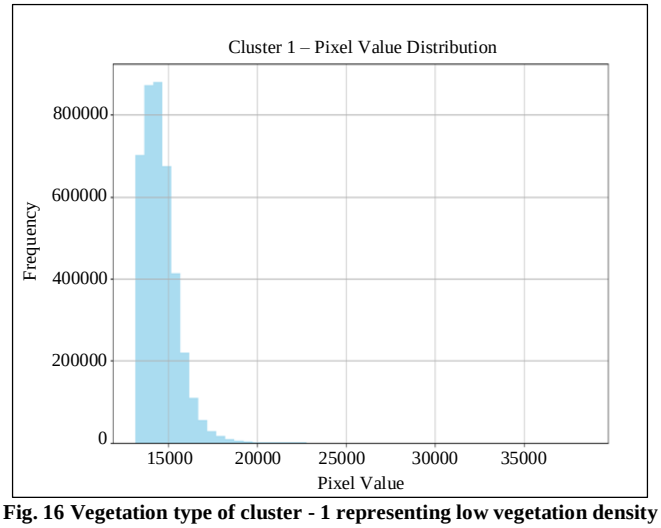

Fig. 16 Vegetation type of cluster - 1 representing low vegetation density on 22.03.2019

7. The Sample Python Code was Used to Do the Unsupervised Classification for the Landsat8 Band 5 for the Vegetation Analysis

```
Python 2.7.5 (default, May 15 2013, 22:43:36) [MSC v.1500
32 bit (Intel)] on win32
Type “copyright”, “credits”, or “license()” for more
information.
import rasterio
from sklearn.model_selection import train_test_split
import numpy as np
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
# The file path is set to the satellite imagery location
raster_file =
r’C:\Users\ANAND\Desktop\IOT\Sybmolised.tif’
# The file is opened to read the file for further processing,
with rasterio.open(raster_file) as src:
# Read the raster data
raster_data = src.read()
# Reshape the raster data to a 2D array (rows x columns,
bands)
reshaped_data = raster_data.reshape((raster_data.shape[0], -
1)).T
# Separating the imagery for training and testing sets
train_data1, test_data1 = train_test_split(reshaped_data,
test_size=0.3, random_state=42)
# Reshape the training and testing data back to the original
shape
train_data1 = train_data.T.reshape((raster_data.shape[0],
src.height, src.width))
test_data1 = test_data.T.reshape((raster_data.shape[0],
src.height, src.width))
# Write the training and testing datasets to new raster files as
per the data
with rasterio.open(‘training_data1.tif’, ‘w’, **src.profile) as
dst_train:
```

```
dst_train.write(train_data1)
with rasterio.open('testing_data1.tif', 'w', **src.profile) as
dst_test:
dst_test.write(test_data1)
# Location to the Path of the Landsat 8 image data
image_path =
r'C:\Users\ANAND\Desktop\IOT\Sybmolised.tif'
# Landsat 8 image data for band 5 is loaded for the analysis
with rasterio.open(image_path) as src:
band5 = src.read(1)
# The data band 5 is reshaped to 1D array
X = band5.flatten().reshape(-1, 1)
# Normalizing the feature vectors
X_normalized = (X - X.min()) / (X.max() - X.min())
# K-means clustering is performed,
k = 3  # Number of clusters
kmeans = KMeans(n_clusters=k, random_state=42)
kmeans.fit(X_normalized)
# cluster labels is reshaped to the original image data
cluster_labels = kmeans.labels_.reshape(band5.shape)
# The classified image is visualised as a final plot
plt.figure(figsize=(10, 8))
plt.imshow(cluster_labels, cmap='viridis')
plt.colorbar(label='Cluster')
plt.title('Unsupervised Classification Result (Vegetation
Analysis)')
plt.xlabel('Column')
plt.ylabel('Row')
plt.show()
```

The procedure follows: The programme performs unsupervised classification on satellite image data using K-means clustering to identify various land cover categories, specifically vegetation. The software begins by importing the following libraries. The matplotlib.pyplot library is used for visualisation and imported as plt. The KMeans class from the sklearn.cluster module is used for K-means clustering. The train_test_split function is taken from the sklearn-model_selection module. The numpy library is used for numerical computations and imported as np. The coordinates of the satellite imagery file (Sybmolised.tif) are provided. Using the rasterio.open() function, the application retrieves and stores the data from the raster file in the raster_data variable. In order to generate the "reshaped_data" two-dimensional array, the raster data is transformed.

Each column in this array corresponds to a band in the image, while each row represents a pixel. Using the train_test_split function, the reshaped data is divided into distinct training and testing sets. Seventy percent of the data consists of train_data1, while thirty percent comprises test_data1. Resizing the training and testing data to match the original dimensions of the raster data, which consists of multiple bands, src. width columns, and src. height rows, is achieved. Create the training and testing datasets: The training and testing datasets are written to new raster files (training_data1.tif and testing_data1.tif) using rasterio.open(). To load the Landsat 8 image data, the file path for both the satellite imagery file and the Landsat 8 image data file is specified.

The band 5 variable is populated with the extracted data from the Landsat 8 image initiated by the software. During data preprocessing, the band 5 data undergoes compression, converting it into a column vector (X) from a one-dimensional array. By ensuring that all values lie within the range of [0, 1], min-max scaling normalises the feature vectors. The K-means Clustering: The normalised feature vectors (X_normalized) are subjected to the K-means clustering procedure with a predetermined number of clusters (k = 3). Once acquired, the cluster labels are adjusted to correspond with the dimensions of the original image data (band5.shape). Visualisation: The matplotlib.pyplot.imshow() function displays the image based on the cluster labels representing the classified image. The 'Viridis' colormap is utilised to depict discrete clusters, while an accompanying color bar is a critical reference. To enhance clarity, the plot is enhanced by including a title, x-label, and y-label. Plotting can be accomplished by utilising the plt. show() function.

8. The Statistical Plots are Generated for the Vegetation Analysis, and Three Plots are Made Based on the Three Clusters, 0, 1, and 2, and a Sample Code is Given below

Three statistical plots are taken for the vegetation analysis, and three plots are made based on the three clusters, 0, 1, and 2.

```
import rasterio
from sklearn.model_selection import train_test_split
import numpy as np
from sklearn.cluster import KMeans
import matplotlib.pyplot as plt
# New path set to the raster imagery
raster_file =
r'C:\Users\ANAND\Desktop\IOT\Sybmolised.tif'
# The raster file is opened for reading the file
with rasterio.open(raster_file) as src:
    # Reading the raster data
    raster_data = src.read()
    # Reshaping the raster data to a 2D array
    reshaped_data = raster_data.reshape((raster_data.shape[0],
-1)).T
# Separating the raster data into training and testing sets
train_data, test_data = train_test_split(reshaped_data,
test_size=0.3, random_state=42)
# Reshaping the training and testing data back to the original
shape
train_data = train_data.T.reshape((raster_data.shape[0],
src.height, src.width))
test_data = test_data.T.reshape((raster_data.shape[0],
```

```
src.height, src.width))
# Writing the training and testing datasets to new raster files
with rasterio.open('training_data.tif', 'w', **src.profile) as
dst_train:
    dst_train.write(train_data)
with rasterio.open('testing_data.tif', 'w', **src.profile) as
dst_test:
    dst_test.write(test_data)
# New path is set to the Landsat 8 image file
image_path =
r'C:\Users\ANAND\Desktop\IOT\Sybmolised.tif'
# Load Landsat 8 image data for band 5
with rasterio.open(image_path) as src:
    band5 = src.read(1)
# Reshape the band 5 data to 1D array
X = band5.flatten().reshape(-1, 1)
# Normalize the feature vectors
X_normalized = (X - X.min()) / (X.max() - X.min())
# Perform K-means clustering
k = 3  # Number of clusters
kmeans = KMeans(n_clusters=k, random_state=42)
kmeans.fit(X_normalized)
 Reshaping the cluster labels to the original image shape
cluster_labels = kmeans.labels_.reshape(band5.shape)
# Plot the classified image with different colours for each
cluster
plt.figure(figsize=(10, 8))
plt.imshow(cluster_labels, cmap='viridis',
interpolation='none', aspect='auto')
plt.colorbar(ticks=np.arrange(k), label='Cluster')
plt.title('Unsupervised Classification Result (Vegetation
Analysis)')
plt.xlabel('Column')
plt.ylabel('Row')
# Calculating and plotting the statistical distribution for each
cluster
for i in range(k):
    plt.figure(figsize=(8, 6))
    plt.hist(X[kmeans.labels_ == i], bins=50, color='skyblue',
alpha=0.7)
    plt.title(f'Cluster {i+1} - Pixel Value Distribution')
    plt.xlabel('Pixel Value')
    plt.ylabel('Frequency')
    plt.grid(True)
    plt.show()
```

The above code is explained in detail as follows: The given code provides a thorough procedure for performing vegetation analysis using satellite imageries. The first step is to load essential libraries such as rasterio, scikit-learn, numpy, and matplotlib. The processing of raster data is, therefore, the primary emphasis of the code, beginning with loading a raster image file called "Sybmolised.tif" through rasterio. The data is reorganised into a two-dimensional array to make subsequent analysis easier. In the following step, the data is divided into training and testing sets using the train_test_split function available in scikit-learn. This function is crucial for both the training and assessment of the model. The training and testing data are then written to new raster files after being reshaped back to their initial shape at the beginning of the process.

The next step is to load a different raster image file called "Sybmolised.tif" to read band 5 data. After that, the data is normalised and flattened into a one-dimensional array. The normalised data is then put through a K-means clustering process with k equal to three groups to categorise different types of vegetation based on their spectral properties. After the generated cluster labels are moulded to the shape of the initial image, a classed image is plotted, with each cluster represented by a distinct colour according to the classification.

Furthermore, the algorithm goes beyond simple visualisation by calculating and presenting each cluster's statistical distribution of pixel values using histograms. This provides insight into the properties of the vegetation captured by each cluster. This approach incorporates data preparation, machine learning, and visualisation techniques to gain useful insights from satellite imagery for vegetation analysis.

9. Geospatial Solutions that are Based on the Cloud Computing

The management and analysis of geographic data can be strengthened using cloud-based geospatial solutions, which take advantage of the cloud's scale and adaptability. The provision of computing resources on demand is made possible by these technologies, which makes it possible for users to tailor their resource utilization to their specific requirements. Security and scalability aspects included in these apps are provided to enhance the efficiency and dependability of cloud-based geospatial applications. As a consequence of this, businesses that are dependent on data analysis and geographic information systems increasingly regard these solutions to be indispensable.

10. Combining Artificial Intelligence with Geographical Information Systems

It is feasible to improve the capabilities of both systems by merging Geographical Information System (GIS) and Artificial Intelligence (AI). This is accomplished by maximizing the benefits that each system has to offer. AI will be able to provide more advanced geographic analysis and predictive modelling opportunities due to the intelligent automation of GIS.

Assigning resources, responding to catastrophes and building cities are just a few examples of real-world applications that might considerably benefit from merging GIS and AI. In recent years, the coronavirus disease outbreak (COVID-19) has emerged as one of the most complex international problems.

This previous literature study uses a multilayer perceptron artificial neural network topology to evaluate the relative significance of putative explanatory variables (n = 75) in relation to COVID-19 prevalence and mortality, given the absence of global studies on the spatiotemporal modelling of the virus. Ten variable importance analysis methodologies were used to determine the relative importance of the explanatory factors.

The primary conclusions showed that a small handful of variables remained the most critical variables during all time. Population density and unemployment were two of the most important characteristics with the highest relevance ratings to COVID-19 prevalence. Health-related factors, such as the availability of hospital beds and the incidence of diabetes, are significant predictors of COVID-19-related mortality. The results of this study may offer insightful information to public health policymakers that will help them track the spread of disease and make more informed decisions (Kianfar et al., 2022).

11. Geographical Analytics: The Latest Developments

Lakes characterized by low water mobility are especially vulnerable to Bisphenol A (BPA) accumulation, which threatens aquatic life in numerous human-polluted watersheds. While prior research has examined the detrimental impacts of BPA concentrations on marine organisms, the absence of comprehensive data prevents us from assessing the ecological peril in watersheds. 164 BPA data points were collected from Taihu Lake to determine the spatiotemporal distribution and associated hazards of BPA. Following that, machine learning models were constructed utilising Support Vector Machines (SVM), Random Forests (RF), and Least Square Regression (LSR). Following this, monthly watershed projection maps for temperate lakes were generated.

By virtue of its enhanced robustness against chaotic data, the RF model exhibits superior performance compared to the other two methodologies. The RF model demonstrated respectable predictive capability on the modelling dataset, as evidenced by its RMS errors of 17,499 and 39,645 on the training set and 0.607 and 0.927, respectively, on the validation set. The cartographic representations indicated that regions prone to human intervention contained the most elevated concentrations of BPA pollution.

Moreover, an increase in precipitation could potentially facilitate the migration of BPA into aquatic ecosystems. In addition, 42 BPA data points were obtained from Dianchi Lake and projected by the model. Despite a decline in the accuracy of the model's predictions, the findings indicated that most predicted data points were within a tenth of the measured data. Upon assessing the ecological hazards in both lakes, our attention can be directed towards the most perilous regions.

By conducting a thorough examination of the spatiotemporal distribution of an innovative trace pollutant that disrupts endocrine systems in aquatic habitats, the previous research proposed a novel approach to evaluate the ecological hazard posed by bisphenol A impartially (Wang et al., 2024). Agricultural nutrient runoff significantly contributes to river and coastal water system degradation. If issues with water quality can be better understood by water quality modelling, then suitable measures to improve water quality can be implemented. Nutrient model calibration based on complicated processes necessitates a plethora of input parameters and incurs high computing costs.

In comparison to process-based models, ML approaches have lately demonstrated comparable levels of accuracy and may even surpass them when describing non-linear interactions. From 2016 to 2020, observations from 242 catchments in Estonia were utilized to train Random Forest (RF) models that can estimate yearly N and P concentrations. The data set included 469 TN observations and 470 TP observations. Soil, terrain, land cover, and climatic variables were among the eighty-two considered. The amount of dependent features in the models was reduced using a feature selection technique.

Using the SHAP method, the authors could extract the most valuable predictors. The TN model's R2 value of 0.83 and the TP model's R2 value of 0.52 show they are just as practical as the previous process-based models utilized in the Baltic region. But our models make getting this kind of data more accessible, so they're more useful in domains where process-based approaches can't work due to insufficient input data. Thus, the models facilitate decision-making assistance for regional water management plans by precisely estimating national nutrient losses and representing the spatial diversity of nutrient discharge (Virro et al., 2022).

12. Analytics for Prediction within the Framework of Geospatial Data with Case Studies

Socioeconomic evaluation requires a deep understanding of population distribution. Dasymetric mapping, frequently used, estimates populations at fine-grid scales using primitive administrative models. Due to the diverse data distribution, the training area and estimation domain are different sizes.

Artificial neural networks effectively attenuated scale heterogeneity by detecting gridded components such as digital terrain models, road networks, building footprints, and land use as dependent variables. This was achieved by considering population density as an independent variable. Hong Kong studies from 2016 to 2021 showed many benefits of the

suggested methodology. The root mean square error was reduced by 19.4% compared to existing approaches. Our technique worked better for larger census units, while the pre-trained model accurately projected population at other times. Land usage was helpful in population estimation. When land use data was replaced with random values, measurement accuracy plummeted by almost 89.0%.

Some attributes lost 2.7% to 13.9% of their measurement accuracy. Traditional cities had the highest population decreases between 2016 and 2021, while newly developed regions had the most significant population rise. Median population density decreased while average population density increased as the study progressed, indicating population concentration (Lu and Weng, 2024).

An attempt is made to forecast the Vertical Total Electron Content (VTEC) in central Anatolia, Turkey, utilizing artificial neural networks. The VTEC dataset was supplied with 19 permanent GPS stations by the International Global Navigation Satellite System (IGS) and the Turkish National Permanent GPS Network Active (TUSAGA-Aktif). The coordinates for the research region are as follows: 36.0 degrees north, 42.0 degrees north, 32.6 degrees east, and 37.5 degrees east.

A perceptron Neural Network (NN) that consists of seven input neurons and an extra layer is created to account for oscillations in the ionosphere's Voltage-Transmission-Averaged Current (VTEC). Within the neural network model framework, the TUSAGA-Aktif GPS stations ANMU and KURU are employed. The neural network model's hidden layer, comprised of 41 neurons, demonstrated the lowest Root Mean Square Error (RMSE) across 50 simulation experiments.

The superior performance of NN VTEC is apparent when considering the correlation coefficients, absolute and relative errors, hourly and quarterly GPS VTEC forecasts, and other relevant factors. This article also shows that NN VTEC is higher than the global IRI 2016. When comparing the Total Electron Content (TEC) forecast with the geographical contribution of the station-based GPS network, it is clear that the KURU station aligns more strongly with the proposed Neural Network (NN) model than the ANMU station (Özkan, 2023).

Surface-wave tomography can image Earth's crustal velocity structure and upper mantle. This study proposed utilizing CNN-based Deep Learning (DL) SfNet to generate the vS model from group velocity dispersion curves and the Rayleigh wave phase. Visible surface-wave tomography can assess Earth's crust and upper mantle velocity structure. This article shows how to build the vS model using SfNet, a CNN-based Deep Learning (DL) technique, group velocity dispersion curves, and Rayleigh wave phase. Visible surface-wave tomography can show Earth's upper mantle and crust velocity structure. The Rayleigh wave phase and group velocity dispersion curves illustrate that SfNet, a deep learning method based on Convolutional Neural Networks (CNNs), can generate the vS model. After applying the approaches to a dataset on the Chinese mainland, ChinaVs-DL1.0, a reference velocity model with fewer dispersion anomalies, was created. The DL technique can invert vS models with enormous surface-wave dispersion data due to its accuracy and efficiency (Wang, Song and Li, 2023).

13. Results and Discussion

For the study, resources were gathered from four different datasets related to the municipal region of Salem, India. It was determined that using these datasets, in conjunction with applying machine learning techniques to satellite imageries, was useful. Following data partitioning into a training set comprising 70% of the total and a test set comprising 30% of the data, the machine learning model was trained using the training data. Examining the early stage of vegetation in Salem City using K-means clustering was the primary objective of the research.

The study's findings, represented in Images 1 through 4, indicated that arid regions saw annual expansion, with a particularly noticeable increase between February 2019 and February 2024. Beginning in 2024, vegetation cover increased considerably, coinciding with the drop in vegetation development 2019.

Many different kinds of vegetation were present in this development, including grass, bushes, mixed vegetation, and higher shrubbery. This growth occurred at both medium and low elevations. The amount of vegetation in the Yercuad forest regions was reduced due to a forest fire that broke out on February 25, 2019, according to an examination of data from 2019. The statistical graphs in Figures 5–12 demonstrate that Cluster 3 is related to the regions with the least vegetation cover. This was discovered through the analysis of the data. It has been observed that the level of vegetation cover has been consistently low since 2019, although there is evidence to suggest that it may begin to increase by 2024. Cluster 2, on the other hand, was characterised by a diverse assortment of plant life, with similar plant species dispersed over the landscape among the higher-elevation shrubs.

14. Conclusion

In conclusion, the research that was conducted in Salem, India, utilised techniques from the field of machine learning as well as satellite data analysis to analyse the vegetation dynamics that were present in the metropolitan area. During the research, K-means clustering was utilised to identify patterns and temporal changes in the distribution of vegetation. Based on the findings, it was seen that the number of bare areas rose on an annual basis, with a significant rise occurring between February 2019 and February 2024.

Beginning in 2024, there was a significant increase in the coverage of various plant types, including mixed grassland and shrubby vegetation. This was the case even though vegetable growth had slowed down in 2019. In addition, the analysis highlighted the effects of a forest fire that broke out in February 2019 on particular regions of the Yercuad forest, including the reduction in vegetation. In addition, the statistical analysis revealed that each cluster was defined by its distinct vegetation pattern, with Cluster 3 representing places that were distinguished by the least amount of vegetation coverage.

In conclusion, these findings contribute to the extensive body of information that is already accessible on the dynamics of Salem City's vegetation and highlight the relevance of ongoing management and monitoring programmes to maintain and improve the area's vegetative resources.

Acknowledgments

The authors thank Saveetha School of Engineering and Saveetha Institute of Medical and Technical Sciences (formerly Saveetha University) for providing the necessary infrastructure to make this work successful.

References

[1] Min Chen et al., "Artificial Intelligence and Visual Analytics in Geographical Space and Cyberspace: Research Opportunities and Challenges," *Earth-Science Reviews*, vol. 241, pp. 1-10, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Ruilong Wei et al., "Combining Spatial Response Features and Machine Learning Classifiers for Landslide Susceptibility Mapping," *International Journal of Applied Earth Observation and Geoinformation*, vol. 107, pp. 1-13, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Jenny M. Suter et al., "Comparing Observed and Unobserved Fishing Characteristics in the Drift Gillnet Fishery for Swordfish," *Fisheries Research*, vol. 256, pp. 1-11, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[4] Min Xu et al., "Estimating Estuarine Primary Production Using Satellite Data and Machine Learning," *International Journal of Applied Earth Observation and Geoinformation*, vol. 110, pp. 1-14, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[5] Negar Shabanpour et al., "Integration of Machine Learning Algorithms and GIS-Based Approaches to Cutaneous Leishmaniasis Prevalence Risk Mapping," *International Journal of Applied Earth Observation and Geoinformation*, vol. 112, pp. 1-17, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[6] Wei-chun Zhang et al., "Soil Total and Organic Carbon Mapping and Uncertainty Analysis Using Machine Learning Techniques," *Ecological Indicators*, vol. 143, pp. 1-11, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Johannes H. Uhl et al., "Towards the Automated Large-Scale Reconstruction of Past Road Networks from Historical Maps," *Computers, Environment and Urban Systems*, vol. 94, pp. 1-26, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[8] Duy-Phien Tran, and Yuei-An Liou, "Creating a Spatially Continuous Air Temperature Dataset for Taiwan Using Thermal Remote-Sensing Data and Machine Learning Algorithms," *Ecological Indicators*, vol. 158, pp. 1-23, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[9] Henry Martin et al., "Trackintel: An Open-Source Python Library for Human Mobility Analysis," *Computers, Environment and Urban Systems*, vol. 101, pp. 1-15, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[10] Anne-Sophie Onody et al., "Use of Satellite Imagery to Categorize Vegetation on the French Railway Network (SNCF Réseau)," *Transportation Research Procedia*, vol. 72, pp. 1451-1458, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[11] Ishwarya Srikanth, and Madasamy Arockiasamy, "Deterioration Models for Prediction of Remaining Useful Life of Timber and Concrete Bridges: A Review," *Journal of Traffic and Transportation Engineering (English Edition)*, vol. 7, no. 2. pp. 152-173, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[12] Deepti Yadav et al., "Identification of Most Useful Spectral Ranges in Improvement of Target Detection Using Hyperspectral Data," *Egyptian Journal of Remote Sensing and Space Science*, vol. 22, no. 3, pp. 347-357, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[13] F.R. López Serrano et al., "Artificial Intelligence-Based Software (AID-FOREST) for Tree Detection: A New Framework for Fast and Accurate Forest Inventorying Using LiDAR Point Clouds," *International Journal of Applied Earth Observation and Geoinformation*, vol. 113, pp. 1-20, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[14] Fareed Majeed et al., "A Novel Artificial Intelligence Approach for Regolith Geochemical Grade Prediction Using Multivariate Adaptive Regression Splines," *Geosystems and Geoenvironment*, vol. 1, pp. 1-16, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] J. Sánchez-Morales, E. Pardo-Igúzquiza, and F.J. Rodríguez-Tovar, "Terrain Methods on Spectral Analysis for Paleoclimate Interpretations: A Novel Visualization Technique Using Python," *Computers & Geosciences*, vol. 175, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[16] Rita Beigaitė, Jesse Read, and Indrė Žliobaitė, "Multi-Output Regression with Structurally Incomplete Target Labels: A Case Study of Modelling Global Vegetation Cover," *Ecological Informatics*, vol. 72, pp. 1-10, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Emily J. Yang et al., "Machine Learning to Support Citizen Science in Urban Environmental Management," *Heliyon*, vol. 9, no. 12, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Mukesh Singh Boori et al., "Machine Learning for Yield Prediction in Fergana Valley, Central Asia," *Journal of the Saudi Society of Agricultural Sciences*, vol. 22, no. 2, pp. 107-120, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[19] Maarten J. van Strien, and Adrienne Grêt-Regamey, "Unsupervised Deep Learning of Landscape Typologies from Remote Sensing Images and other Continuous Spatial Data," *Environmental Modelling and Software*, vol. 155, vol. 1-12, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[20] Nima Kianfar et al., "Spatio-Temporal Modeling of COVID-19 Prevalence and Mortality Using Artificial Neural Network Algorithms," *Spatial and Spatio-temporal Epidemiology*, vol. 40, pp. 1-16, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[21] Yilin Wang et al., "Estimating the Temporal and Spatial Distribution and Threats of Bisphenol A in Temperate Lakes Using Machine Learning Models," *Ecotoxicology and Environmental Safety*, vol. 269, pp. 1-10, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[22] Holger Virro et al., "Random Forest-Based Modeling of Stream Nutrients at National Level in a Data-Scarce Region," *Science of the Total Environment*, vol. 840, pp. 1-19, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[23] Weipeng Lu, and Qihao Weng, "An ANN-Based Method C Population Dasymetric Mapping to Avoid the Scale Heterogeneity: A Case Study in Hong Kong, 2016–2021," *Computers, Environment and Urban Systems*, vol. 108, pp. 1-13, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[24] Ali Özkan et al., "An Artificial Neural Network Model in Predicting VTEC over Central Anatolia in Turkey," *Geodesy and Geodynamics*, vol. 14, no. 2, pp. 130-142, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[25] Feiyi Wang, Xiaodong Song, and Mengkui Li, "A Deep-Learning-Based Approach for Seismic Surface-Wave Dispersion Inversion (SfNet) with Application to the Chinese Mainland," *Earthquake Science*, vol. 36, no. 2, pp. 147-168, 2023. [CrossRef] [Google Scholar] [Publisher Link]