Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

Drinking Water Quality Prediction Using Machine Learning

Aishwarya K
Dept. of CSE – Data Science ACS College of Engineering, VTU
Bangalore, India
1AH22CD003@acsce.edu.in

Jeevitha C P
Dept. of CSE – Data Science
ACS College of Engineering, VTU
Bangalore, India
1AH22CD024@acsce.edu.in.

Supriya Prakash Tuppad
Dept. of CSE – Data Science
ACS College of Engineering, VTU
Bangalore, India
in H22CD054@acsce.edu.in

Yashodha S M
Dept. of CSE – Data Science ACS College of Engineering, VTU
Bangalore, India
1AH22CD061@acsce.edu.in

Dr. D. Gandhimathi
Associate Professor, Dept. of CSE – Data Science ACS College of Engineering Bangalore, India

Abstract—Water you can drink without harm touches every part of life-health,environment,how towns move forward. Yet dirty runoff grows stronger each year, fed by factory spills, farms sprays seeping through soil, household pipes leaking into streams – testing what flows must now happen faster, everywhere. Lab checks show truth,yes,yet arrive slow, drain money, skip sudden shifts in quality.

Among ways to tackle this, one catches attention- a system using machine learning to predict tap water safety by tracking signs such as acidity, murkiness, dissolved oxygen, conductivity, minerals, nitrates, and hints of gut microbes. Before anything runs, data goes through careful shaping- missing bits fixed, ranges matched, fresh indicators formed, strange readings tossed-aimed all aimed at smoother, clearer forecasts. Not stuck on a single approach, it tried different routes: forests of spilt decisions, step by step improvements engines, and a sharp optimizer named XGBoost tweaking countless dials. Performance came down to how well fits aligned and errors stayed low, checked via R squared scores and mean deviation sizes. In the end, scattered web of branching choices proved toughest, holding firm with better accuracy, less wobble under pressure.

Ahead of the curve, a web tool made with Streamlit activates the adjusted model so users can type in water details and receive fast predictions on safety.Hidden within this system lies an automatic method for monitoring water quality – nimble, light on resources, low-cost to operate. Should the task involve protecting natural ecosystems or connecting rural pipelines, perhaps even sharing updates with urban digital grids- it slips neatly into place.

Keywords – Machine Learning, Random Forest, Data Processing, Streamlit, Environment monitoring

1. INTRODUCTION

Water keeps life going, still many lack safe sources. As towns swell and production rises, pure water slips away. Crops thirst for it, businesses rely on it, humans require it. Rivers turn foul, illness creeps in. Poison travels where clean flow should be.

Out of rivers, water travels in bottles held by collectors heading toward testing rooms. Inside those rooms, microscopes and sensors search for hidden organisms along with pollutants floating in samples. Getting it right means waiting- it takes time since skilled hands must guide every procedure on delicate equipment. When instruments fail or schedules thin out mid season, wait periods stretch without warning. Surprise leaks show up late. Skip the idea of fast warnings – the system crawls, missing splits the moment they strike.

Lately, machines are getting sharper at catching patterns hidden in oceans of data. Not simply crunching digits

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

anymore, they study old reports on actual water quality instead. Thanks to that experience, tools today can hint at contamination well ahead – before anything obvious occurs. Early signals mean faster reactions from communities waiting endlessly for test outcomes isn't always required now.

Out of nowhere, a clever gadget dives into water details using patterns found in numbers. Not only does it scan like pH and particles, but also judges whether sipping is safe. If temperature nudges up, clarity might dip – warnings pop without delay. Built right into web window, the core system runs smooth, no training required. Simple clicks bring sharp answers, especially handy for folks just starting out.

2. LITERATURE SURVEY

Lately, folks began exploring whether machine learning might forecast water quality more effectively. Some studies tried different routes just to speed up progress. A technique built by Mustafa Yurtsever alongside Murat Emeç paired Support Vector Regression with XGBoost - aimed at tap water checks - and delivered solid precision. Together, those systems coped surprisingly well with uneven inputs, hinting that mixing tactics could boost prediction strength.

Peering into machine guesses about rivers, Astha Sharma teamed up with peers to test gear such as Decision Trees, methods tracking close examples, also voting-based forest models along India's waterways. Out of all options, the team-picked tree approach dealt best with knots in data since snarls didn't knock it sideways. Elsewhere, Osim Kumar Pal stacked systems mimicking brains beside math tools tracing lines, plus basic random-chance predictors. The brainy nets often hit closer to truth but sucked down far more energy while running.

Now predictions for changing water quality have improved, due to research led by Juntao Liu. Rather than rely on older techniques, the group used a unique neural network structure - processing data flows in two directions using streamlined recurrent cells. With information moving ahead and behind across layers, subtle trends emerge more clearly. Each point in time carries echoes from what came before and hints of what follows. That dual view sharpens early warnings when environments like aquaculture systems shift unexpectedly. In fast-changing scenarios, this method stood out with consistent accuracy. Water quality forecasts often come out right in studies, still almost none work in real time or welcome user input. Only now is something arriving that gets it correct, free for anyone, running the moment it's needed.

3. EXISTING SYSTEM

These days, testing how clean water is usually involves taking it away to a lab. Whether pulled from rivers, lakes, or straight out of faucets - each sample goes through fixed routines that check chemicals and living matter. It works well enough. Yet problems creep in the moment containers are carried off-site. Findings show up slowly. A wait of several days isn't unusual. In some cases, nearly a month passes. Back at the lab, gear works well - though stuck in one spot. Results might hold up, still they arrive too late when speed matters. Issues spread during delays. Waiting on live choices stretches out. This delay slows reactions more than expected.

Hours tick by while samples move from collection to delivery, then sit untouched until labs respond. Not until machines hum and specialists step in does any progress show - costs rising with every check. Far from city edges, where roads thin and services vanish, help rarely arrives at all. Delays like these expose how slow the usual process really is. Distance becomes a wall when care cannot travel beyond crowded streets.

Something stops progress. Predicting issues before they happen just isn't possible with current methods. Water tests show present conditions only, never hint at future risks. Problems must appear first - only then does anyone notice.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

Late data means reactions always lag behind events. Fixing things early fails when information arrives post-incident.

Still, the present system eats hours while draining funds, especially since updates must flow nonstop - meaning an automatic overhaul could fit better. Yet speed matters most now, because delays pile up fast without real-time feedback built in. Efficiency drops if every step needs watching, even when performance seems fine on paper at first glance.

4. PROPOSED SYSTEM

Out front, a new way to watch water quality uses machines that learn. Not stuck in the past with manual tests, they study what happened before to guess how safe things will be next. All pieces fit into one view, not laid out piece by piece. Answers show quickly - no waiting days for reports. While it seems quiet on the surface, all parts work together under the hood.

4.1 System Architecture

[User Input: Water quality parameters (pH, Turbidity, TDS, etc.)]
↓
Data Intake &Validation (Initial Checks)
↓
Data Cleaning & Preprocessing (Handling Missing /Noise)
↓
Feature Engineering (Transforming Raw Inputs)
↓
Trained Forecasting Model (Learn from historical Data) → Prediction Output (Water quality Status) → Result Interpretation Layer→ Safe / Unsafe Tag Display
↓
User Interface (Visual Indicators + Clear Feedback)
↓
Continuous Monitoring & Updates (Adaptive Improvements Over Time)

Cloudy stuff floats in water alongside tiny living things we cannot see. Its sourness or bitterness shifts depending on what mixes inside. Oxygen swims through it, letting life survive below the surface. Electricity sneaks along when minerals dissolve into streams. Nitrates pile up slowly, often from outside sources creeping in. Bacteria leave faint marks detectable only through close inspection. Each piece tells a story about safety before any tool sees it. Machines learn better once numbers behave predictably. Raw details get smoothed out first, shaped carefully without losing truth. Patterns emerge clearer after quiet adjustments behind the scenes.

One algorithm beats the rest once every option is tested for how well it predicts. Since getting close to real outcomes counts, the best performer shows up clearly in numbers - higher R^2, smaller RMSE. Plugged into a website made with Streamlit, it grabs water measurements at input spots, shoots back forecasted results right away.

Right off the bat, things move quicker once you're in motion. Waste shrinks slowly - so expenses shrink too. Fewer mistakes appear even when oversight fades. Places that used to stay shut now let people through. Out of nowhere, scaling slips into larger workflows without a hitch. As things grow, watching over systems stretches easily into new zones. From your regular smart devices, live updates travel along links that just work. While old tools keep running, everything ties together - no fuss, no noise.

5. METHODOLOGY

5.1 Dataset Collection

The data set has been made from authentic data of water quality, comprising important parameters like pH, turbidity, dissolved solids, etc., which help decide if water is fit for consumption or not. In each case, the values, along with the label indicating their authenticity during the

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

testing of water samples, are provided. This will make sure that the data is authentic in nature.

5.2 Preprocessing and Feature Extraction

The collected dataset is carefully pre-processed to boost model efficiency. Data gaps are intelligently imputed instead of being discarded to minimize data losses. Errors or inconsistent inputs are excluded from the dataset before other operations.

Numeric variables are normalized or standardized to retain consistency in their range, whereas categorical variables—such as water type or water color flags—are encoded for machine understanding. Outliers are statistically detected through methods such as Interquartile Range (IQR), and outliers are eliminated when needed to avoid biasing model predictions.

For improved prediction accuracy, extra features can be created by applying statistical operations, such as mean value and gap of variation.

5.3 Model Architecture and Training

The final dataset is then segregated into train and test sets to ensure unbiased assessment of the results. Various machine learning models have been considered, such as Random Forest Regressor, Gradient Boosting, and XGBoost models that can capture complex non-linearities present in the data.

Model assessment is done by calculating various performance measures, such as R^2 Score, which gives an idea about the variance captured, and RMSE, to estimate how well the model has performed with respect to prediction. Tuning the hyperparameters of these models leads to the best model being identified from performance measures.

5.4 Model Optimization and Selection

In all the models that have been trained, the one with the greatest accuracy and stability is chosen as the ultimate model. Techniques such as gradient boosting and XGBoost tend to yield excellent results because they continuously learn from their mistakes and are capable of handling even distorted data distributions.

5.5 System Integration and Real-Time Inference

The model will be incorporated into a single pipeline in which data is passed from one end to another without any delay. The user-defined values for water will go through the same preprocessing and feature generation techniques as applied during the training process.

Inferences will be generated continuously, meaning there will be no delay between modules. This will ensure high precision since every module delivers its processed information directly to the next module.

5.6 Deployment and User Interaction

Deployment is achieved by means of Streamlit to provide an easy-to-use, browser-based application for the user. People have the ability to enter water quality parameters via a web page and obtain predictions concerning their safety.

The output is provided in a clear manner that can be visually seen, for example, “Safe” or “Unsafe.” This application is scalable and can easily be maintained without any difficulties.

6. REAL-WORLD APPLICATIONS

Water Safety Monitoring:

This system will allow users to quickly analyze water quality in the home or community environment without needing complicated lab testing.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

Support in Rural/Remote Areas:

For those living in places where there are few testing facilities, it will become easier to analyze water quality in simple terms.

Environmental Monitoring:

This system can be used by government bodies or any other organization that needs to monitor water quality.

Integration with Smart Water Management Systems:

The model developed can also be used in IoT based water monitoring systems.

Awareness on Drinking Water Safety:

By transforming difficult to interpret data into understandable predictions and visual information, the awareness regarding drinking water safety can be raised.

7. CONCLUSION

This approach predicts tap water safety by analyzing real-time sensor data with smart algorithms. Rather than waiting for traditional lab work, outcomes appear instantly when conditions shift.

Among online resources, combining smart algorithms with a straightforward web design helps plenty of people access useful insights. Because Random Forest performs so reliably, forecasts on water condition show solid results - proof sits right inside the data. Practice proves it too; actual tests confirm these systems deliver what they promise.

One tiny drop might start something large if someone pays attention. When needs shift, this system keeps moving without slowing down. Safe drinking water remains possible since alerts come ahead of time. Patterns get noticed by machines, which helps humans respond sooner. Problems that were unseen appear before damage grows wider. Over time, progress slips in through constant updates. Wellness shows up when numbers move without blocks.

8. FUTURE WORK

Most folks overlook how sensors keep feeding new data while things move. Yet predictions sharpen dramatically whenever advanced networks handle the heavy lifting. Live details stream in constantly if systems stay hooked to active gadgets. Patterns hidden from regular eyes emerge clearly under deeper scrutiny. Growth fits naturally into place only after backbones adjust for bigger demands. When people use phones to reach tools, more folks get in. Control shifts around when it rides on mobile signals. Where you stand matters less once apps fit in pockets.

REFERENCES

[1] M. rtsever and M. Emec, “Potable Water Quality Prediction Using Artificial Intelligence and Machine Learning Algorithms,” 2023.

[2] Sharma et al., “Water Quality Prediction Using Machine Learning Models,” 2024.

[3] O. K. Pal, “The ality of Drinkable Water Using Machine Learning Techniques,” 2022.

[4] J. Liu et al., “ urate Prediction Scheme of Water Quality in Smart Mariculture,” 2020.

[5] V. S. G. Devi, “Random Forest for Water Quality Prediction,” 2019.