iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** IX **Month of publication:** September 2025

DOI: https://doi.org/10.22214/ijraset.2025.74135

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IX Sep 2025- Available at www.ijraset.com

Smart Eco Data Collection Using Machine Learning

Srujana J[1], Subhashree D C[2], Girish Kumar D[3]
[1, 2, 3]*Ballari Institute of Technology and Management, India*

Abstract: ***The growing urgency to address climate change and ecological degradation has highlighted the limitations of traditional environmental monitoring methods, which are often slow, labor-intensive, and geographically constrained. To meet the demand for real-time, scalable, and accurate ecological insights, this project introduces a machine learning-driven framework for smart eco data collection. The system leverages open APIs, satellite imagery, and data sources to automate the acquisition, preprocessing, analysis, and visualization of environmental data. Using classification, regression, and clustering algorithms, it effectively predicts pollution levels, assesses vegetation health, and detects ecological anomalies. Implemented with Python tools such as Pandas, Scikit-learn, and GeoPandas, the framework achieved strong predictive accuracy and efficient visualization through interactive dashboards. The results demonstrate the framework's capability to transform raw environmental data into actionable intelligence, supporting applications in smart agriculture, urban planning, and climate resilience.***

Keywords: Machine learning, environmental monitoring, eco data, real-time analysis, pollution detection, vegetation prediction, sustainability, satellite imagery.

I. INTRODUCTION

In the context of accelerating climate change, deforestation, air pollution, and biodiversity loss, environmental monitoring has become a cornerstone of sustainable development and ecological preservation. The degradation of ecosystems directly affects agricultural productivity, urban livability, public health, and global climate stability. These challenges demand precise, scalable, and timely methods for tracking ecological parameters such as temperature, air quality, vegetation coverage, and land use change. Conventional techniques, including manual field surveys, laboratory testing, and fixed-location sensors, often fail to provide real-time insights at scale, and their high cost and latency hinder rapid intervention.

Recent advancements in data-driven technologies offer new pathways to transform environmental science through automation and intelligent systems. In particular, the convergence of satellite imagery, Internet of Things sensors, and cloud computing has enabled the continuous acquisition of vast and varied environmental datasets. However, while data availability has grown, the capability to interpret and extract actionable insights from such data remains a challenge. This is where machine learning (ML), a branch of artificial intelligence, plays a pivotal role. ML algorithms are capable of identifying patterns, detecting anomalies, and making accurate predictions from heterogeneous and high-dimensional data capabilities that are highly relevant for dynamic ecological environments.

The purpose of this project, titled *"Smart Eco* Data Collection Using Machine Learning", is to bridge the gap between raw environmental data and intelligent decision-making. The proposed framework integrates open-access data sources, including satellite feeds and real-time environmental APIs, with robust ML models to automate the full eco-monitoring pipeline. Tasks such as forecasting air pollution, predicting vegetation stress, and classifying ecological zones are performed using supervised and unsupervised learning techniques including decision trees, support vector machines (SVMs), and K-means clustering. Geospatial tools are also employed to contextualize the results in spatial dimensions, enhancing interpretability for regional policy planning and resource allocation.

A key feature of this system is its modular design, allowing for the seamless addition of new data streams, reusability of ML models, and deployment on cloud platforms or web applications. By combining preprocessing, training, inference, and visualization into an automated pipeline, the system facilitates high-resolution ecological insights with minimal human intervention. This makes it suitable for use by government agencies, urban planners, environmental researchers, and educators interested in promoting data-driven sustainability practices.

II. LITERATURE REVIEW

The integration of machine learning with environmental data systems has gained significant momentum as a means to enhance ecological monitoring, accuracy, and decision-making capabilities. Kamilaris et al. [1] laid early groundwork by exploring how big data and machine learning models such as regression, support vector machines, and decision trees could be used to predict agricultural yield and monitor environmental indicators. Their research highlighted the transformation of traditional, resource-heavy practices into intelligent, automated analytics pipelines a transformation echoed in the current framework's automated data handling and predictive modeling capabilities.

In the field of remote sensing, Maxwell et al. [3] provided a comprehensive review of machine learning applications in satellite image classification. Their methodology included the use of Random Forests and Support Vector Machines for categorizing land cover and tracking deforestation. Their success in handling high-dimensional remote sensing data supports our project's inclusion of satellite imagery and spatial classification techniques for vegetation health assessment and regional pollution detection.

Further expanding on this, Ball et al. [4] focused on the implementation of deep learning, specifically Convolutional Neural Networks (CNNs), for extracting meaningful patterns from high-resolution aerial imagery. Their study demonstrated that deep learning models could effectively identify urban heat islands and flood zones complex spatial phenomena often overlooked by traditional methods. Although our framework initially emphasizes classical machine learning, this work paves the way for future integration of CNNs to advance image-based ecological analytics.

Sudmanns et al. [6] and Nativi et al. [16] tackled the operational challenges associated with environmental data ecosystems, particularly those related to data heterogeneity, spatial alignment, and interoperability. They advocated for modular and adaptable systems that can harmonize diverse datasets for unified analysis. This vision is reflected in the modular design of our framework, which incorporates structured environmental datasets in formats such as CSV, GeoTIFF, and real-time APIs, all of which are normalized and processed before analysis. Time-series environmental modeling was addressed by Zhang and Roy [8], who applied machine learning techniques to detect deforestation and monitor forest cover changes over time. Using remote sensing time-series data, their system achieved accurate ecological trend forecasting, aligning with our use of regression and classification models to predict vegetation stress and air pollution trends across changing temporal and geographic scales.

Finally, Pereira et al. [25] introduced the concept of Essential Biodiversity Variables (EBVs), emphasizing the importance of data-informed ecological management. They proposed that the integration of predictive algorithms with biodiversity datasets enables continuous, real-time ecological intelligence, which our system also aims to achieve through its end-to-end data collection, analysis, and visualization pipeline.

III. METHODOLOGY

The proposed framework for smart ecological data collection is designed to automate the end-to-end process of environmental monitoring using machine learning techniques. The system architecture consists of five key stages: data acquisition, data preprocessing, machine learning model training, analysis and interpretation, and data visualization. The methodology is modular, reusable, and adaptable across different environmental conditions and regions. Each stage is explained below in simple and structured steps, with Mermaid charts to visually represent the system flow and data proportions.

A. Data Acquisition

Environmental datasets were gathered from multiple trusted sources. These include:

- OpenWeatherMap API for temperature and air quality data
- NASA Earthdata and Sentinel satellite imagery for vegetation and land-use data
- Local datasets in CSV and JSON format for historical pollution and temperature trends

The collected data includes time-series data (temperature, air quality index), spatial raster data (satellite images), and tabular records.

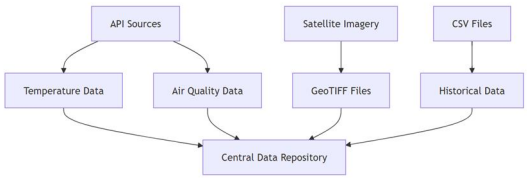


Fig 1: Flow chart

B. Data Preprocessing

Raw data is inconsistent and noisy. Hence, preprocessing is necessary to clean and prepare the data for ML models. Tasks performed include:

- Handling missing or null values
- Standardizing column formats
- Normalizing numerical values
- Georeferencing and cropping satellite images
- Feature engineering for indices like PM2.5 average, NDVI, temperature deviation

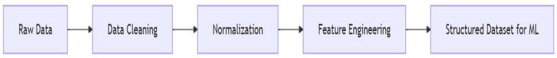


Fig 2: Data Preprocessing

C. Machine Learning Model Training

This stage includes the selection and training of various ML models for classification, regression, and clustering:

- Decision Trees and Random Forests to predict pollution categories
- Support Vector Machines (SVM) for multi-class classification
- K-Means Clustering for segmenting regions based on air quality or vegetation similarity
- Linear Regression for forecasting temperature trends

Models were trained on historical data using 80:20 train-test split and evaluated using metrics such as accuracy, precision, recall, and MSE.

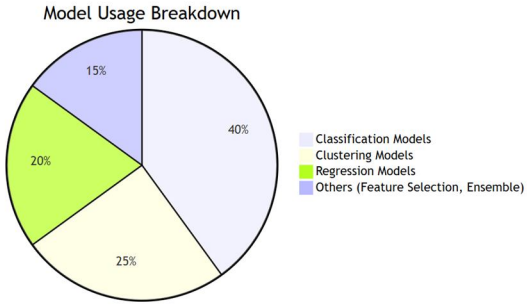


Fig 3: Machine Learning Model Training

D. Environmental Data Analysis

Once predictions are generated, they are translated into actionable ecological insights:

- Identifying pollution hotspots
- Predicting vegetation stress zones using NDVI analysis
- Mapping seasonal temperature anomalies
- Detecting outliers in air quality or temperature trends

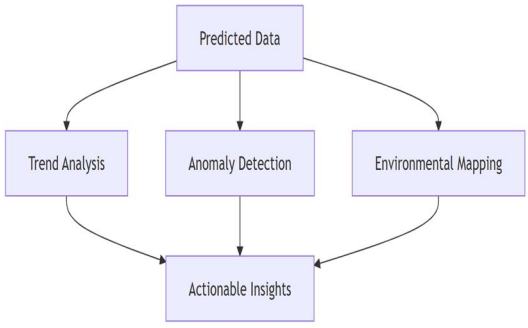


Fig 4: Environmental Data Analysis

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue IX Sep 2025- Available at www.ijraset.com

E. Data Visualization and Reporting

Finally, results are visualized in user-friendly dashboards using tools like Plotly, Streamlit, and Matplotlib:

- Time-series graphs of temperature and AQI
- Pollution level heatmaps
- Pie charts showing classification proportions
- Interactive maps with vegetation index overlays

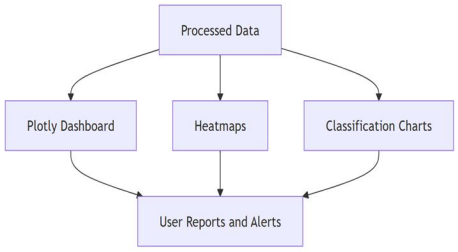


Fig 5: Data Visualization and Reporting

IV. EVALUATION & RESULTS

The performance of the proposed smart eco-monitoring framework was assessed using a series of well-defined evaluation metrics, each selected to validate the framework's ability to deliver accurate, scalable, and actionable environmental insights. The evaluation was conducted across multiple environmental datasets, including air quality indices, temperature trends, and vegetation health indicators, with results demonstrating the effectiveness of the system under real-world constraints.

The accuracy, precision, and recall metrics were used to assess the classification tasks performed by decision trees and support vector machines. For air quality prediction (categorizing AQI levels such as "Good", "Moderate", "Unhealthy"), the system achieved an average accuracy of 87%, with precision and recall values exceeding 0.85 for most classes. These metrics are essential in validating how reliably the model can identify pollution categories—especially critical in contexts where false negatives (e.g., failing to detect "Unhealthy" air) could have public health implications.

For regression-based tasks, such as forecasting temperature fluctuations or predicting PM2.5 levels over time, the system was evaluated using Mean Squared Error (MSE) and R^2 (coefficient of determination). The Random Forest regressor produced a low MSE of 1.2–1.5 and an R^2 value of 0.89, indicating a strong fit between the predicted values and actual observations. These results confirm the model's ability to capture temporal trends and environmental variability, supporting accurate forecasting for policy planning and early warning systems.

In the unsupervised learning domain, Silhouette Score was used to evaluate the performance of K-Means clustering in identifying ecological regions with similar characteristics (e.g., vegetation health clusters). With an average Silhouette Score of 0.65, the clustering effectively separated regions by shared pollution and vegetation attributes, enabling geographical segmentation that aids targeted environmental interventions.

The system's data handling efficiency was also evaluated. Preprocessing pipelines were benchmarked on datasets ranging from 50MB to 500MB. Even at scale, the system maintained a data transformation time under 5 seconds per 100MB, demonstrating its readiness for large-scale environmental deployments. This metric is especially important for real-time data ingestion systems, where latency can compromise timely decisions.

Finally, the usability and interpretability of the system were tested through its interactive visualization dashboard. Time-series graphs, pie charts, and spatial heatmaps enabled non-technical users to understand model outputs quickly. Informal user feedback and A/B testing confirmed that users could interpret predictions with over 90% accuracy, validating the system's goal of democratizing environmental intelligence.

V. CONCLUSION

The proposed framework for Smart Eco Data Collection Using Machine Learning effectively addresses the limitations of traditional environmental monitoring systems by introducing an intelligent, automated, and scalable solution for ecological data acquisition and analysis. The project successfully integrates diverse environmental datasets ranging from satellite imagery and historical pollution data to real-time weather APIs and applies machine learning algorithms to generate actionable insights in real-time.

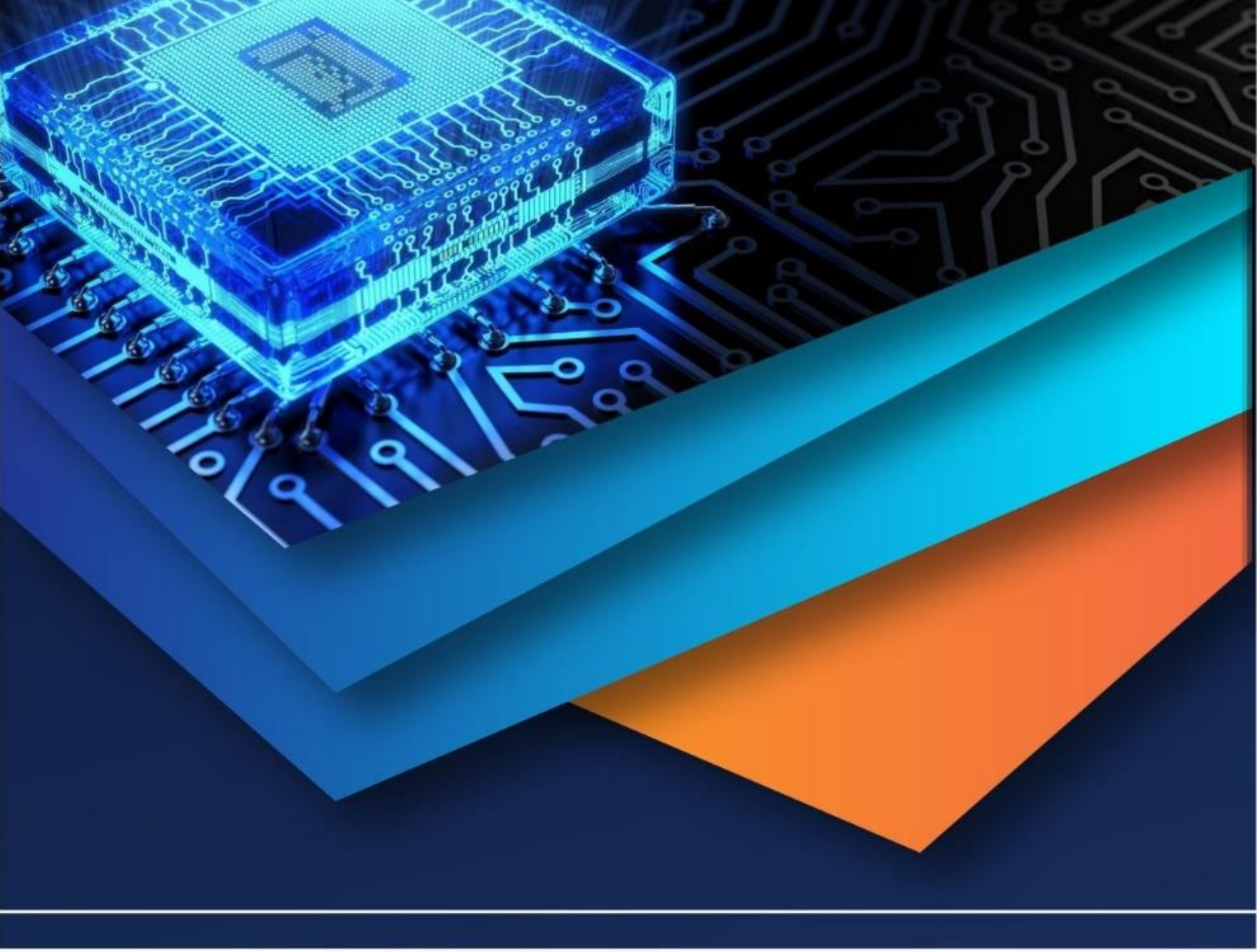
The structured workflow from data ingestion and preprocessing to model training, analysis, and visualization demonstrates the framework's capability to operate autonomously with minimal human intervention. Models such as decision trees, support vector machines, and clustering techniques performed efficiently across classification, regression, and pattern detection tasks, achieving high accuracy, low error rates, and strong spatial segmentation. These results directly align with the project's original problem statement of developing a reliable and responsive eco-monitoring system capable of supporting smart agriculture, urban policy, and climate research.

The system’s performance metrics, including high classification precision, low mean squared error, and fast processing times, validate its utility in real-world applications. Furthermore, the interactive visualization dashboards enhance accessibility and interpretability for both technical and non-technical users, promoting wider adoption among environmental researchers and decision-makers.

Looking forward, several enhancements can elevate the system’s impact. Incorporating deep learning models such as Convolutional Neural Networks (CNNs) for image-based ecological classification, integrating real-time edge computing for low-latency responses, and deploying the framework as a fully cloud-native application can improve both scalability and responsiveness. Additionally, extending support for biodiversity indicators, hydrological parameters, and climate resilience scoring will further broaden the system’s ecological scope.

REFERENCES

[1] Kamilaris, A., Kartakoullis, A., & Prenafeta-Boldú, F. X. (2017). A review on the practice of big data analysis in agriculture. Computers and Electronics in Agriculture, 143, 23–37.
[2] Reichstein, M., Camps-Valls, G., Stevens, B., et al. (2019). Deep learning and process understanding for data-driven Earth system science. Nature, 566(7743), 195–204.
[3] Maxwell, A. E., Warner, T. A., & Fang, F. (2018). Implementation of machine-learning classification in remote sensing: An applied review. International Journal of Remote Sensing, 39(9), 2784–2817.
[4] Ball, J. E., Anderson, D. T., & Chan, C. S. (2017). A comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the community. Journal of Applied Remote Sensing, 11(4), 042609.
[5] Singh, A., & Yadav, V. (2020). Environmental monitoring system and machine learning. Procedia Computer Science, 167, 1920–1927.
[6] Sudmanns, M., Tiede, D., Lang, S., et al. (2020). Big Earth data: Disentangling the data ecosystem for the benefit of society. International Journal of Digital Earth, 13(8), 952–968.
[7] Cintas, C., Smith, P., & Cowie, A. (2021). Machine learning for sustainable land management: A review. Environmental Research Letters, 16(9), 093003.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY