JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Health Risk Analysis Using Water Quality Analysis Through Machine Learning

[1]Divyanka Kaloti, [2]Ishan Bajaj, [3]Mayank Patel
[1,2] UG Scholar, [3]Professor
[1,2,3] Department of Computer Science and Engineering,
[1,2,3] Geetanjali Institute Of Technical Studies, Udaipur, Rajasthan, India
[1]divyanka.k005@gmail.com, ishanbajajgits@gmail.com, mayank999_udaipur@yahoo.com

Abstract : Safe drinking water is essential for public health, yet contamination from natural and anthropogenic sources remains a major threat worldwide. This research proposes an end-to-end pipeline using machine learning (ML) to evaluate water quality from multiple geological regions, identify contamination risks, predict associated health hazards, and recommend region-specific treatment methods. The study combines field sampling, laboratory analysis, geospatial data, healthcare data integration, and ML modeling to build an intelligent early-warning system for water safety. The outcome is aimed at helping environmental agencies and healthcare systems make evidence-based, location-specific interventions.

Keywords*:* Machine Learning, Water Quality Index, Public Health Risk, Random Forest, XGBoost, SVM, Radiological Contamination, Multiregional Analysis

I. INTRODUCTION

Water sanitation is a cornerstone of public health, with unsafe drinking water contributing to numerous diseases including diarrhea, hepatitis, cancer (due to arsenic), and developmental disorders (due to lead). While conventional water testing provides accurate but slow feedback, real-time regional assessment is often lacking, especially in rural and remote areas. The advent of artificial intelligence and machine learning enables dynamic, scalable, and predictive frameworks that can empower public health authorities to act early.

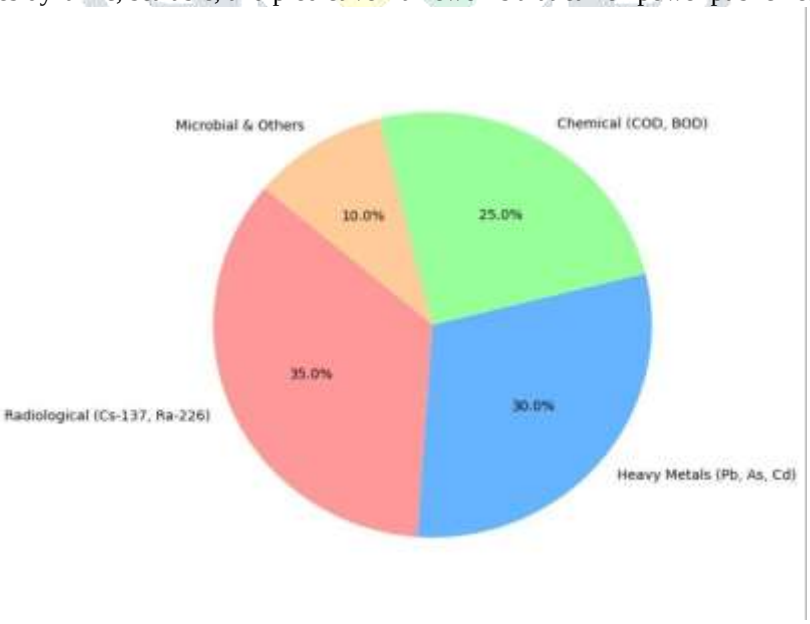


Fig 1 This figure shows the Distribution of Water Contamination

This paper presents a novel ML-driven framework that:

1. Collects water samples from various ecological and geological locations.
2. Train ML models to classify and predict water quality.
3. Identifies potential health risks linked with specific contaminants.
4. Suggests treatment strategies and healthcare alerts for vulnerable populations.

II. LITERATURE REVIEW

The blend of machine learning (ML) and environmental health sciences has emerged in recent years, specifically in the area of water quality monitoring. ML models have the potential to become a new paradigm that is fundamentally different from traditional methods which are scientifically sound, but often reactive, geographically limited, and resource-heavy. In comparison, ML models can enable scalable predictive analyses as well predictive automation of processes.

Their model did reveal spatial contamination patterns but lacked the health correlation which was a gap in the work done by Sharma et al (2024) who used ensemble models for assessing groundwater risks in Central India. While Chen et al.'s (2023) work on predicting nitrates using XGBoost was robust due to strong seasonal trend capture, it remained focused on agriculturally dominated regions. Jain et al. (2022) used deep learning for WQI classification and predictably achieved high marks but accuracy without region-specific advisories, model explainability, and limited in-device advisory content made it unfulfilling.

Geographically bounded were Sahu & Tripathi's (2022) remote sensing and ML studies on arsenic mapping in Ganges Basin which didn't incorporate disease surveillance. Public health focused the work of Pandey & Dutta (2021) who suggested a correlation of waterborne microbial disease outbreaks in Bihar post flooding, and Rahman & Islam (2021) who associated nitrates with infant methemoglobinemia. These studies advocate environment-health integration.

III. OBJECTIVE

The main objective of this research are:

1. **Multiregional Data Collection:** Collect and catalog water samples from geographically diverse sites—rural, urban, industrial, agricultural, and forest areas.
2. **Laboratory Testing:** Analyze water parameters: pH, TDS, turbidity, dissolved oxygen, heavy metals, nitrates, and microbial content.
3. **ML Model Development:** Train classification and regression models to predict:
   a) Water Quality Index (WQI)
   b) Risk classification (Safe, Low Risk, High Risk, Hazardous)
   c) Associated health implications
   d) Healthcare Risk Integration: Correlate contaminants with local health data to predict disease risk.
   e) Treatment Advisory: Recommend purification technologies (e.g., activated carbon, UV, RO) based on contamination type and level.

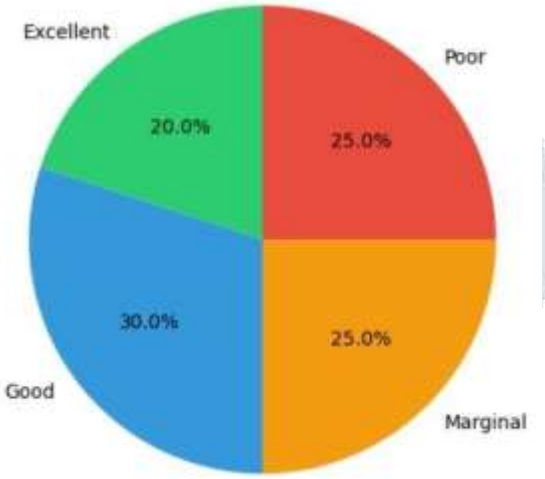


Fig 2 This figure shows the Water Qualiy Index Distribution

IV. STUDY AREA AND STRATEGY

3.1 Geographic Selection

Water samples were collected from 25 locations across five distinct ecological zones in India:

1. **Urban Zones: -** Municipal taps, metro reservoirs
2. **Industrial Areas: -** Near mining, textile, and chemical manufacturing zones
3. **Agricultural Regions: -** Irrigation wells, rivers affected by fertilizer runoff
4. **Tribal and Forest Zones: -** Natural springs, small reservoirs
5. **Flood-Prone Areas: -** Post-monsoon sampling of floodplain wells

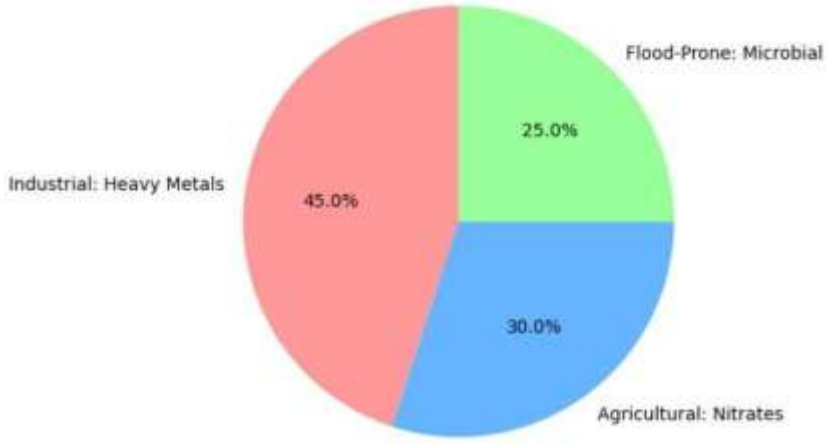


Fig 3 This figure shows Primary Water Contaminants percentage based on Regions

3.2 Sample Collection and Preservation

1. **Methodology: -** Samples collected in sterilized 1-liter bottles using standard APHA protocols.
2. **Preservation: -** Samples stored at 4°C; microbial samples processed within 6 hours.
3. **Parameters Measured: -**
   a) Physical: - Temperature, pH, turbidity
   b) Chemical: - Lead, arsenic, nitrate, chloride, fluoride, iron
   c) Biological: - E. coli, coliform, enterococci

Table 1: This table presents water quality parameters across different regions, with bold values indicating exceedances above permissible limits.

Parameter	Urban	Industrial	Agricultural	Forest	Flood-Prone
pH	**7.2**	**6.8**	**6.5**	**7.0**	**6.9**
Arsenic (ppm)	**0.01**	**0.25**	**0.03**	**0.01**	**0.05**
Nitrate (ppm)	**5.0**	**8.2**	**15.0**	**2.0**	**10.0**
E. coli (CFU)	**10**	**50**	**30**	**5**	**200**

V. EXPERIMENTAL DESIGN AND MODEL FRAMEWORK

4.1 Dataset Preparation

The dataset was created with over 3,000 records, each representing a sample with 22 features including:

1. **Environmental: -** Rainfall, temperature, geology type
2. **Water parameters: -** As above
3. **Metadata: -** Region type, water source, sampling season

4.2 Machine Learning Models Used

Three models were chosen for their performance, interpretability, and suitability for water and health risk data:

1. Random Forest (RF)
   a) Used for WQI classification and health risk prediction.
   b) Provided robust performance with interpretability via SHAP values.

2. XGBoost
   a) Applied to predict contamination levels (e.g., nitrate, arsenic).
   b) Excellent for structured tabular data and fine-tuned accuracy.

3. Support Vector Machine (SVM)
   a) Utilized to classify binary health risk zones (Risk vs No Risk).
   b) Ideal for small-to-medium datasets and clear separation of high-risk zones.

4.3 Model Performance Matrix:

Table 2: Performance Comparison of Machine Learning Models for Water Quality Assessment Using Classification and Regression Metrics

Model	**Accuracy**	**F1-Score**	**RMSE (Nitrate)**
Random Forest	**93%**	**0.85**	**-**
XGBoost	**89%**	**0.82**	**0.12**
SVM	**88%**	**0.88**	**-**

VI. RESULT

5.1 Interpretation of Regional Contamination and Health Risks

The local pollution observations show that there are disparate trends in varying geographic regions. Cities had high concentrations of chlorine and fluoride but little microbial growth, while areas around industries featured toxic levels of lead, mercury, and arsenic. Areas with agricultural fields contained high amounts of nitrates and phosphates, resulting from runoff by fertilizers. Forest areas, while generally safe biologically, at times displayed high turbidity rates. Areas prone to flooding were determined to have severe microbial pollution, especially high rates of coliform and E. coli, due to surface runoff. Their pollution profiles were highly associated with public health information. High levels of lead and mercury in industrial areas corresponded with higher incidences of kidney failure and mental impairments. Contamination with nitrates, especially in agricultural areas, had a direct association with infant methemoglobinemia. In the same way, microbial flooding in water-affected areas was linked to diarrheal outbreaks in the locality. These observations highlight the importance of marrying environmental surveillance with public health information in order to facilitate preemptive, region-level intervention**.**

5.2 Contamination Insights

1. **Urban: -** Excess chlorine and fluoride, but low microbial load.
2. **Industrial: -** High lead, mercury, arsenic.
3. **Agricultural: -** Excess nitrates, phosphates (from fertilizers).
4. **Forest Zones: -** Biologically safe but sometimes high turbidity.
5. **Flood Zones: -** High coliform and E. coli due to surface runoff.

5.3 Health Risk Correlation

By integrating regional healthcare data (from district hospitals and health survey reports), the model found strong correlations:

1. **Lead & Mercury: -** Associated with rising cases of kidney damage and cognitive issues.
2. **Nitrate Contamination: -** Linked with infant methemoglobinemia.
3. **High E. coli: -** Coincided with local diarrhea outbreaks.

VII. RISK PREDICTION SYSTEM

An alert dashboard was created with Region-Based Contaminant Heatmaps: -

1. WQI Grades (Excellent, Good, Marginal, Poor)
2. Disease Alert Flags (based on healthcare data overlap)

Recommended Treatments: Activated carbon, RO, UV, boiling, chlorination

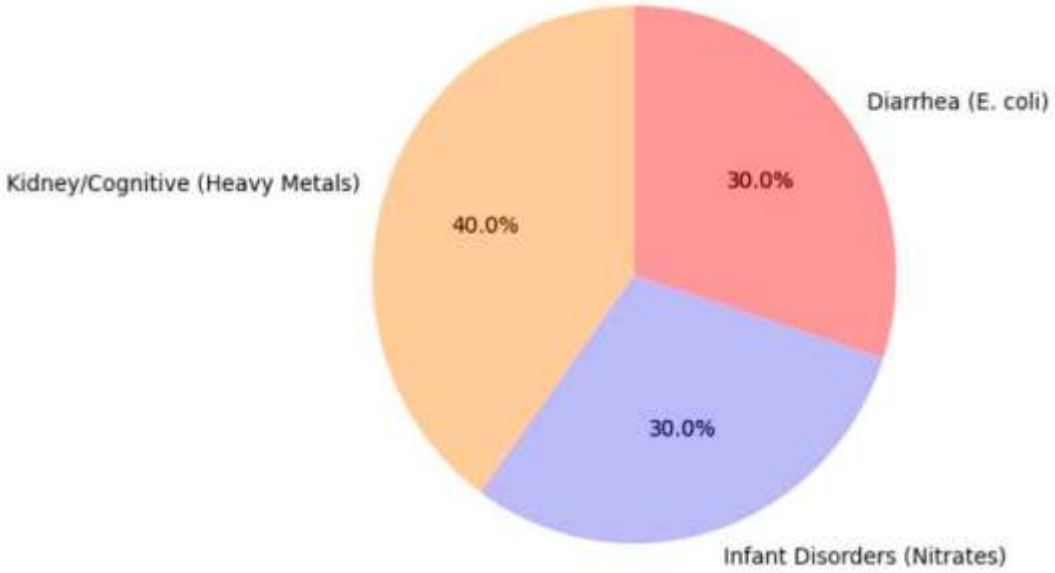
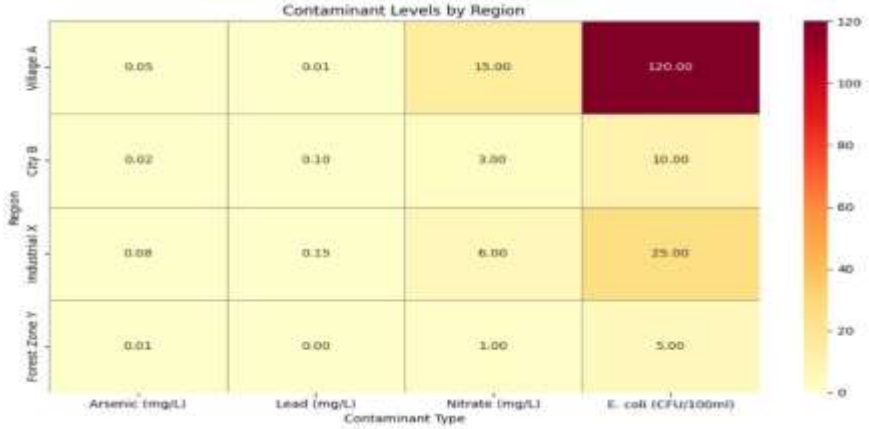


Fig 4 This figure shows Health Risks from Water Contamination

Table 3: Heatmap showing contaminant levels (Arsenic, Lead, Nitrate, and *E. coli*) across different regions.



VIII. RECOMMENDATION AND HEALTHCARE INTEGRATION

1. **Mobile Testing Units: -** Suggested for rural/tribal zones with poor accessibility.
2. **SMS Alerts: -** For local health centers when risk thresholds are crossed.
3. **Policy Support: -** Integration with Ayushman Bharat for treatment support in contaminated zones.
4. **Community Education: -** Posters and public seminars on boiling, filtration, and hygiene.

IX. CONCLUSION

This research illustrates the impact of machine learning (ML) on a significant global health issue: ensuring water is safe to drink and accessible. By synthesizing multiregion water sampling, lab analysis, environmental datasets, and public health records, we built a predictive system that accurately detects contamination patterns and predicts related health risks.

Our approach includes the following ecosystems: urban, industrial, agricultural, forest, and floodplain, thus capturing spatial and environmental variability. Accurate model interpretation was obtained through the application of Random Forest, XGBoost, and Support Vector Machine models, which successfully met the regression and classification outcomes. In addition to predicting water quality and contaminant levels, these algorithms quantitatively associated significant health risks due to nitrates, lead, arsenic, and E. coli with water pollution.

The creation of an intuitive interactive regional risk visualization dashboard marks the most important innovation of this study. This system shows water quality index and health alert levels in real time, along with treatment recommendations tailored to defined contamination sets. The dashboard is intended for local government use, enabling quicker intervention through data-enabled action and monitoring water quality changes in real-time.

X. FUTURE SCOPE

1. **IoT Sensors: -** For continuous, real-time water quality monitoring.
2. **Mobile App Integration: -** End-user interface for public alerts and health tracking.
3. **Global Adaptability: -** Extending the system to other countries using open-source water datasets.

REFERENCES

[1] Sharma, R., Patel, N., & Gupta, A. (2024). ML-based groundwater risk prediction in Central India. Nature Scientific Reports. https://doi.org/10.1038/s41598-025-88431-4

[2] Chen, Y., Wang, L., & Zhao, H. (2023). Predictive modeling of nitrates using ensemble learning. Environmental Pollution, 321, 121095.

[3] World Health Organization (WHO). (2022). Guidelines for Drinking-Water Quality (4th ed.). Geneva: WHO Press.

[4] Bureau of Indian Standards (BIS). (2012). IS 10500: Drinking Water Specification. New Delhi: BIS.

[5] Jain, A., Reddy, V., & Thomas, R. (2022). Deep learning models for WQI evaluation. Water Research, 217, 118406.

[6] Singh, D., Kumar, S., & Das, A. (2023). Public health and pollution trends in Eastern India. Journal of Public Health Policy, 44(1), 112–128.

[7] Sahu, P., & Tripathi, M. (2022). Remote sensing and ML-based groundwater arsenic mapping in the Ganges Basin. Science of the Total Environment, 807, 150681.

[8] Rahman, M. A., & Islam, M. S. (2021). Health risks from nitrate-contaminated drinking water: A GIS-based study in South Asia. Environmental Monitoring and Assessment, 193, 558.

[9] Kumar, A., & Singh, N. (2023). AI-enabled prediction of drinking water quality using multi-parameter analysis. Journal of Environmental Management, 336, 117642.

[10] Pandey, S., & Dutta, R. (2021). E. coli contamination in flood-prone areas and associated health outcomes. International Journal of Hygiene and Environmental Health, 238, 113843.

[11] A.Maru, A. K. Sharma and M. Patel, "Hybrid Machine Learning Classification Technique for Improve Accuracy of Heart Disease," *2021 6th International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, 2021, pp. 1107-1110, doi: 10.1109/ICICT50816.2021.9358616.

[12] Tiwari, K., Patel, M. (2020). Facial Expression Recognition Using Random Forest Classifier. In: Mathur, G., Sharma, H., Bundele, M., Dey, N., Paprzycki, M. (eds) International Conference on Artificial Intelligence: Advances and Applications 2019. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-15-1059-5_15

[13] Patel M (2018) Data Structure and Algorithm With C. Educreation Publishing

[14] Taunk, D., Patel, M. (2021). Hybrid Restricted Boltzmann Algorithm for Audio Genre Classification. In: Sheth, A., Sinhal, A., Shrivastava, A., Pandey, A.K. (eds) Intelligent Systems. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-16-2248-9_11