International Journal of Scientific Research and Engineering Development-– Volume X Issue X, Year
Available at www.ijsred.com

RESEARCH ARTICLE OPEN ACCESS

WISE-Guard: An Explainable Multi-Pathway Spatio-Temporal Framework for Water-Related Disease Risk Prediction and Water Treatment Decision Support

smv@mcehassan.ac.in | angelpraises217@gmail.com | ashwinicj256@gmail.com | manishmaniee07@gmail.com | suhasgowda3939@gmail.com | samikshajaind@gmail.com
Sushma M V, Angel Praises A P, Ashwini C J, Manish L, Suhas S L, D Samiksha Jain
Department of Computer Science and Engineering (AI & ML),
Malnad College of Engineering, Hassan, Karnataka, India

ABSTRACT
Water-related diseases remain a persistent public-health challenge, particularly in communities where safe water, sanitation, hygiene infrastructure and timely disease surveillance are limited. These conditions do not arise through a single transmission pathway: waterborne infections are acquired through contaminated water; water-washed diseases are associated with inadequate water for hygiene; water-based infections involve aquatic intermediate hosts; and water-related vector diseases are transmitted by vectors that breed in or near water. At the same time, water treatment decisions depend on the contaminant, water matrix and treatment performance rather than on a generic water-quality score. This paper proposes WISE-Guard, an explainable, multi-pathway decision-support framework that integrates water-quality observations, meteorological and hydrological indicators, sanitation and hygiene proxies, vector and aquatic-host indicators, geospatial relationships, and aggregated disease surveillance. A shared data-validation and representation layer feeds separate prediction heads for the four water-related disease pathways, while a distinct water-assessment module estimates water-quality status and supports evidence-based treatment recommendations. Random Forest and XGBoost provide tabular baselines; a PyTorch spatio-temporal model learns geographical and temporal dependencies. Missingness indicators, source-quality scores, probability calibration and SHAP-based explanations are incorporated to communicate limitations and support human review. The paper defines a reproducible implementation and evaluation protocol using temporal, spatial and combined holdout validation. Since this is a proposed framework, no experimental performance is claimed. The intended contribution is an integrated and auditable architecture that connects disease-risk estimation with pathway-appropriate prevention and water-treatment decision support without treating a model prediction as a diagnosis or a water safety certification.

Index Terms— Water-related diseases, waterborne disease prediction, water quality, water treatment, machine learning, XGBoost, spatio-temporal learning, graph neural networks, explainable AI, public-health surveillance.

 ©IJSRED: All Rights are Reserved

I. INTRODUCTION

Water is connected to human health through drinking, domestic hygiene, food production, recreation and ecological systems. Unsafe water and inadequate sanitation can contribute to infectious disease, while water scarcity can undermine hygiene and create conditions that favour transmission. Environmental changes such as heavy rainfall, flooding, drought, temperature variation and changes in water storage can alter exposure patterns. These relationships are especially important in rural and peri-urban settings where monitoring coverage and healthcare access may be uneven.

The phrase "water-induced diseases" is used broadly in this paper to refer to diseases whose occurrence or transmission is materially associated with water, sanitation, hygiene or aquatic environments. A useful classification distinguishes waterborne, water-washed, water-based and water-related vector-borne pathways [1], [2]. These categories are epidemiologically different. For example, cholera is associated with ingestion of water or food contaminated with Vibrio cholerae; trachoma is linked to inadequate hygiene and water access; schistosomiasis involves freshwater snails as intermediate hosts; and malaria or dengue involves vectors that breed in or around water. Consequently, a single model that treats all these diseases as equivalent labels would obscure their mechanisms and may produce misleading recommendations.

Water-quality monitoring and disease surveillance have advanced through sensor networks, Internet of Things (IoT) platforms and machine learning. IEEE studies have examined sensor-based water monitoring, water-quality index prediction, potability classification and the estimation of microbial contamination [3]–[7]. Recent work has also explored direct waterborne-disease forecasting using temporal learning and spatially connected community surveillance [8], [9]. Separately, artificial intelligence has been investigated for treatment-process prediction and operational optimisation [10]. These strands provide a foundation, but their outputs are not interchangeable: a water-quality prediction is not a pathogen test, and a disease-risk estimate is not a clinical diagnosis.

This paper proposes WISE-Guard (Water Intelligence and Surveillance for Environmental-health Guard), a multi-pathway, explainable framework. It combines common environmental and geographic representations with separate prediction heads for waterborne, water-washed, water-based and water-related vector pathways. A separate water-assessment and treatment-support module maps verified water-quality evidence to candidate treatment processes and operational checks. The design deliberately keeps treatment selection distinct from disease prediction, because treatment must be selected according to the contaminant and validated process performance.

The contributions are: (1) a unified data model for environmental, water, hygiene, vector/host and health-surveillance information; (2) a multi-task prediction architecture with pathway-specific outputs; (3) a confidence-aware missing-data strategy and calibrated risk estimates; (4) an evidence-constrained water-treatment recommendation layer; and (5) a reproducible evaluation protocol that tests temporal forecasting and geographic transfer. The novelty claim is an integration and evaluation design, not the invention of GNNs, LSTMs, XGBoost, SHAP or water treatment.

II. BACKGROUND AND DISEASE TAXONOMY

A. Waterborne pathway: Waterborne diseases are transmitted when pathogens are ingested in contaminated water, although food and person-to-person pathways may also contribute to some outbreaks. Relevant examples include cholera, typhoid fever, hepatitis A and E in appropriate epidemiological settings, shigellosis, giardiasis and cryptosporidiosis. Risk indicators can include faecal contamination, sanitary failures, source protection, turbidity, rainfall, flooding and reported cases. Routine physicochemical sensors cannot directly establish the presence or absence of every pathogen; microbiological sampling and laboratory confirmation remain important.

B. Water-washed pathway: These conditions are associated with insufficient water quantity or access for personal and domestic hygiene. Examples commonly discussed in the water-and-health literature include trachoma, scabies and some faecal–oral infections whose transmission is reduced by improved hygiene. Relevant predictors therefore include continuity and quantity of household water supply, access to handwashing facilities, sanitation, crowding and hygiene-service coverage. A drinking-water disinfection recommendation alone is not an adequate intervention for this pathway.

C. Water-based pathway: These diseases involve an aquatic organism or intermediate host in the pathogen life cycle. Schistosomiasis is a key example involving freshwater snails; dracunculiasis is another historically important example. Exposure may depend on freshwater contact, host habitat suitability, water-body ecology, sanitation and local transmission history. The appropriate response may involve safe-water access, exposure reduction, sanitation and disease-specific control programmes rather than household chlorination alone.

D. Water-related vector pathway: In this group, vectors breed in or near water and transmit pathogens. Examples include malaria, dengue and Japanese encephalitis, with the specific vector ecology differing by disease. Predictors may include rainfall, temperature, humidity, standing-water indicators, land use, vector surveillance and population susceptibility. A water purification module cannot prevent vector transmission; environmental management, vector surveillance and public-health control measures are needed.

 ©IJSRED: All Rights are Reserved

The four-pathway structure follows established public-health classifications [1], [2]. Some diseases may fit more than one environmental category, and local epidemiology determines the relevant pathway. WISE-Guard therefore allows disease-specific labels and multiple pathway indicators rather than forcing every observation into one mutually exclusive class.

III. RELATED WORK AND RESEARCH GAP

A. IoT and water-quality prediction: IEEE research has integrated sensors measuring parameters such as temperature, pH, turbidity and total dissolved solids with cloud or edge analytics. An IoT/ML framework reported the use of LSTM, support-vector regression, multilayer perceptrons and classifiers including SVM, XGBoost, decision trees and Random Forest for water-quality index and class prediction [3]. Another IoT monitoring study used Random Forest, artificial neural networks and LightGBM to analyse sensor measurements and address missing observations [4]. These studies demonstrate the value of continuous measurements but focus primarily on water parameters rather than multiple human disease pathways.

B. Water potability and microbial indicators: IEEE studies have evaluated machine-learning approaches for water potability and microbial contamination. SoftMax-ELM optimised with an adaptive crow-search strategy was evaluated for potability classification [5]. An AI/ML mobile-app study estimated E. coli levels in surface waters using sensor variables and compared SVR, Random Forest and XGBoost approaches [6]. Such models can support screening, but a predicted indicator level should not be represented as a direct pathogen assay or a complete assessment of chemical and microbiological safety.

C. Disease forecasting: Recent IEEE conference work has explored LSTM-based prediction of waterborne-disease risk from historical and water-quality data [8]. Another study describes a rural early-warning architecture using spatio-temporal graph learning and ensemble models [9]. These works are directly relevant and mean that "using AI for waterborne disease prediction" is not, by itself, a novel claim. WISE-Guard instead broadens the task to multiple water-related transmission pathways and explicitly separates disease forecasting from water-treatment decision support.

D. Treatment process intelligence: Research has considered AI for drinking-water treatment operations, including coagulation/flocculation, membrane filtration, adsorption and disinfection-related process management [10]. IEEE work has also described predictive water-treatment monitoring and chemical-use optimisation using plant and SCADA information [11]. Treatment recommendations must nevertheless remain constrained by water chemistry, pathogen class, process validation, operating conditions and local standards.

E. Research gap: The reviewed work indicates an opportunity for a framework that (i) represents distinct water-related disease pathways rather than one generic label; (ii) links environmental exposure, water quality and aggregated health surveillance; (iii) evaluates transfer across time and geography; (iv) communicates uncertainty and missing-data quality; and (v) translates water-quality evidence into treatment-process options without confusing these options with vector or hygiene interventions. This is a synthesis-derived research gap, not proof that no prior system has ever combined any of these components.

IV. PROPOSED WISE-GUARD FRAMEWORK

A. System overview: WISE-Guard contains six logical layers: data acquisition; data governance and preprocessing; feature engineering; pathway-specific disease prediction; water-quality and treatment decision support; and an explainable alert interface. Inputs are aligned to a common spatial unit (for example, village, ward or district) and time interval (for example, week). The chosen resolution must match the reporting quality and intended intervention.

B. Data sources: The framework may ingest (1) water measurements such as pH, turbidity, conductivity, total dissolved solids, temperature, dissolved oxygen and residual disinfectant where available; (2) microbiological observations such as E. coli or thermotolerant coliform results; (3) meteorological and hydrological observations such as rainfall, temperature, humidity, flood occurrence and water level; (4) hygiene and sanitation indicators such as water-service continuity, sanitation coverage and handwashing access; (5) vector and aquatic-host surveillance where relevant; (6) geospatial features such as water-source locations, land cover and connectivity; and (7) aggregated, de-identified disease counts from authorised surveillance sources.

C. Data governance: Health records should be aggregated to an approved geographic/time unit and de-identified before model development. Access to individual-level health records requires appropriate legal authority, institutional approval and safeguards. Data-use agreements, provenance, collection method, reporting delay, units, detection limits and quality flags should be retained. Public datasets may not contain all required variables; unavailable features must be omitted or treated as unavailable, not fabricated.

D. Preprocessing: Records are harmonised by timestamp, geography, unit and source. Duplicate records are removed; physically implausible values are flagged for review rather than automatically deleted; outliers are investigated; and missingness indicators are retained. Imputation is fitted only on training data to prevent leakage. For time-series inputs, forward-looking interpolation is prohibited in retrospective forecasting tests. Source-quality weights reflect sensor calibration, laboratory method, sampling frequency and reporting completeness.

 ©IJSRED: All Rights are Reserved

E. Feature engineering: Candidate features include rolling rainfall totals, rainfall anomalies, temperature and humidity lags, flood indicators, water-quality trends, source type, sanitation access, water-service interruptions, vector indices, aquatic-host suitability and lagged disease incidence. Features must be selected separately for each pathway. For example, standing-water and vector indices are relevant to vector-borne disease forecasting, whereas household water quantity and hygiene access are more relevant to water-washed disease risk.

V. ALGORITHMIC DESIGN

A. Baseline models: Scikit-learn Logistic Regression, Random Forest and calibrated tree ensembles provide interpretable or robust tabular baselines. XGBoost is used for nonlinear tabular prediction and feature interaction modelling. These models are trained separately for each pathway and, where sample sizes permit, for selected disease outcomes. Baseline comparisons are essential because a complex neural model may not outperform gradient-boosted trees on modest tabular datasets.

B. Spatio-temporal model: The PyTorch implementation uses a graph representation in which nodes represent geographic units and edges encode justified relationships, such as adjacency, shared water-distribution networks or hydrological connectivity. Edge definitions should not rely on proximity alone when water flow or vector movement follows different patterns. At each time step, a graph message-passing layer aggregates neighbouring information; a temporal encoder (such as a gated recurrent unit or LSTM) processes historical embeddings. A pathway-specific prediction head produces a risk score for each outcome.

C. Multi-task pathway heads: Let X(i,t) denote the feature vector for location i at time t and A denote the graph adjacency matrix. The shared encoder produces Z(i,t)=fθ(X(i,t),A). Separate heads produce p(k)(i,t+h)=sigmoid(gφk(Z(i,t))), where k indexes waterborne, water-washed, water-based and water-related vector pathways, and h is the forecast horizon. Multi-label outputs are permitted because a location may face more than one pathway risk. Disease-specific outputs should be added only when adequate, reliable labels exist.

$$p_k(i,t+h) = \sigma(g_\{\varphi_k\}(f_\theta(X_\{i,t\}, A))) \quad (1)$$

D. Confidence-aware missingness: For each feature j, include both an imputed value x̃j and a missingness flag mj. A source-quality vector q records measurement or reporting reliability. The model receives [x̃, m, q] rather than imputed values alone. A simple observation-coverage score may be computed as C(i,t)=Σj wj rj / Σj wj, where rj indicates valid observation availability and wj reflects predeclared feature importance. Coverage is reported alongside risk and is not itself interpreted as a calibrated probability.

$$C(i,t) = (\Sigma_j\ w_j\ r_\{i,t,j\}) / (\Sigma_j\ w_j) \quad (2)$$

E. Calibration and uncertainty: Model probabilities are calibrated using a validation set that is separated chronologically from training data. Platt scaling or isotonic regression may be used where sample size permits. Uncertainty can be estimated using ensembles, bootstrap resampling or conformal prediction with assumptions and coverage limitations reported. A high predicted risk with poor data coverage should trigger a "review required" status rather than an automatic high-confidence alert.

F. Explainability: SHAP explanations are generated for tabular baselines and, where computationally feasible, for the final model. Explanations identify influential features for a prediction but do not establish causality. Reports should distinguish predictive association from causal effect. Counterfactual statements should be avoided unless the model and intervention assumptions support them.

G. Training objective: For pathway k, use a weighted binary cross-entropy or an appropriate count/time-series loss. Class weights or focal loss may address rare outbreaks, but resampling must occur only within training folds. A combined objective can be expressed as L = Σk λk Lk + β Lreg, where λk weights pathway tasks and Lreg controls model complexity. Hyperparameters are tuned using training/validation periods only.

VI. WATER-QUALITY ASSESSMENT AND PURIFICATION DECISION SUPPORT

A. Separate assessment from disease prediction: WISE-Guard includes a water-assessment module that evaluates measured parameters against applicable national or WHO guideline values and site-specific water-safety plans. It must distinguish sensor-derived indicators from laboratory microbiological results. A normal pH or turbidity reading does not prove the absence of pathogens, pesticides, heavy metals or other contaminants.

B. Treatment recommendation logic: Treatment selection follows a rule-based, evidence-constrained decision tree. First, identify the contaminant class and confirm the analytical evidence. Second, assess turbidity, organic load, salinity, pH, flow and operational constraints. Third, identify candidate treatment barriers and verify that the complete treatment train is validated for the target hazard. Fourth, require operational monitoring and post-treatment verification. If contaminant identity is unknown or measurements are incomplete, the system should recommend confirmatory testing and precautionary public-health guidance rather than a specific unverified treatment.

C. Microbial hazards: Coagulation/flocculation and clarification may reduce particle-associated organisms and turbidity; filtration provides a physical barrier; and disinfection can inactivate susceptible pathogens when dose, contact time, pH and water quality are controlled. Chlorination is not effective against Cryptosporidium oocysts under normal drinking-water disinfection conditions, and turbidity can reduce disinfection effectiveness [12]. Therefore, the framework must not

 ©IJSRED: All Rights are Reserved

recommend chlorination as a universal solution. Treatment performance should be based on validated technology specifications and local water-safety requirements [12], [13].

D. Chemical hazards: Chemical contamination requires contaminant-specific treatment. Depending on the substance and concentration, options may include activated carbon adsorption, ion exchange, oxidation, membrane processes or source substitution. Boiling and ordinary chlorination do not remove dissolved metals or salts. The model should not prescribe a generic treatment based on a composite score; it should identify the need for laboratory confirmation and refer to a qualified water-treatment professional.

E. Pathway-specific intervention mapping: Waterborne risk may trigger source inspection, microbiological sampling, safe-water advisories and verified treatment review. Water-washed risk may trigger water-access, hygiene and sanitation interventions. Water-based risk may trigger exposure reduction, aquatic-host surveillance and disease-specific control. Water-related vector risk may trigger standing-water management, vector surveillance and public-health vector control. This mapping is central to avoiding the false assumption that purification alone addresses all water-related diseases.

F. Human oversight: Recommendations are advisory. A qualified water professional or public-health authority must approve operational decisions. The interface records evidence, timestamp, model version, confidence, data gaps and the rationale for each suggested action. It must never label water "safe" solely because an ML model predicts a low risk.

VII. IMPLEMENTATION PLAN

A. Software stack: Python is the primary language. Pandas and NumPy support data preparation; Scikit-learn supports preprocessing, baselines, model selection and calibration; XGBoost supports boosted-tree models; PyTorch supports graph and temporal neural networks; SHAP supports feature attribution; GeoPandas and Folium or an equivalent GIS stack support spatial analysis and mapping; and Streamlit or a comparable web framework can present the decision-support dashboard.

B. Suggested project structure: data/raw/ stores source snapshots with access controls; data/processed/ stores validated derived datasets; src/ingestion/ handles source adapters; src/preprocessing/ handles cleaning and leakage-safe imputation; src/features/ creates lag and geographic features; src/models/baselines.py implements Scikit-learn and XGBoost models; src/models/spatiotemporal.py implements the PyTorch model; src/calibration/ handles probability calibration; src/explainability/ produces SHAP reports; src/treatment_rules/ stores versioned treatment logic; app/ contains the dashboard; and tests/ validates data contracts and model behaviour.

C. Development sequence: (1) define geography, time step and target labels with a public-health collaborator; (2) obtain permitted datasets and document provenance; (3) build a data-quality report; (4) establish simple baselines; (5) implement pathway-specific targets; (6) train the PyTorch spatial-temporal model only if the data volume and graph structure justify it; (7) calibrate and explain predictions; (8) implement treatment rules from authoritative guidance; (9) conduct retrospective and prospective shadow evaluation; and (10) deploy only after expert review and governance approval.

D. Minimum viable prototype: For an undergraduate implementation, begin with a district- or block-level weekly dataset and two targets: waterborne disease incidence/outbreak risk and water-quality non-compliance. Demonstrate the four pathway heads in the architecture, but train only those for which credible labels are available. The full multi-pathway system should be presented as the research design, while unsupported disease heads remain untrained or explicitly marked as future work. This is more defensible than creating synthetic labels and presenting them as real outcomes.

VIII. EXPERIMENTAL DESIGN AND EVALUATION

A. Research questions: RQ1: Does adding environmental and water-quality features improve pathway-specific forecasting over historical-incidence baselines? RQ2: Does spatial learning improve performance in connected regions compared with non-spatial models? RQ3: Does missingness-aware modelling improve robustness when monitoring data are incomplete? RQ4: Are calibrated risk estimates and explanations sufficiently reliable for human review? RQ5: Does the treatment decision layer select recommendations consistent with authoritative treatment guidance and measured contaminant classes?

B. Baselines: Compare persistence or seasonal-naïve forecasting, Logistic Regression, Random Forest, XGBoost, LSTM/GRU and the proposed graph-temporal model. Compare water-quality predictions separately from disease predictions. Do not compare a WQI regression score directly with a disease-outbreak classification score.

C. Validation: Use chronological splits (train on earlier periods, validate on a later period, test on the most recent holdout) and spatial holdouts (entire districts or blocks excluded from training). A combined test should evaluate later observations in unseen areas. All imputation, scaling, feature selection and oversampling must be fitted within training folds. Prevent leakage from post-outbreak reports, revised case totals and future weather observations.

D. Metrics: For classification, report sensitivity/recall, specificity, precision, F1-score, ROC-AUC and PR-AUC. For rare outbreaks, PR-AUC and sensitivity at a fixed false-alarm rate are particularly informative. For probabilistic predictions, report Brier score, calibration

 ©IJSRED: All Rights are Reserved

curves and expected calibration error. For count forecasts, report MAE, RMSE and a suitable count deviance. Also report lead time, false alerts per area-period, missed outbreaks and subgroup performance where data permit.

E. Ablation: Evaluate (1) historical disease data only; (2) plus climate; (3) plus water-quality variables; (4) plus sanitation/hygiene or vector/host indicators as appropriate; (5) plus graph learning; (6) plus missingness and source-quality features; and (7) calibrated full model. Treatment recommendations should be evaluated separately through expert review and scenario-based verification against documented treatment guidance.

F. Robustness and fairness: Simulate missing sensor intervals and delayed disease reporting using realistic patterns. Test sensor drift and distribution shift. Report performance across urban/rural settings and levels of data completeness. Avoid interpreting lower reporting rates as lower disease incidence. If under-reporting is severe, predictions should be accompanied by wider uncertainty and a data-quality warning.

G. Results status: This manuscript specifies a proposed methodology and evaluation plan. No experimental dataset, trained model, measured performance, or clinical/public-health effectiveness result is asserted here. Accuracy, F1-score, AUC, lead time and treatment performance must be populated only after reproducible experiments. Synthetic data may be used for software testing, but must be labelled synthetic and excluded from claims about real-world predictive performance.

IX. ETHICAL, OPERATIONAL AND LIMITATION CONSIDERATIONS

Disease surveillance data are sensitive even when aggregated, particularly in small communities. Data minimisation, de-identification, access control, retention limits and institutional approval are required. Predictions may stigmatise communities or create unnecessary alarm if uncertainty and reporting bias are ignored. Alerts should be phrased as requests for verification or investigation, not definitive outbreak declarations.

The framework depends on the availability, quality and representativeness of environmental and health data. Sensor measurements are indirect proxies for pathogen presence; microbiological testing remains necessary. Disease categories overlap, surveillance systems may under-report cases, and climate–disease relationships vary geographically. A model trained in one region may fail in another due to differences in infrastructure, ecology, healthcare access and reporting practices.

WISE-Guard is a research decision-support proposal, not a medical diagnostic tool, water certification system or autonomous treatment controller. Any future operational use requires external validation, public-health oversight, water-treatment expertise, cybersecurity review and prospective monitoring. Treatment advice must follow current national standards and a site-specific water safety plan.

X. CONCLUSION AND FUTURE WORK

This paper has presented WISE-Guard, an explainable multi-pathway framework for water-related disease-risk prediction and water-treatment decision support. Its central design choice is to distinguish waterborne, water-washed, water-based and water-related vector transmission rather than compressing them into a single disease class. The framework combines environmental, water-quality, sanitation, vector/host, geospatial and aggregated surveillance information; applies pathway-specific prediction heads; and reports calibrated risk with data-quality and explanation information.

The water-treatment module is intentionally separate from disease prediction. It uses contaminant-specific evidence and validated treatment guidance to identify candidate treatment barriers, while directing hygiene, aquatic-host and vector risks to appropriate non-purification interventions. The proposed Python implementation uses Scikit-learn and XGBoost for baselines and tabular models, with PyTorch for spatial-temporal learning.

Future work will focus on assembling authorised longitudinal datasets, selecting a feasible regional study area, validating pathway-specific labels with public-health experts, implementing the baseline pipeline, and testing whether spatial-temporal learning improves generalisation. Prospective shadow evaluation and expert review will be required before any public-health deployment. The principal expected contribution is a transparent, testable framework that connects environmental intelligence to differentiated prevention decisions.

A. Unit of analysis: The recommended unit is a geographic area observed at a regular time interval, such as a village-week, ward-week or district-week. The selected unit should be consistent across environmental and health sources and should not imply finer precision than the source data support. For example, district-level disease counts should not be downscaled to village-level labels without a defensible statistical method. Each record should contain a location identifier, interval start and end, source identifiers, feature values, missingness flags and data-quality metadata.

B. Feature schema: A practical initial schema includes location_id, week_start, rainfall_mm_7d, rainfall_mm_30d, mean_temperature, mean_relative_humidity, flood_flag, source_type, pH, turbidity_NTU, conductivity_uScm, TDS_mgL, dissolved_oxygen_mgL, free_chlorine_mgL where measured, E_coli_result, sanitation_coverage, household_water_access, water_service_interruptions, standing_water_index, vector_index, aquatic_host_indicator, population_denominator, and disease-specific reported counts. The schema must preserve measurement units, method, detection limit, timestamp and

 ©IJSRED: All Rights are Reserved

provenance. Variables not collected in a region should remain absent with a missingness flag.

C. Outcome labels: A disease-outbreak label must be defined before model training with public-health input. One option is a disease-specific threshold based on an established outbreak definition or a statistically justified increase over an expected baseline. A second option is to forecast case counts and derive an alert using a documented threshold. The threshold should not be selected to maximise test-set performance. Labels should distinguish suspected, presumptive and laboratory-confirmed cases where surveillance systems provide that distinction. The framework should avoid merging syndromic diarrhoea with laboratory-confirmed cholera or typhoid.

D. Forecast horizon: The prediction horizon should be selected according to the intended action. A one-week horizon may support rapid sampling and local investigation; a two- to four-week horizon may support resource planning but may have greater uncertainty. The horizon, observation window and reporting delay should be stated explicitly. Features must be available at the forecast issue time. A report that arrives after the target period cannot be used as a predictor for that forecast.

E. Dataset audit: Before modelling, produce a data dictionary, source coverage map, missingness heat map, distribution plots, class prevalence summary, date-range report and duplicate/consistency checks. Record how many geographic units and time intervals are available for each disease pathway. If the number of positive events is too small, use descriptive surveillance analysis or a simpler baseline rather than fitting a high-capacity neural network.

XI. DATASET DESIGN AND TARGET CONSTRUCTION

Algorithm 1 describes the proposed workflow. It is intentionally modular so that the tabular baseline can be completed before the graph-temporal component. The same chronological partitions and target definitions must be used for every model. Model selection is performed on validation data, while the final test set is opened only once for the final comparison.

Algorithm 1: WISE-Guard training and inference

Input: time-indexed environmental data X; graph A; pathway labels Yk; source-quality metadata Q; forecast horizon h. Output: pathway risks Pk, water-quality flags WQ, coverage C and explanation E.

1) Validate schema, units, timestamps, spatial identifiers and source provenance. 2) Aggregate records to the declared location/time unit. 3) Create chronological train, validation and test partitions; reserve geographic holdouts. 4) Fit imputation, scaling and feature selection on training data only. 5) Create missingness indicators, lagged variables and rolling environmental features using past observations only. 6) Train per-pathway Logistic Regression, Random Forest and XGBoost baselines. 7) Train the PyTorch graph-temporal encoder and pathway-specific heads where event volume supports it. 8) Tune hyperparameters on validation data and calibrate probabilities using a separate calibration split or cross-validation confined to training/validation. 9) Evaluate temporal, spatial and combined holdouts. 10) At inference, validate current data, calculate coverage and source-quality status, produce pathway risk estimates, apply calibration and uncertainty checks, generate SHAP explanations and log the model/data versions. 11) Pass water-quality evidence to the separate treatment-rule module. 12) Present alerts for expert review; do not issue autonomous diagnosis or treatment control.

A. Graph construction: Define nodes as the chosen geographic units. Candidate edges include adjacency, shared water-supply networks, upstream/downstream hydrological connections or empirically justified vector/ecological connectivity. Edge types may be stored separately. Avoid connecting locations simply because they are close if a river barrier, watershed divide or disconnected supply system makes the relationship implausible. The graph should be versioned and tested for isolated nodes and unstable connections.

B. Temporal windows: For each location, arrange observations in chronological order. Lag windows should be chosen with domain input and validated without looking at the final test period. Missing intervals should not be silently treated as zero. Include explicit missingness and elapsed-time indicators. For recurrent networks, use masks or sequence packing where appropriate. Compare the neural approach with XGBoost because deep learning may be less stable on small datasets.

C. Imbalance handling: Outbreak events are often rare. Report event prevalence and use class-weighted losses or threshold selection based on operational sensitivity and false-alarm tolerance. Synthetic oversampling, if considered, must be applied only to training folds and should not be used to create artificial time sequences that leak information. Precision-recall curves and event-based evaluation are more informative than accuracy alone in highly imbalanced settings.

D. Alert thresholds: A probability threshold should be selected with public-health stakeholders based on the cost of missed events, the capacity to investigate alerts and the acceptable false-alert burden. A three-level low/moderate/high label may be used for display, but the underlying probability, calibration status, observation coverage and threshold version should remain visible. Thresholds may differ by pathway and should be documented.

XII. WISE-GUARD TRAINING AND INFERENCE PROCEDURE

Table I summarises illustrative treatment-support logic. It is not a universal operating prescription. Actual selection depends on laboratory confirmation, water chemistry, flow rate, system design, local standards and validated

 ©IJSRED: All Rights are Reserved

performance. The recommendation engine should display the evidence needed before a treatment option is considered.

TABLE I — ILLUSTRATIVE CONTAMINANT-TO-RESPONSE MAPPING

• Elevated turbidity / suspended solids: consider source protection, settling or coagulation/flocculation followed by filtration as appropriate; verify turbidity after treatment and ensure disinfection conditions are met. Turbidity alone does not identify a pathogen.

• Suspected faecal microbial contamination: initiate confirmatory microbiological sampling and a water-safety response. A validated treatment train may include clarification/filtration and disinfection, with operational checks for dose, contact time, pH and residual where applicable. Do not assume a single sensor can confirm pathogen removal.

• Protozoan concern, including Cryptosporidium: select a validated physical removal barrier and verify performance; ordinary free-chlorine disinfection is not a reliable inactivation barrier for Cryptosporidium oocysts [12].

• Dissolved salts / salinity: consider source substitution, blending or a suitable membrane/desalination process after full water chemistry assessment. Boiling does not remove dissolved salts and may concentrate them.

• Arsenic, fluoride, nitrate or heavy metals: require contaminant-specific laboratory analysis and a validated adsorption, ion-exchange, membrane or other technology as appropriate. Boiling and ordinary chlorination are not reliable removal methods for these dissolved chemical hazards.

• Unknown contaminant or conflicting readings: do not automatically select a treatment. Flag the source for laboratory investigation, inspect the sensor and apply the responsible authority's precautionary advice.

A. Treatment evidence record: Each recommendation should include the target contaminant, source measurement and method, guideline or standard reference, treatment process considered, known limitations, required operating conditions, verification test, and date of the rule-set review. A rule must not be enabled until reviewed by a qualified water-treatment professional.

B. Water-safety verification: A model output is not a compliance certificate. The system should require post-treatment verification and retain laboratory results, sensor calibration status and operational records. Where a water safety plan exists, the rule engine should align with its hazard analysis, control measures and monitoring programme.

C. Disease prevention mapping: For water-washed risks, recommend assessment of water quantity, reliability and hygiene access rather than merely disinfection. For water-based risks, present exposure and host-surveillance indicators. For vector-related risks, present breeding-site and vector surveillance information. The user interface should make clear why a water purification action is or is not relevant to each pathway.

XIII. WATER TREATMENT DECISION MATRIX AND SAFETY CONTROLS

A. Data ingestion module: Each connector should convert its source into a documented internal schema. Automated checks should reject invalid units, impossible timestamps, duplicated identifiers and unknown geography codes. Raw source snapshots should be immutable and access-controlled. A manifest should record retrieval date, licence or access terms, version, transformations and checksums where practical.

B. Model module: Scikit-learn pipelines should bundle preprocessing and estimators so that transformations are fitted only on training folds. XGBoost parameters, random seeds, software versions and feature lists should be saved. PyTorch training should log graph version, sequence length, optimiser, learning rate, batch size, early-stopping rule and checkpoint hash. Model cards should document intended use, excluded use cases, known limitations and evaluation populations.

C. Testing: Unit tests should cover unit conversion, aggregation, missingness flags, lag generation, graph connectivity and threshold logic. Integration tests should verify that the dashboard displays the same calibrated probability and coverage values produced by the model service. Regression tests should detect changes in predictions after software updates. A treatment-rule test suite should include known scenarios for turbidity, microbial hazards, protozoa and dissolved chemical contaminants.

D. Reproducibility: Publish a data dictionary, preprocessing specification, model configuration, evaluation scripts and non-sensitive derived data where licences permit. If source data cannot be redistributed, document the access process and provide code that can run against an authorised local copy. Report the exact date range, geography, sample counts, number of outbreak events, exclusion criteria and missing-data profile.

E. Dashboard design: Each geographic unit should show current data coverage, last observation time, pathway-specific risk, forecast horizon, calibration status, key explanatory variables and suggested verification steps. Maps should avoid visually implying precise local risk when source data are district-level. Alerts should include a clear timestamp and expiry or review date so that stale warnings are not mistaken for current conditions.

F. Security and operations: Apply role-based access, encryption in transit and at rest, audit logging and secure secret management. Avoid displaying identifiable health information. Define an incident process for erroneous alerts, compromised sensors or data outages. Maintain a human override and a mechanism to withdraw or correct a misleading alert.

 ©IJSRED: All Rights are Reserved

XIV. SOFTWARE MODULES, TESTING AND REPRODUCIBILITY

A. Construct validity: The term water-related disease covers several pathways, and the selected indicators may only approximate exposure. For example, rainfall is not a direct measure of pathogen concentration, and turbidity is not a universal proxy for microbiological contamination. Disease-specific definitions and expert review are necessary.

B. Internal validity: Reporting delays, changes in testing practice, public-health interventions and concurrent outbreaks can influence observed associations. Leakage can occur when revised case totals or future environmental observations are used in retrospective features. Strict timestamp audits and prospective shadow testing reduce, but do not eliminate, these risks.

C. External validity: A model trained in one state, climate zone or water-supply system may not generalise to another. Spatial holdouts, calibration by region and local validation are required. The model should be retrained only under documented governance, not automatically in response to every new observation.

D. Operational validity: An alert is useful only if the receiving authority can verify it and respond. A highly sensitive model that produces too many false alerts may overwhelm limited field teams. Evaluation should therefore include workload-related measures such as alerts per district-week and positive predictive value at the chosen operating threshold.

E. Expected contribution: WISE-Guard is expected to provide a common, transparent architecture for differentiating water-related disease pathways, forecasting risk from heterogeneous environmental and surveillance information, and connecting water-quality evidence to treatment options. The contribution must ultimately be demonstrated through comparative experiments, expert-reviewed intervention logic and, ideally, a prospective pilot. Until those evaluations are complete, the work should be described as a proposed framework rather than a validated public-health system.

XV. THREATS TO VALIDITY AND EXPECTED CONTRIBUTION

REFERENCES

[1] World Health Organization, “Introduction to water-related infectious diseases: Classification of water-related infectious diseases,” WHO training material, 2022. [Online]. Available: https://www.who.int/docs/librariesprovider2/default-document-library/module_1.1_introduction-wrid_ds_220831.pdf

[2] P. R. Hunter, “Addressing Risk for Waterborne Disease,” in Global Issues in Water, Sanitation, and Health, National Academies Press / NCBI Bookshelf. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK28460/

[3] “Toward Design of Internet of Things and Machine Learning-Enabled Frameworks for Analysis and Prediction of Water Quality,” IEEE Access, vol. 11, pp. 101055–101086, Sep. 2023, doi: 10.1109/ACCESS.2023.3315649.

[4] “Machine Learning Approach to IoT-Based Water Quality Monitoring,” in Proc. 2023 IEEE 5th Eurasia Conf. Biomedical Engineering, Healthcare and Sustainability (ECBIOS), 2023, doi: 10.1109/ECBIOS57802.2023.10218420.

[5] “Prediction of Water Quality Using SoftMax-ELM Optimized Using Adaptive Crow-Search Algorithm,” IEEE Access, vol. 11, pp. 140900–140913, Dec. 2023, doi: 10.1109/ACCESS.2023.3339564.

[6] “AI/ML-Based Water Quality Monitoring Mobile App for Predicting E. coli in Surface Waters,” in Proc. 2025 19th Int. Conf. Ubiquitous Information Management and Communication (IMCOM), 2025, doi: 10.1109/IMCOM64595.2025.10857585.

[7] “AI and IoT-Assisted Water Quality Monitoring and Prediction,” in Proc. 2024 8th Int. Conf. Electronics, Communication and Aerospace Technology (ICECA), 2024, doi: 10.1109/ICECA63461.2024.10801021.

[8] “Early Detection and Warning of Water-Borne Diseases using Deep Learning,” in Proc. 2026 Fourth Int. Conf. Augmented Intelligence and Sustainable Systems (ICAISS), 2026, doi: 10.1109/ICAISS68683.2026.11526699.

[9] “Smart Community Health Monitoring and Early Warning System for Water-Borne Diseases in Rural,” in Proc. 2026 5th Int. Conf. Communication, Computing and Electronics Systems (ICCCES), 2026, doi: 10.1109/ICCCES62661.2026.11436313.

[10] “Applications of artificial intelligence (AI) in drinking water treatment processes: Possibilities,” Chemosphere, vol. 356, Art. no. 141958, 2024, doi: 10.1016/j.chemosphere.2024.141958.

[11] “Intelligent Water Quality Monitoring and Prediction System,” in Proc. 2025 6th Int. Conf. Inventive Research in Computing Applications (ICIRCA), 2025, doi: 10.1109/ICIRCA65293.2025.11089780.

[12] World Health Organization, Guidelines for Drinking-water Quality: Fourth Edition Incorporating the First and Second Addenda, Geneva, 2022. See also WHO, Guidelines for Drinking-water Quality, fourth edition incorporating the second and third addenda, 2026. [Online]. Available: https://www.who.int/publications/i/item/9789240121225

[13] World Health Organization, Evaluating Household Water Treatment Options: Health-Based Targets and Microbiological Performance Specifications, Geneva, 2011. [Online]. Available: https://www.who.int/publications/i/item/9789241548229

[14] World Health Organization, “Vector-borne diseases,” Fact sheet, Sep. 26, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/vector-borne-diseases

 ©IJSRED: All Rights are Reserved