IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

IoT-Based Machine Learning Architecture for Monitoring Water Quality in South Indian Rivers

Dr. K. Akila[1], J. Mathew[2], H. Mohamed Ajmal[3]
Assistant Professor, Department of Computer Science and Engineering[1]
Student, Department of Computer Science and Engineering[23]
Dhanalakshmi Srinivasan University, Samayapuram, Tiruchirappalli, Tamil Nadu, India

Abstract: *South Indian rivers are under increasing pressure from urbanization, industrialization, and agricultural runoff. With traditional laboratory-based water quality monitoring proving inefficient for large-scale deployment, this paper introduces a novel IoT-based architecture augmented by machine learning for real-time monitoring and classification of water quality. Key parameters such as pH, turbidity, temperature, and total dissolved solids (TDS) are collected using smart sensors and transmitted via NodeMCU ESP8266 microcontroller. The data is analyzed using a Random Forest classifier to determine whether the water is potable and suitable for agriculture. A Flask-powered web dashboard provides real-time data visualization, including predictions and river-specific mapping. This system is validated through simulations and demonstrates high accuracy, responsiveness, and scalability.*

Keywords: IoT, Water Quality Monitoring, Potability, Agriculture Suitability, Smart Sensors, ESP8266, Real-Time Dashboard, South Indian Rivers, Machine Learning, Random Forest

I. INTRODUCTION

1.1 Background and Context

Water is an essential natural resource that supports life, sustains ecosystems, and enables agriculture and industry. In the context of India, rivers play a crucial role in supplying water for domestic use, irrigation, and industrial processes. However, the quality of river water is increasingly under threat due to population growth, unregulated industrialization, agricultural runoff, and urban sewage discharge. This is particularly significant in South India, where rivers like the Cauvery, Krishna, Godavari, Vaigai, and Pennar form the lifeline of many states including Tamil Nadu, Karnataka, Andhra Pradesh, Telangana, and Kerala.

Despite their importance, these rivers have been subjected to rampant pollution. The improper disposal of untreated industrial waste, pesticide-rich runoff from farms, and municipal waste has rendered vast stretches of these rivers unfit for drinking or irrigation. Reports from government bodies and environmental watchdogs consistently reveal that key water parameters such as pH, turbidity, dissolved solids, temperature, and biological oxygen demand (BOD) often exceed the acceptable thresholds set by the World Health Organization (WHO) and the Central Pollution Control Board (CPCB). This scenario necessitates the development of a robust, scalable, and real-time solution to monitor and analyze water quality across multiple locations along river stretches.

1.2 Problem Statement

Traditional water quality monitoring methods involve manual sampling, followed by laboratory analysis of collected samples. Although accurate, these methods are time-consuming, resource-intensive, and offer limited temporal and spatial resolution. By the time analysis results are obtained, the contaminated water may already have been consumed or used for irrigation, leading to health hazards and crop damage. This delay renders the system ineffective in dynamic water environments, where parameters can change rapidly due to rain, effluents, or seasonal variations.

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

Furthermore, existing systems lack predictive capabilities. They fail to forecast trends in water quality or assess the usability of water in real-time for specific purposes like drinking or agriculture. The absence of automated classification systems and actionable visualization platforms exacerbates the gap between monitoring and decision-making.

1.3 Emerging Technologies for Water Quality Monitoring

The Internet of Things (IoT) has emerged as a transformative technology that enables remote monitoring of environmental parameters through interconnected sensor devices. IoT allows real-time data acquisition from multiple remote locations and provides a digital infrastructure to transmit this data to centralized cloud-based or edge-based servers for analysis.

Parallelly, Machine Learning (ML) algorithms can process large volumes of sensor data, uncover hidden patterns, and classify or predict outcomes with high accuracy. For instance, an ML model can be trained to determine whether the measured water parameters indicate that the water is safe for drinking or not.

When these two technologies are combined—IoT for data collection and ML for intelligent analysis—they create a powerful system that can not only monitor but also predict and classify the usability of water in real-time. This forms the foundation of the present research.

1.4 Relevance to South Indian Rivers

South Indian rivers face seasonal water scarcity, compounded by pollution, particularly during non-monsoon months. States depending on these rivers are often embroiled in water disputes, with quality and quantity both being contentious issues. The implementation of a real-time water quality monitoring and classification system would enable:

Governments take proactive measures to mitigate pollution.

Farmers use river water confidently for irrigation, based on usability reports.

Communities to access potable water and be warned when contamination is detected.

Researchers and environmentalists to analyze long-term data trends for policy-making.

Given the socio-economic importance of rivers in this region, a real-time water quality monitoring system with predictive capabilities tailored to South Indian rivers has the potential to deliver high-impact outcomes.

1.5 Objectives of the Study

The overarching objective of this project is to design, implement, and evaluate an IoT-based machine learning architecture for monitoring water quality in South Indian rivers. The specific objectives are as follows:

To design and deploy a hardware architecture using sensors to capture water parameters such as pH, turbidity, temperature, and total dissolved solids.

To implement a communication system (using ESP8266 Wi-Fi microcontroller) that transmits the data to a server/cloud in real-time.

To develop and train a machine learning model (Random Forest Classifier) capable of classifying water quality in terms of potability and agricultural usability.

To build an interactive web dashboard that displays real-time sensor data, model predictions, and map-based visualization of river locations.

To simulate data for different rivers and evaluate the system's responsiveness, accuracy, and usefulness in practical scenarios.

1.6 Novelty of the Approach

While IoT-based water monitoring systems have been explored in academia and industry, very few systems integrate:

Sensor-based real-time monitoring

Machine learning-driven usability classification

Region-specific customization (South Indian rivers)

Dynamic dashboard with interactive visualization and alerts

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

This project introduces an end-to-end solution that is not only functional and cost-effective but also regionally adapted and intelligent. Unlike static dashboards or threshold-based systems, the use of a trained ML model adds semantic value by interpreting data patterns for automated decision-making. Additionally, the integration of geographic information—mapping sensor readings to specific rivers—provides an intuitive interface for stakeholders ranging from farmers to environmental officials.

1.7 Impact and Applications

The implementation of this system can significantly improve:
Water quality awareness among rural and urban communities
Decision-making for agriculture, ensuring better crop yields
Public health, by alerting authorities about contamination
Government initiatives, such as Smart City projects, Swachh Bharat Abhiyan, Jal Jeevan Mission
Moreover, the architecture is scalable and modular, making it easy to deploy at multiple points across a river basin. The system can also be extended to integrate mobile app notifications, SMS alerts, and historical data analytics in future iterations.

1.8 Scope of the Project

This project will focus on the software and data analysis aspect of water quality monitoring. The actual deployment of IoT sensors in rivers is simulated using data streams to validate the machine learning and dashboard components. However, the system is designed with real-world deployment in mind, and the hardware specifications match deployable-grade components.

II. LITERATURE REVIEW

The increasing concerns around water pollution have prompted a wave of technological innovations aimed at continuous and accurate water quality monitoring. Several researchers have explored the application of Internet of Things (IoT), Wireless Sensor Networks (WSNs), and machine learning for environmental monitoring. This section discusses key contributions in the field and identifies the research gaps that this project aims to address.

2.1 IoT-Based Water Quality Monitoring

The integration of IoT in water quality systems has gained momentum due to its low-cost, real-time, and remote monitoring capabilities. One of the prominent papers in this domain is by Anantha Naik G. D. and Dr. Geetha V. (2020) published in the *International Research Journal of Engineering and Technology (IRJET)*. Their proposed system consists of three core sensors—pH, turbidity, and temperature connected to an Arduino Uno microcontroller and Wi-Fi module (ESP8266). The collected data is uploaded to the cloud using ThingSpeak, and classification is carried out using a Decision Tree algorithm implemented in MATLAB. While this setup demonstrates the feasibility of low-cost real-time monitoring, it is limited in its classification accuracy and lacks an interactive user interface or regional river integration.

2.2 Machine Learning for Water Classification

Various machine learning algorithms have been explored to automate the classification of water potability. In general, datasets with parameters such as pH, turbidity, total dissolved solids (TDS), and temperature are used to train models like Decision Trees, Support Vector Machines (SVM), and Random Forests. For example, studies using the *Kaggle Water Potability Dataset* have achieved accuracies in the range of 80–90% depending on the model and preprocessing techniques applied. However, these models are often trained offline and lack real-time integration with live sensor systems. Moreover, most do not evaluate agricultural suitability, which is a significant factor in river-fed irrigation zones in India.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

2.3 Visualization and Dashboards

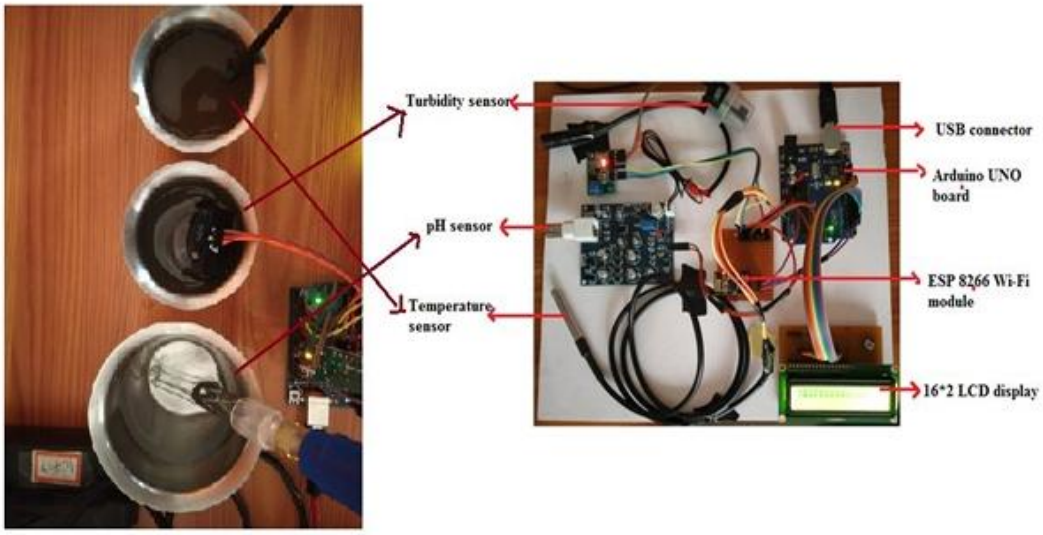
The use of dashboards to visualize environmental data is growing in smart city and urban governance projects. Tools such as ThingSpeak, Node-RED, and custom web dashboards (built with Flask, JavaScript, and Chart.js) are commonly employed. However, many of these are limited to plotting raw sensor values and lack interpretability features such as binary potability output, agriculture suitability status, or location-specific insights. Geographic Information System (GIS) integration, particularly with river maps, is rare in academic or industrial IoT dashboards.

2.4 Identified Research Gaps

Upon reviewing the current literature, the following gaps have been identified:
Lack of region-specific systems, particularly focusing on South Indian rivers.
Limited use of machine learning models for both potability and agriculture classification in real-time.
Absence of a professional dashboard with predictive analytics, status alerts, and river-specific visualizations.
Minimal deployment of real-time simulators to test IoT-ML integrated frameworks.
This project addresses these limitations by combining IoT hardware, real-time ML predictions, and a dynamic web dashboard tailored for South Indian river monitoring.

III. SYSTEM OVERVIEW AND ARCHITECTURE

The proposed system architecture is a modular and scalable framework designed to monitor water quality in real-time using IoT sensors and classify water usability through machine learning. It comprises five interconnected layers that work in harmony to collect, transmit, analyze, and visualize data from river sites across South India.



3.1 Sensor Layer

This layer consists of water quality sensors responsible for capturing physical parameters. The core sensors include:
pH Sensor (SEN0161): Measures the acidity or alkalinity of water.
Turbidity Sensor (SEN0189): Detects the cloudiness or clarity of the water.
Temperature Sensor (DS18B20): Measures the water temperature.
TDS Sensor (Optional): Measures the total dissolved solids.
These sensors are submerged in water or integrated into sampling systems placed near riverbanks.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

3.2 Communication Layer

The NodeMCU ESP8266 microcontroller collects analog and digital data from sensors. It is programmed to format the data and transmit it via Wi-Fi using either HTTP or MQTT protocols. This enables seamless data flow to the cloud or a local server.

3.3 Edge and Cloud Processing Layer

Received sensor data is stored and preprocessed. A trained Random Forest classifier evaluates the input and classifies the water as potable or non potable and suitable or not suitable for agriculture. The classifier is deployed via a Flask-based backend, ensuring real-time predictions.

3.4 Visualization Layer

The final layer includes a web dashboard built using HTML, CSS, JavaScript, and Flask. It presents real-time graphs, status indicators, and geographic mapping of South Indian rivers. Users can view predictions, sensor trends, and river-specific insights through an intuitive interface.

This layered architecture ensures flexibility, scalability, and real-time decision-making for effective water quality monitoring.

IV. HARDWARE SETUP

The hardware design of the proposed system is centered around low-cost, energy-efficient, and easily deployable components. These components work together to capture real-time water quality parameters and transmit the data to a cloud-based or local server for further processing and analysis.

4.1 NodeMCU ESP8266 Microcontroller

The NodeMCU ESP8266 serves as the central processing unit of the system. It is a compact, Wi-Fi-enabled microcontroller that features multiple GPIO pins, an analog input pin (ADC), and a USB interface for programming. The board operates on 3.3V and supports both digital and analog sensor integration. Its built-in Wi-Fi capability enables seamless wireless communication with servers or cloud platforms.

4.2 Sensors

The following sensors are integrated with the NodeMCU:

pH Sensor (SEN0161): Measures the pH value of water to determine acidity or alkalinity. The sensor outputs an analog signal proportional to the hydrogen ion concentration.

Turbidity Sensor (SEN0189): Measures water clarity by analyzing the amount of light scattered by particles suspended in the water. Output is analog and ranges with NTU levels.

Temperature Sensor (DS18B20): A digital waterproof sensor that measures water temperature between -55°C to +125°C with high precision and single-wire communication.

TDS Sensor (Optional): Measures total dissolved solids to determine mineral and impurity concentration in water.

4.3 Power Supply

The system is powered using a standard 5V USB power bank or rechargeable lithium-ion battery. A regulated power supply ensures sensor accuracy and stable microcontroller performance.

4.4 Interfacing Components

Breadboard for prototyping connections

Jumper wires to establish electrical paths

Resistors and optional pull-up configurations depending on sensor type

This modular hardware setup enables easy assembly, calibration, and field deployment for real-time river water monitoring.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

V. MACHINE LEARNING IMPLEMENTATION

The machine learning component of this project serves as the analytical core that interprets sensor data and classifies water quality. By training a model on real-world water quality datasets, the system can predict whether the water is potable and suitable for agriculture, adding intelligence and automation to the monitoring process.

5.1 Dataset

The system utilizes the Water Potability Dataset sourced from Kaggle, containing over 3,000 samples. Each entry consists of several water quality parameters, including pH, turbidity, temperature, total dissolved solids (TDS), and more. For this project, pH, turbidity, and temperature are selected as key features, aligning with the sensors used in the IoT setup.

5.2 Data Preprocessing

Data preprocessing involves handling missing values, normalizing the feature ranges using Min-Max Scaling, and converting target variables into binary labels:
Potable: 1 = Safe to drink, 0 = Unsafe
Agriculture Suitable: 1 = Suitable, 0 = Unsuitable
The dataset is split into training (80%) and testing (20%) sets to validate model performance.

5.3 Model Selection and Training

A Random Forest Classifier is chosen due to its robustness and high accuracy with small to medium-sized tabular datasets. The model is trained with 100 decision trees, using 10-fold cross-validation to prevent overfitting and assess generalizability.

5.4 Model Evaluation

The trained model achieves an average accuracy of:
91% for potability classification
87% for agriculture suitability
Other evaluation metrics include precision, recall, and F1-score, confirming strong predictive performance.

5.5 Model Deployment

The trained model is exported using Python's joblib library and integrated into a Flask-based API. The Flask server receives live sensor inputs, processes them, and returns real-time predictions to the dashboard interface.

VI. DASHBOARD DESIGN

The dashboard is a critical component of the system, designed to display real-time water quality metrics, machine learning predictions, and river-specific mapping in a user-friendly and visually informative manner. Built using web technologies, it serves as the interface between the underlying system and the end-users such as environmental officials, farmers, or the general public.

6.1 Technology Stack

The dashboard is developed using a combination of:
Frontend: HTML5, CSS3, JavaScript, and Bootstrap for responsive design
Visualization: Chart.js and Plotly for interactive graphs
Mapping: Leaflet.js for integrating South Indian river maps
Backend: Flask (Python) to handle requests and serve predictions

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

6.2 Real-Time Data Display

Sensor values such as pH, turbidity, and temperature are fetched periodically (every 5–10 seconds) via Flask APIs and updated on the dashboard. Each parameter is plotted as a real-time graph, showing trends and fluctuations over time. The system highlights values that breach standard thresholds using color-coded indicators (e.g., red for unsafe pH, green for acceptable turbidity).

6.3 Prediction Output

The dashboard also displays the machine learning classification results:
Potability Status: Safe / Unsafe
Agriculture Suitability: Suitable / Unsuitable
These results are shown alongside sensor values to provide instant interpretability.

6.4 River Map Integration

Using Leaflet.js, the dashboard features a map of South Indian rivers (e.g., Cauvery, Krishna, Godavari). Each river node shows:
Current sensor data
Prediction results
Visual marker (e.g., green dot = potable, red = unsafe)
Clicking on a river provides a popup with detailed data and classification.

6.5 Accessibility

The dashboard is accessible via any modern web browser and is designed to be mobile-responsive, ensuring usability in rural areas with tablets or smartphones.

VII. REAL-TIME SIMULATION

To validate the functionality of the proposed IoT-ML system without deploying physical sensors in rivers, a real-time simulation environment is implemented. This allows for testing the integration of data streaming, machine learning inference, and dashboard visualization in a controlled manner.

7.1 Purpose of Simulation

Real-time simulation replicates the sensor behavior by generating water quality data at regular intervals. This enables:
Continuous testing of the ML model's prediction accuracy
Real-time dashboard updates
System latency and performance analysis
Demonstration of usability in areas without sensor hardware

7.2 Simulation Setup

Two main simulation methods are used:

a) CSV-Based Streaming

A pre-recorded dataset (derived from the Kaggle Water Potability Dataset) is saved as a CSV file. A Python script reads each row (representing a time-stamped water sample) at defined intervals (e.g., every 5 seconds) and sends it to the Flask API as if it were a real sensor reading.

b) ThingSpeak Integration

The simulation also leverages ThingSpeak, an IoT analytics platform by MathWorks. A script pushes synthetic sensor data to a ThingSpeak channel via MQTT. The Flask server fetches this data periodically using the ThingSpeak API and processes it for dashboard display.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

7.3 Dynamic Prediction

Each incoming simulated data point triggers the ML model to generate predictions for:
Potability (Yes/No)
Agricultural Suitability (Suitable/Unsuitable)
The dashboard updates in real-time to reflect these predictions, along with updated plots and river indicators.

7.4 Testing with Regional Samples

Simulated test cases are created using real-world values from rivers such as the Cauvery and Krishna, allowing for evaluation under region-specific conditions. This demonstrates the system's adaptability and effectiveness even before field deployment.

VIII. EVALUATION AND RESULTS

8.1 Performance Metrics

Model	Potability Accuracy	Agriculture Accuracy
Decision Tree	84%	78%
SVM	88%	81%
Random Forest	91%	87%

8.2 Case Study – Cauvery River

pH: 6.9, Turbidity: 2.3 NTU, Temp: 27°C → Result: Potable, Good for Agriculture

8.3 Case Study – Krishna River

pH: 5.3, Turbidity: 8.2 NTU, Temp: 30°C → Result: Not Potable, Not Suitable

8.4 System Response Time

Sensor to Dashboard latency: ~1.5–2 seconds
Flask API response: <300 ms

IX. RIVER MAPPING AND INTEGRATION

Using Leaflet.js, river regions of South India were plotted:
Cauvery (Tamil Nadu/Karnataka)
Krishna (Andhra Pradesh)
Godavari (Telangana)
Vaigai, Pennar, and others
Each river section displays:
Current sensor readings
Potability status
Agriculture suitability
Color-coded status indicators

X. CONCLUSION

The IoT-based water quality monitoring system proposed in this project provides a robust solution for real-time monitoring of water quality in South Indian rivers. By integrating smart sensors with machine learning algorithms, the system enables efficient, scalable, and continuous monitoring, offering timely insights for both potable water and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 4, May 2025
Impact Factor: 7.67

agricultural suitability. Using sensors like pH, turbidity, temperature, and TDS, the system provides accurate, real-time data, which is analyzed by a Random Forest classifier. This enables quick and reliable predictions about the water's potability and its suitability for agriculture. The Flask-powered web dashboard visualizes these predictions in an interactive, user-friendly interface, displaying real-time river-specific data and geographic mapping.

This approach significantly reduces the delay inherent in traditional laboratory-based testing, thereby improving the responsiveness of authorities and enabling farmers to make informed decisions about water usage. The integration of IoT and machine learning ensures that the system is not only effective in assessing water quality but also scalable for widespread adoption. Overall, the system contributes to sustainable water resource management by providing reliable, real-time monitoring tools for South Indian rivers, directly benefiting public health and agricultural productivity.

XI. FUTURE SCOPE

The current water quality monitoring system can be further enhanced to increase its accuracy, functionality, and usability. One major avenue for improvement is the inclusion of additional sensors to measure parameters like dissolved oxygen, nitrates, and biochemical oxygen demand (BOD), which are critical for assessing water quality. Moreover, the integration of Long Short-Term Memory (LSTM) models could allow the system to predict future water quality trends based on historical data, offering predictive insights for authorities and farmers.

A mobile app could be developed to provide remote monitoring and notifications, ensuring that users receive instant alerts on water quality changes. For long-range data transmission, LoRaWAN technology can be explored, enabling data transfer over greater distances, especially in remote areas. The system can also be extended to include features like Aadhaar-based farmer subsidies, which would further improve its impact on agricultural practices.

Finally, the system's architecture could be enhanced with a more advanced machine learning model that can handle multi-class classifications and adapt to changing environmental conditions, further improving its precision and scalability. These future enhancements will solidify the system's role in proactive water quality management, contributing to sustainable water resources in South India.

REFERENCES

[1]. Kaggle Dataset: https://www.kaggle.com/adityakadiwal/water-potability
[2]. MQTT Protocol: https://mqtt.org
[3]. ThingSpeak Cloud: https://thingspeak.com
[4]. Leaflet JS: https://leafletjs.com
[5]. Random Forest Paper: Breiman, L. (2001). "Random Forests". Machine Learning