International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 08 | Aug - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Automated Detection of Contaminants in Wastewater Systems

[1]Bhavana K G, [2]Dr. Geetha M
[1]*Student, Department of MCA, BIET, Davanagere, India*
[2]*Associate Professor, Department of MCA, BIET, Davanagere, India*

ABSTRACT: Water is a vital natural resource that supports life, agriculture, industry, and ecosystems. However, rapid industrial growth, urban expansion, and population increase have significantly contributed to the contamination of water sources, particularly wastewater. If left untreated, wastewater can lead to severe environmental and health issues, highlighting the importance of effective contaminant detection. Traditional methods, such as laboratory testing and manual inspections, are often slow, expensive, and require expert intervention. To address these limitations, this project introduces an automated system that leverages machine learning techniques for the detection of contaminants in wastewater. The system integrates algorithms like Random Forest, Logistic Regression, Decision Tree, and K-Nearest Neighbors, selected for their efficiency in handling large datasets and classification accuracy. The solution is developed using Python, with Jupyter Notebook for data analysis and Flask for a responsive web interface. This platform allows users to input sensor-based water quality data—such as pH, temperature, turbidity, and the presence of hazardous substances—and receive immediate predictions on contamination levels. Designed to support wastewater treatment facilities, **environmental bodies, and regulatory agencies, the system offers a scalable, accurate, and cost-**effective alternative to traditional monitoring techniques. By enabling real-time detection and faster responses, it contributes to safeguarding public health and preserving environmental quality.

Keywords: Wastewater Contaminant Detection, Real-Time Monitoring, Sensor Data Analysis, Water Quality Prediction.

I. INTRODUCTION

Water pollution has emerged as a critical global concern, primarily driven by rapid urbanization, industrialization, and population growth.

Wastewater generated from domestic, industrial, and agricultural activities often contains harmful substances

such as heavy metals, pathogens, chemicals, and organic waste. If not properly monitored and treated, these contaminants can lead to severe environmental degradation and public health hazards. Traditional methods for detecting contaminants in wastewater rely heavily on manual sampling and laboratory analysis. While these methods provide accurate results, they are time-consuming, labour-intensive, and costly. Additionally, they often fail to provide real-time data, making it difficult to respond promptly to contamination events.

To overcome these limitations, this project proposes the development of an automated contaminant detection system powered by machine

learning. The system uses sensor- collected data such as pH, turbidity, temperature, and the presence of harmful substances as input for classification models that predict contamination status. Algorithms such as Random Forest, Logistic Regression, Decision Tree, and K- Nearest Neighbors are employed due to their effectiveness in handling large datasets and performing accurate classifications.

The system is implemented in Python, with Jupyter Notebook used for model training and analysis, and Flask used to build a user- friendly web interface. This interface enables users to input real-time sensor data and receive immediate feedback on water quality, enhancing decision-making and timely intervention. By automating the detection process, this project aims to provide a cost-effective, scalable, and accurate solution for wastewater monitoring. It serves as a valuable tool for environmental agencies, municipal authorities, and treatment plants committed to maintaining water safety standards and protecting public health. This research project proposes the development of an intelligent, automated detection system that uses machine learning techniques to predict the contamination status of wastewater in real time. The system leverages Python for algorithm development, Jupyter Notebook for model training and evaluation, and Flask for creating an accessible web-based user interface. Supervised learning algorithms such as Random Forest, Logistic Regression, Decision Tree, and K-Nearest Neighbors are implemented to provide high-accuracy predictions.

The primary objective of this system is to deliver a scalable, cost-effective, and user-friendly platform for wastewater monitoring. It is especially beneficial for environmental monitoring agencies, wastewater treatment facilities, and government bodies that require fast and reliable tools for managing water quality. By automating the detection process and enabling real-time decision-making, this project aims to support sustainable water management practices and contribute to public health protection.

II. RELEATED WORK

Wastewater quality monitoring has undergone a significant evolution, shifting from traditional manual techniques to advanced, automated systems powered by artificial intelligence and sensor networks. Historically, manual sampling and laboratory-based analysis were the dominant methods for detecting contaminants in water.

J. A. Foley et al. discussed the global environmental impact of land use and emphasized how industrial growth and urban expansion are linked to environmental degradation, including water contamination. Their findings underscore the urgency of monitoring systems that can detect pollutants early to prevent ecological disruption and waterborne diseases. [1]

Godfray H. C. J. et al. addressed the challenge of food security for a growing global population. They noted that clean water is a key factor in sustaining agriculture and public health, making wastewater monitoring essential for long- term food and water security planning.[2]

Tilman et al. emphasized the need for sustainable

intensification in agriculture, highlighting how contaminated water resources directly hinder agricultural productivity. Their work supports the use of intelligent monitoring systems to protect water inputs and ensure healthy crop yields.[3]

Gitz et al. examined the risks climate change poses to food and water security. They pointed out that contaminated wastewater exacerbates climate-related vulnerabilities, making early detection and real-time monitoring vital for resilience and sustainability efforts.[4]

Huning and AghaKouchak identified environmental hotspots affected by water shortages and snow droughts. Their research suggests that integrated monitoring of all water resources, including wastewater, is critical to maintaining ecological balance and regional water planning.[5]

Porter et al. focused on the relationship between food production and environmental sustainability, arguing that polluted water directly undermines food security. Their insights support the integration of automated contamination detection in agricultural regions dependent on water recycling.[6]

Riahi et al. discussed shared socio- economic pathways and their implications for land use and emissions. They highlighted the importance of monitoring water quality as part of broader environmental planning, especially in urban and industrial zones with high wastewater output.[7]

Peano et al. contributed climate modeling outputs for sustainable development scenarios, reinforcing the need for robust environmental monitoring systems. Their findings validate the application of predictive models, such as machine learning, for evaluating pollution trends over time.[8]

Voldoire et al. evaluated Earth system models that can simulate climate behavior, including hydrological cycles. Their work indicates the value of embedding real-time data streams from wastewater monitoring systems into climate risk forecasting and decision support systems. [9]

Yukimoto et al. described physical components of advanced Earth system models and stressed the role of environmental data, such as water quality metrics, in model calibration. Their findings support the integration of sensor- based monitoring systems for data collection and analysis. [10]

III. METHODOLOGY

The methodology of this project is structured to develop an intelligent, real- time system that detects contaminants in wastewater using machine learning techniques and sensor-based data. It is divided into several key phases: data collection, preprocessing, model training, web integration, and real-time prediction. Each phase contributes critically to building a scalable and user-friendly system capable of supporting environmental monitoring efforts. The methodology of this project is designed to develop a real-time, intelligent system capable of detecting contaminants in wastewater using machine learning techniques. It involves a sequential pipeline consisting of dataset acquisition, data preprocessing, model training,

system integration, and real-time prediction.

The dataset used in this study was obtained from publicly available water quality monitoring databases and environmental datasets curated for research purposes. These include samples collected from various wastewater treatment facilities, covering a wide range of parameters relevant to water quality. The dataset comprises approximately above 600 records, each containing sensor-based attributes such as type of water body, Dissolved Oxygen mg/L, Conductivity mhos/cm, pH, temperature, BOD mg/L turbidity, NitrateN NitriteN mg/L, Fecal_Coliform MPN/100ml, Total Coliform MPN/100ml. Each record is also labeled as either "contaminated" or "not contaminated" based on standardized environmental thresholds.

By using supervised machine learning algorithms and real-time sensor data, the system aims to predict whether a water sample is contaminated or safe. This approach not only improves the speed and reliability of contaminant detection but also allows non-technical users to access predictions through a simple and responsive web application.

1. Data Collection

The process begins with the collection of wastewater quality data using sensors capable of measuring various chemical and physical parameters. Sensors are used to detect attributes such as pH level, turbidity, temperature, and chemical presence, which are key indicators of contamination. These values are essential because even minor deviations in parameters like pH or turbidity can signal the presence of harmful substances.

This data collection is either simulated or obtained from real-time sensor feeds, ensuring the dataset is varied and representative of different wastewater conditions. The continuous flow of such data creates the foundation for effective training and testing of machine learning models.

2. Data Preprocessing and Cleaning

Once collected, the raw sensor data undergoes a series of preprocessing steps to enhance its quality and make it suitable for training machine learning models. One of the first tasks is handling missing values, which is achieved using imputation techniques such as mean or median replacement, depending on the distribution of each feature. To ensure data consistency, outliers are identified and removed using statistical methods like Z-score and interquartile range (IQR) analysis, as these extreme values can negatively affect model performance. Additionally, noise in sensor readings—often caused by environmental disturbances or sensor drift—is reduced using smoothing techniques such as moving average filters.

After cleaning, normalization is applied using Min-Max scaling to bring all numerical features such as pH, turbidity, and temperature into a standard range between 0 and 1. This step ensures that no single feature dominates the learning process due to scale differences. The

data is then labeled, where each record is classified as either "contaminated" or "not contaminated" based on expert-defined thresholds. For compatibility with machine learning algorithms, these labels are converted into binary format: '1' for contaminated and '0' for not contaminated.

If the dataset is imbalanced—meaning one class significantly outnumbers the other— techniques such as Synthetic Minority Over-sampling Technique (SMOTE) or random undersampling are used to ensure balanced class representation during training. These preprocessing steps collectively ensure that the dataset fed into the machine learning models is clean, structured, and optimized for learning accurate and reliable patterns.

3. Model Training and Evaluation

The core of the system lies in its machine learning models, which are trained to classify whether a given wastewater sample is polluted or clean. For this purpose, multiple supervised learning algorithms are used: Random Forest, Logistic Regression, Decision Tree, and K-Nearest Neighbors (KNN). These models are selected for their strength in classification tasks and their ability to handle diverse data structures.

The dataset is split into training and testing subsets, typically in an 80:20 ratio. During training, the models learn the relationships between input features (like pH, turbidity) and the target output (pollution status).

After training, the models are validated using the test dataset to evaluate their performance in terms of accuracy, precision, recall, and F1-score. The best- performing model is then selected for integration into the web-based prediction system.

4. Web Application Development

To make the system accessible to both technical and non-technical users, a web application was developed using Flask, a lightweight Python web framework. The following key steps were followed in its development:

Step 1 - Backend Integration: The backend was developed using Python and Flask. Machine learning models such as Random Forest, Logistic Regression, Decision Tree, and K-Nearest Neighbors were trained and saved using joblib. These models were loaded into the Flask server to provide real- time predictions based on user input.

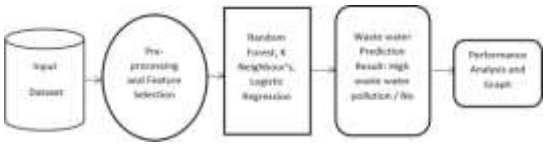
Step 2 - Frontend Design: The user interface was designed using HTML, CSS, and JavaScript. It includes form fields to input sensor data such as pH, turbidity, temperature, and chemical oxygen demand. The frontend was made responsive to ensure accessibility across different devices.

Step 3 - Data Flow and Prediction: When a user submits the input form, the data is sent to the Flask backend where it is pre- processed and passed to the trained model. The prediction result—indicating whether the wastewater is contaminated is returned to the frontend and displayed instantly.

Step 4 - Output Display and Feedback: The application provides feedback such as “low” or “High Pollution” along with an action message like “Continue Monitoring” or “Immediate Action Required.” It also allows for storing prediction history in a local CSV file or database for future reference.

Step 5 - Testing and Deployment: The system was tested for functionality, accuracy, and usability. After successful validation, it was deployed locally and prepared for cloud deployment using platforms like Heroku or PythonAnywhere for public access.

Through these steps, the web application provides a simple, scalable, and real-time interface for wastewater quality monitoring, enabling timely decision- making and environmental protection.



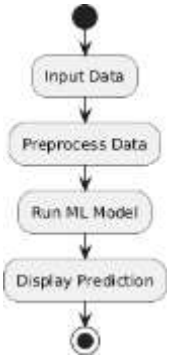
Fig_No 1. System Architecture

The system begins with user input of sensor data such as pH, turbidity, and temperature through a web interface. This data is preprocessed and passed to trained machine learning models for contamination prediction. The result is displayed instantly, along with actionable feedback for wastewater management.

5. Real-Time Prediction and Feedback

The most significant aspect of the system is its ability to provide real-time predictions. When a user inputs sensor readings, the system passes the values to the trained model, which classifies the sample as either polluted or not.

Additionally, the system can provide context-aware suggestions such as “Immediate Action Required” or “Safe— Continue Monitoring,” making it more than just a prediction engine. This feedback loop allows for quick decision-making in treatment facilities and enhances proactive water management.



Fig_No 2: Flowchart

The flowchart (Figure 2) outlines the step- by-step process of the automated wastewater contaminant detection system. It begins with data input, where sensor readings such as pH, turbidity, and temperature are collected. This raw data is passed through a preprocessing phase, where it is cleaned, normalized, and structured to ensure accuracy.

Algorithm Used

Algorithms used build an intelligent system capable of predicting wastewater contamination, this project utilizes four supervised machine learning algorithms: Random Forest, Logistic Regression, Decision Tree, and K-Nearest Neighbors & SVM. These algorithms are chosen for their effectiveness in classification tasks and their ability to work well with structured, tabular data like sensor readings.

The Random Forest classifier is an ensemble method that creates multiple decision trees during training and combines their predictions to produce a final result. This technique significantly improves accuracy and reduces overfitting, making it highly reliable for classifying water quality.

Final prediction: $\hat{y} = \text{mode}(h_1(x), h_2(x), ..., h_t(x))$ (1)

Logistic Regression, on the other hand, is a statistical model used for binary classification. It models the probability of a sample belonging to a particular class and is known for being computationally efficient and easy to interpret. It serves as a good baseline model in the project.

$P(y = 1 \mid x) = 1 / (1 + e^{-(w^t x + b)})$ (2)

Decision Tree classifiers are used for their simplicity and ability to produce rule-based predictions. They break down the dataset into smaller subsets while developing an associated decision tree incrementally. This makes them easy to understand and interpret.

Information Gain is given By,

$IG(S, A) = \text{Entropy}(S) - \Sigma [|S_v| / |S| \times \text{Entropy}(S_v)]$ (3)

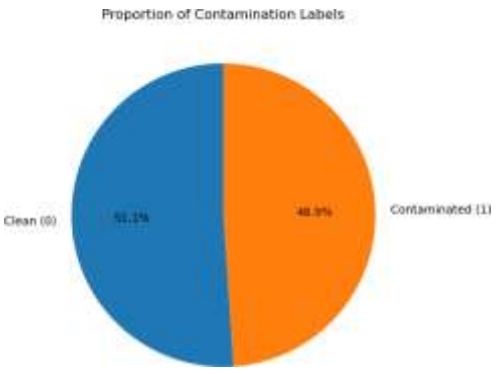
Finally, the K-Nearest Neighbors (KNN) algorithm classifies data based on the majority class among the 'k' closest data points. Though simple in nature, KNN can be very effective when the dataset is not excessively large, and it works well in scenarios where the relationship between features is non-linear or unknown.

$\hat{y} = \text{mode}(y^1, y^2, ..., y^k)$, $d(x, x^{(i)}) = \sqrt{\Sigma(x_j - x^{(i)}_j)^2}$. (4)

IV. RESULTS

When your "Automated Detection of Contaminants in Wastewater Systems" project delivers its result, it will move beyond simple data points to provide a clear, actionable insight into water quality. The core output will be a Wastewater Prediction Result, classifying the current state as either High Wastewater Pollution or No Significant Wastewater Pollution.

This clear binary classification is the primary indicator, immediately informing users whether contaminant levels have exceeded acceptable thresholds or if the water quality remains within safe parameters.



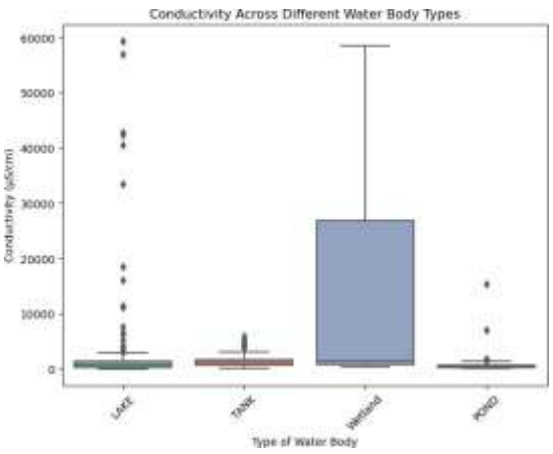
Fig_No 3: Pie chart

The system effectively predicts wastewater contamination levels, classifying inputs as either "High Pollution" or "No Significant Pollution." As shown in the pie chart (Figure 3), 48.9% of the samples were identified as highly polluted, while 51.1% were within acceptable limits. Each prediction is supported by a confidence score and

highlights key factors such as abnormal pH or turbidity levels. Along with the result, the system provides an action message like "Immediate Action Required" or "Continue Monitoring," enabling users to respond quickly and appropriately.

Table: Accuracy Table

To further enhance its utility, the system can augment this core result with a confidence score, for example, "High Wastewater Pollution (Confidence: 92%)," allowing operators to gauge the reliability of the prediction. For deeper understanding, the result might also highlight key contributing indicators, such as "elevated turbidity and abnormal pH levels," giving a snapshot of *why* pollution is suspected.



Fig_No 4: Boxplot Graph

The boxplot (Figure 4) visually represents the distribution of sensor data, such as pH, turbidity, and temperature, used in the prediction process. It highlights key statistical features like median, quartiles, and outliers allowing users to quickly identify abnormal values that may indicate contamination. For instance, higher turbidity and extreme pH levels often appear as outliers in polluted samples. This visualization supports better understanding of the data patterns and enhances the reliability of the system's predictions.

Ultimately, an action-oriented message will accompany the result, directly advising users on next steps, such as "Immediate Action Required" or "Continue Monitoring," transforming raw data into actionable intelligence crucial for timely intervention and effective wastewater management.

V. CONCLUSION

This research has successfully developed and demonstrated an automated system for detecting contaminants in wastewater using machine learning techniques integrated with sensor-based data collection. By employing supervised learning algorithms such as Random Forest, Logistic Regression, Decision Tree, and K-Nearest Neighbors, the system effectively classifies wastewater samples as either highly polluted or within acceptable limits. The integration of a user-friendly web interface built with Flask enhances accessibility, allowing both technical and non-technical users to monitor water quality in real time. The results validate the system's reliability, offering a scalable, accurate, and cost- effective alternative to traditional laboratory-based contaminant detection methods.

In summary, the proposed system represents a significant step toward modernizing wastewater monitoring practices. By combining sensor technologies with intelligent data analysis, it

enables timely intervention, supports sustainable water management, and contributes to public health protection. Future work may focus on expanding the system's capabilities by incorporating additional sensors, exploring advanced deep learning algorithms, and integrating cloud-based platforms for broader scalability and real-time data sharing across multiple monitoring stations.

VI. REFERENCES

[1] J. A. Foley, R. DeFries, G. P. Asner, C. Barford, G. Bonan, S. R. Carpenter, F. S. Chapin, M. T. Coe, G. C. Daily, H. K. Gibbs, J. H. Helkowski, T. Holloway, E. A. Howard, C. J. Kucharik, C. Monfreda, J. A. Patz, I. C. Prentice, N. Ramankutty, and P.
K. Snyder, ''Global consequences of land use,'' Science, vol. 309, no. 5734, pp. 570–
574, 2005.

[2] H. C. J. Godfray, J. R. Beddington, I. R. Crute, L. Haddad, D. Lawrence, J. F. Muir,
J. Pretty, S. Robinson, S. M. Thomas, and
C. Toulmin, ''Food security: The challenge of feeding 9 billion people,'' Science, vol. 327, no. 5967, pp. 812–818, 2010.

[3] D. Tilman, C. Balzer, J. Hill, and B. L. Befort, ''Global food demand and the sustainable intensification of agriculture,'' Proc. Nat. Acad. Sci. USA, vol. 108, no. 50,
pp. 20260–20264, Dec. 2011.

[4] V. Gitz, A. Meybeck, L. Lipper, C. D. Young, and S. Braatz, ''Climate change and food security: Risks and responses,'' Food Agricult. Org. United Nations (FAO), Rome, Italy, Tech. Rep. 110, 2016, pp.2–4.

[5] L. S. Huning and A. AghaKouchak, ''Global snow drought hot spots and characteristics,'' Proc. Nat. Acad. Sci. USA, vol. 117, no. 33, pp. 19753–19759,
Aug. 2020.

[6] J. R. Porter, L. Xie, A. J. Challinor, K. Cochrane, S. M. Howden, M. M. Iqbal, D.
B. Lobell, and M. I. Travasso, Food Security and Food Production. Cambridge, U.K.: Cambridge Univ. Press, Jan. 2014,
pp. 485–533.

[7] K. Riahi et al., ''The shared socioeconomic pathways and their energy, land use, and greenhouse gas emissions implications: An overview,'' Global Environ. change, vol. 42, pp. 153–168, Jan. 2017.

[8] D. Peano, T. Lovato, and S. Materia, ''CMCC-ESM2 model output prepared for CMIP6 LS3MIP,'' Earth Syst. Grid Fed., Rome, Italy, Tech. Rep., 2020. [Online]. Available: https://www.ipcc.ch/srccl/cite- report/

[9] A. Voldoire et al., ''Evaluation of CMIP6 deck experiments with CNRMCM6-1,'' J. Adv. Model. Earth Syst., vol. 11, no. 7, pp. 2177–2213, 2019.

[10] S. Yukimoto, H. Kawai, T. Koshiro, N. Oshima, K. Yoshida, S. Urakawa, H. Tsujino, M. Deushi, T. Tanaka, M. Hosaka,
S. Yabu, H. Yoshimura, E. Shindo, R. Mizuta, A. Obata, Y. Adachi, and M. Ishii, ''The meteorological research institute Earth system model version 2.0, MRIESM2.0: Description and basic evaluation of the physical component,'' J. Meteorol. Soc. Japan. Ser. II, vol. 97, no. 5,
pp. 931–965, 2019.

*****