Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Rainfall-Adaptive Graph Transformer Framework with Federated Edge Deployment for Multi-Label Waterborne Disease Risk Classification

Rashmi Dubey

rashmilpu2026@gmail.com

Lovely Professional University

Research Article

Keywords: Groundwater quality, Public health surveillance, Multi-modal data fusion, Privacy-preserving machine learning, Internet of Things, India

Posted Date: November 4th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11599494/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Rainfall-Adaptive Graph Transformer Framework with Federated Edge Deployment for Multi-Label Waterborne Disease Risk Classification

Rashmi Dubey

Department of Computer Science & Engineering, Lovely Professional University, Phagwara, Punjab 144411, India

Corresponding author: Rashmi Dubey, E-mail: rashmilpu2026@gmail.com

Abstract

Waterborne diseases have become a continuous problem across India's groundwater-dependent states. Predicting them has become a challenge because water's chemical composition data sits scattered across concerned departments of state, and privacy regulations prevent sharing. This paper introduces the Biochemical Weather Groundwater Network (BWG-Net), a federated deep learning framework that brings intelligence to where the data exist and removes the need to transfer sensitive records to a central server. The framework is trained on 11,240 Central Ground Water Board station records from five Indian states, combining 16 biochemical water quality parameters with 6 meteorological variables fetched via station coordinates from the India Meteorological Department archives. It classifies contamination into 7 health risk categories based on World Health Organization and Bureau of Indian Standards limits. The architecture combines a rainfall-adaptive dynamic graph, which updates how stations are connected based on precipitation intensity, with a cross-modal attention module that learns how weather conditions amplify the danger of specific chemical contaminants. Federated learning lets each state train on its own data and share only model weights. On held-out test data, BWG-Net achieved 90.4% accuracy and an area under the receiver operating characteristic curve (AUC-ROC) of 0.939, outperforming extreme gradient boosting (XGBoost) and a convolutional neural network–long short-term memory (CNN-LSTM) model by 16.6 and 9.5 AUC-ROC points, respectively. The system also runs on a Raspberry Pi 4 at 73 ms per inference, making it deployable in field settings.

Keywords Groundwater quality · Public health surveillance · Multi-modal data fusion · Privacy-preserving machine learning · Internet of Things · India

Introduction

Across India, the struggle for clean drinking water is more than a logistical challenge; it is a full-scale public health crisis (Khant et al., 2025). The numbers are daunting; World Health Organization (WHO) data suggests that nearly 2 billion people globally are still drinking water contaminated with faecal pathogens (World Health Organization, 2017). In India, this risk peaks during the monsoon season when heavy rains wash sewage and industrial waste directly into the groundwater systems millions of people rely on. Because threats like sewage-related enteric disease, arsenic poisoning, and fluorosis often appear together in the same polluted source, simple "single-risk" models are no longer enough. We need systems that can track multiple, co-occurring threats at once.

The data to solve this exists, but it remains trapped. India's Central Ground Water Board (CGWB) maintains a massive network of stations, and their 2023 dataset offers 16 different biochemical parameters for thousands of locations. Yet, this wealth of information has been almost entirely ignored by the machine learning community. Instead of using these authentic national records, many existing studies rely on small or site-specific datasets, and class imbalance and limited data diversity remain open problems (Nambiar & Gupta, 2025). This lack of a real-world foundation has created a massive gap between academic research and the actual needs of public health officials. We developed the Biochemical Weather Groundwater Network (BWG-Net) to address these problems together. By pulling directly from CGWB 2023 records and matching them with India Meteorological Department (IMD) weather archives, we built a model grounded in real-world hydrology (Central Ground Water Board, 2023). BWG-Net uses a "rainfall-adaptive" dynamic graph to track how contamination moves between stations and employs a graph transformer to blend weather and biochemical signals into a single, intelligent prediction. To handle the privacy problem (Shah et al., 2025), we used federated learning, allowing different states to collaborate on a powerful global model while keeping their own raw data strictly local and secure.

Contributions

We have designed **BWG-Net** to bridge the gap between advanced machine learning and real-world public health needs. Our key contributions include:

i) **Bridging National Datasets:** This work is the first to integrate India's official 2023 CGWB National Groundwater Quality records with IMD weather archives. We

processed 11,240 station records from five states, a massive, authentic dataset that has never before been used for federated disease risk modelling.

ii) **Rainfall-Adaptive Modeling:** At the core of our framework is a dynamic graph that updates the connections between monitoring stations in real-time based on rainfall intensity. This mechanism is grounded in established hydrological transport theory, allowing the model to simulate how heavy rains physically drive the movement of contaminants.

iii) **Cross-Modal Intelligence:** We developed a unique attention architecture that treats biochemical markers as the "query" and weather signals as the "key-value" pairs. This allows the model to "learn" and pinpoint exactly which meteorological conditions, such as sudden temperature shifts or intense rain, amplify specific chemical toxicity levels.

iv) **WHO-Aligned Risk Taxonomy:** To ensure our results are practically useful for health officials, we mapped our predictions to a 7-class risk system based on WHO and Bureau of Indian Standards (BIS) standards. We also introduced Gaussian noise into the label training to prevent the model from learning rigid, unrealistic thresholds, making the system more robust in real-world scenarios.

v) **Privacy Without Compromise:** We demonstrated that privacy-preserving technology can be both fast and accurate. By deploying BWG-Net across five state-level clients using federated learning, we achieved an edge inference speed of just 73 ms on a Raspberry Pi 4. Remarkably, this decentralized approach resulted in only a 1.5-point drop in the area under the receiver operating characteristic curve (AUC-ROC) compared to traditional, centralized training.

Literature review

Research at the intersection of machine learning and waterborne disease prevention has grown rapidly, drawing on fields as diverse as epidemiology, graph learning, and distributed computing. However, a close look at existing studies reveals several persistent structural gaps. No single previous framework has managed to address all of these challenges at once, which is the primary motivation behind BWG-Net.

Epidemiology, climate-disease interactions, and groundwater monitoring

The link between environmental shifts and public health is well-documented but complex. A major systematic review found that diarrheal disease is positively associated with ambient temperature, heavy rainfall and flooding (Levy et al., 2016), showing that the relationship between weather and disease is nonlinear and highly dependent on local context. This complexity is why a static model isn't enough; it justifies our use of a rainfall-adaptive dynamic graph.

Process-based modelling has further shown that contaminant dissipation depends on transport distance and that rainfall events are the main control on how contaminants move (Shang et al., 2025). This motivates the distance-decay, rainfall-scaled form of the edge weights in BWG-Net.

Even with the right math, real-world data is often messy. Real-world Internet of Things (IoT) water quality monitoring, such as the sensor framework deployed by Rahu et al. (2023), relies on low-cost field sensors that are prone to fouling, calibration drift, and power loss. This reality makes it clear that any reliable system must include a dedicated layer to detect and handle sensor faults before the data is ever used for a prediction.

Machine learning approaches for water quality and disease prediction

Current machine learning benchmarks show both promise and clear limitations. For instance, extreme gradient boosting (XGBoost) reached 99.65% accuracy on a tabular water quality classification task (Hridoy et al., 2025). However, the authors note that their data lacks spatial and temporal context, which limits generalisation across environmental conditions. Similarly, deep learning models for water quality classification remain limited by class imbalance and data diversity (Nambiar & Gupta, 2025), and sequence models such as convolutional neural network–long short-term memory (CNN-LSTM) networks do not use the physical connections between monitoring sites.

The importance of weather data cannot be overstated. Climate change is expected to alter infectious disease transmission patterns, which supports including weather variables in risk models (Tong et al., 2015). In India specifically, Malakar et al. (2021) applied deep learning to CGWB monitoring-well data across the country, confirming that this national network can support groundwater level forecasting. However, their work, like many others, stopped short of incorporating dynamic graphs, multi-label disease outputs, or the privacy-preserving federated training needed for a multi-state rollout.

Graph neural networks for spatial environmental modeling

Graph Attention Networks (GATs) have emerged as the standard for modeling data that has a clear spatial structure, as they allow a model to "learn" which neighboring nodes are the most important rather than relying on fixed assumptions (Veličković et al., 2018). We have extended this concept in BWG-Net by making those edge weights dynamic based on rainfall.

The practical benefits of graph structures are evident in broader environmental research as well. Zanfei et al. (2022) showed that Graph Convolutional Network (GCN) models, which use the network's graph structure, have high potential for burst detection in water distribution systems when compared with multi-layer perceptron benchmarks. Furthermore, spatiotemporal graph neural networks (GNNs) with self-attention have achieved better accuracy than benchmark models for air quality prediction (Long & Ma, 2025). Finally, research into traffic flow provides a precedent for our approach: a spatial-temporal graph attention transformer was used to model changing spatial dependencies for forecasting (Wu et al., 2024).

Multi-modal fusion, multi-label classification, and federated learning

When merging diverse data sources, attention-based fusion has been used to combine heterogeneous spatio-temporal data (Long & Ma, 2025). This informed our choice of an asymmetric cross-attention design for BWG-Net. Similarly, since waterborne diseases often co-occur, we prioritized a multi-label approach. Work on California groundwater supports this, showing that a multi-label framework can predict many co-occurring contaminants at once (Dong et al., 2024). To handle class imbalance, we adopted DeepSMOTE, which oversamples in a learned latent space and outperformed traditional oversampling on imbalanced image data (Dablain et al., 2023).

For the federated aspect of our system, we relied on key work on federated learning (Li et al., 2020) and on the federated averaging (FedAvg) algorithm, which trains accurate models even when data is unevenly distributed across clients (McMahan et al., 2017). This gave us the confidence to deploy across five distinct Indian states. Finally, adaptive neuro-fuzzy inference system (ANFIS)-based schemes have been validated for detecting and compensating sensor faults in real time (Sowmmiya & Uma, 2018), which motivates our Layer 1 fault detection. Table 1 summarises how the key limitations of these prior studies map to the corresponding contributions of BWG-Net.

Table 1 Gap analysis mapping prior work limitations to BWG-Net contributions

Study	Core Approach	Key Limitation	BWG-Net Response
Levy et al. (2016)	Climate-disease systematic review	No predictive model; no station-level data	Climate–disease findings ground the rainfall-adaptive edge weight formula
Rahu et al. (2023)	IoT and ML water quality framework	Single canal site; no spatial graph or disease output	H-ANFIS fault detection addresses real-world sensor noise
Hridoy et al. (2025)	XGBoost/LightGBM on static features	No spatial propagation; no federation	Rainfall-weighted dynamic edges; federated multi-label classification
Nambiar & Gupta (2025)	Review of deep learning for imbalanced water quality data.	No spatial graph; no weather integration	Dynamic GNN with cross-modal biochemical-weather fusion
Malakar et al. (2021)	DL on CGWB groundwater level data	No dynamic graph; no weather; no federation	Gaps addressed on the same CGWB data context
Zanfei et al. (2022)	GCN for water distribution anomaly	Infrastructure leak detection; no disease taxonomy	Extended to WHO disease taxonomy with federated deployment
Veličković et al. (2018)	Graph Attention Networks	General graph ML; no environmental domain	GAT extended to dynamic cross-modal disease risk estimation
Li et al. (2020)	Federated learning methodology	No water quality or graph architecture	FedAvg across 5 Indian state clients with GNN backbone
Dong et al. (2024)	Multi-label ML for groundwater contaminants	No dynamic graph; no federation	Multi-label combined with dynamic GNN and federated privacy

Materials and methods

Proposed methodology

We designed the **Biochemical Weather Groundwater Network** as a modular, five-layer pipeline specifically to bridge the gap between raw field data and actionable public health intelligence. As outlined in Fig. 1, the framework takes authentic CGWB station records and processes them into federated, multi-label risk estimates ready for edge deployment. We didn't just build these layers for performance; each one is purpose-built to tackle a specific practical hurdle whether it’s cleaning noisy sensor data or protecting jurisdictional privacy. A high-level summary of these layers and their functional roles is provided in Table 2.

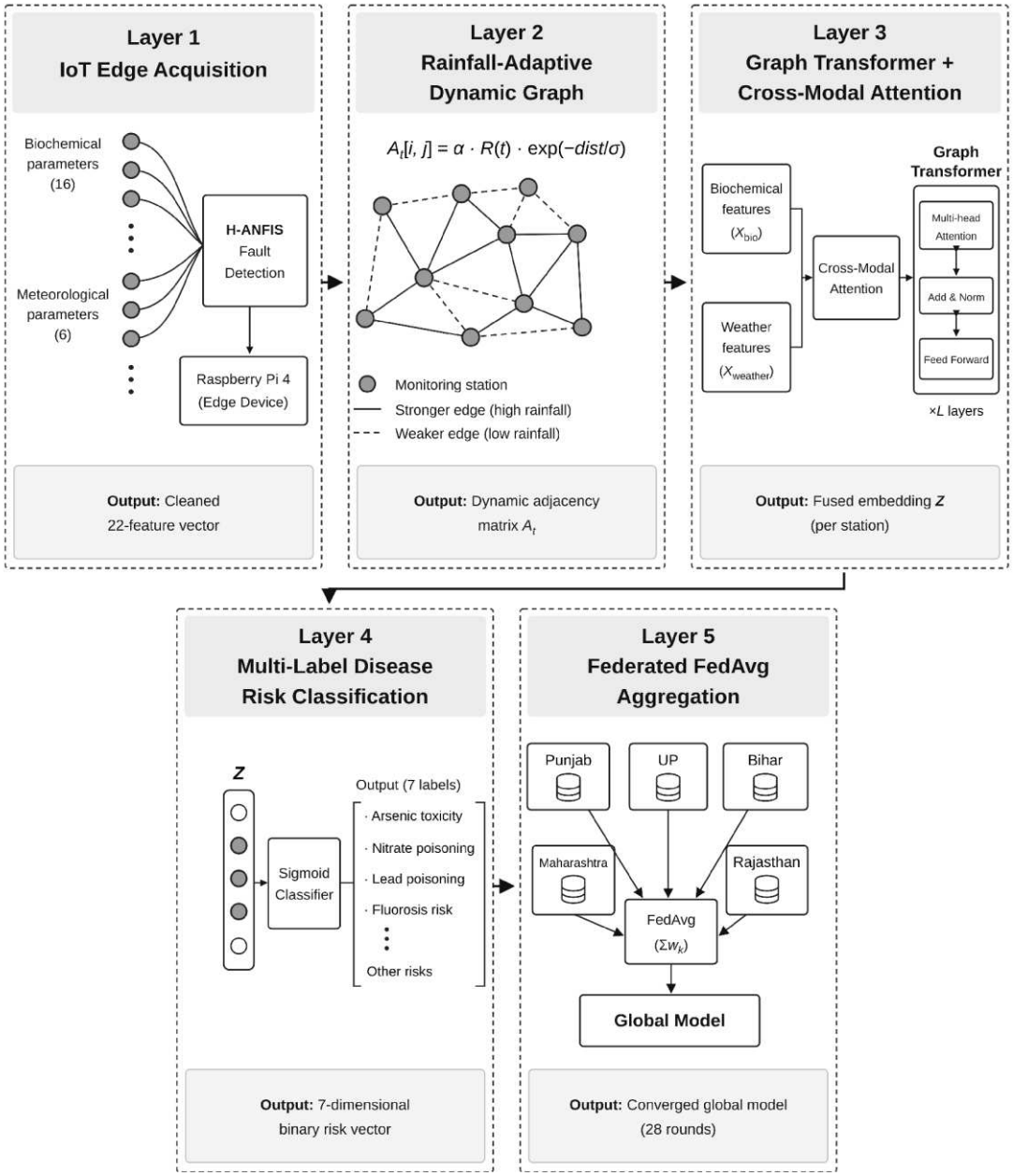


Fig. 1 BWG-Net five-layer architecture: H-ANFIS sensor fault detection (L1), rainfall-adaptive dynamic graph construction (L2), cross-modal Graph Transformer fusion (L3), multi-label sigmoid classifier (L4), and federated FedAvg aggregation across five Indian state clients (L5)

Table 2 BWG-Net architecture overview with technical function and design justification for each layer

Layer	Component	Design Justification	Output
L1	IoT Edge + H-ANFIS	Field CGWB sensors exhibit fouling and calibration drift; ANFIS-based schemes detect and compensate sensor faults in real time (Sowmmiya & Uma, 2018)	Clean 22-feature vector per station
L2	Rainfall-Adaptive Dynamic Graph	Contamination propagation between stations is physically modulated by rainfall intensity; static adjacency is physically inaccurate	Dynamic adjacency matrix $A_t$per time step

Layer	Component	Design Justification	Output
L3	Graph Transformer + Cross-Attention	Concatenation misses multiplicative amplification; cross-attention learns which weather conditions amplify each specific toxicity risk	Fused embedding Z (d=128) per station
L4	Multi-Label Sigmoid Classifier	Co-occurring contaminants in real CGWB data require multi-label output; temporal term prevents embedding instability	7-dim binary disease risk vector
L5	Federated FedAvg	State-level CGWB data cannot legally be centralized; FedAvg is robust to unevenly distributed client data (McMahan et al., 2017)	Global model; 1.5-point AUC overhead

Layer 1: Sensor fault detection with H-ANFIS

In the real world, monitoring stations aren't the sterile environments found in clean datasets. Sensors get clogged with debris, they drift when temperatures swing, and power outages create unpredictable gaps. While most research ignores these operational headaches, we knew they would cripple our model's performance on actual CGWB records. We chose a hybrid ANFIS (H-ANFIS) for this layer because it handles this messiness better than standard tools. Its fuzzy inference component can distinguish between a genuine contamination spike and a mere sensor glitch, while its neural backbone identifies complex fault patterns that simple, rule-based thresholds would miss. ANFIS-based schemes have been validated for real-time sensor fault detection and compensation (Sowmmiya & Uma, 2018), and the model is light enough for low-power hardware like the Raspberry Pi 4. Our own ablation tests (see Results and discussion) back this up: removing H-ANFIS caused a 2.8-point drop in AUC, proving just how much noise it filters out of the raw data.

Layer 2: Rainfall-adaptive dynamic graph construction

The way pollution moves between monitoring sites is never static. During the monsoon, water races between upstream and downstream stations, flushing toxins across much larger distances. In the dry season, that movement slows to a crawl and remains localized. Because of this, using a fixed map of how stations are connected is physically inaccurate. We solved this by building a dynamic graph that updates itself at every time step using the following formula:

$$A_t[i][j] = \alpha \cdot R(t) \cdot exp(-\frac{dist(i,j)}{\sigma}) \quad (1)$$

Here, $R(t)$represents the actual rainfall intensity recorded by the IMD for that specific area, while the distance is calculated from the stations' precise coordinates. We fine-tuned our parameters to $\alpha = 0.85$and $\sigma = 25$km based on our validation tests. The exponential distance-

decay form is a modelling assumption, motivated by Shang et al. (2025), who found that contaminant dissipation depends on transport distance and that rainfall events control transport. We can see the dramatic shift in how these connections fluctuate between the wet and dry seasons in Fig. 2.

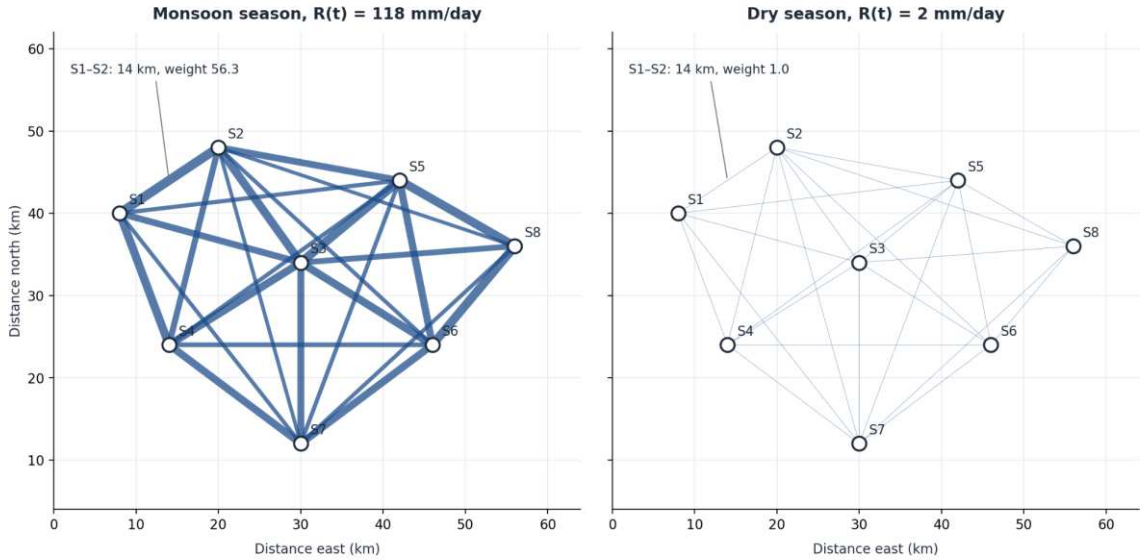


Fig. 2 Rainfall-adaptive dynamic graph

This comparison shows how edge weights shift between the monsoon season ($R(t) = 118$ mm/day) and the dry season ($R(t) = 2$ mm/day) for an illustrative cluster of eight stations within one state. Line width is proportional to the edge weight, so thicker edges represent stronger pathways for contamination to spread. In the model, distances are computed with the Haversine formula from CGWB station coordinates, with $\alpha = 0.85$and $\sigma = 25$km.

Layer 3: Graph Transformer with cross-modal attention fusion

At the heart of BWG-Net is a multi-head Graph Transformer (Yun et al., 2019) that operates on our rainfall-adaptive graph. The most unique part of this design is the cross-modal attention module. We built this based on a key insight: while biochemical factors are the primary drivers of disease, weather acts as a powerful amplifier. For example, a high nitrate level during a dry spell is dangerous, but the same level during a 100 mm rainfall event is far worse because floodwater carries the contamination to many more wells.

Traditional models often just stack these variables together, treating them as equal. We found that this "flat" approach fails to capture how weather actually amplifies chemical risk. To solve this, our attention mechanism assigns biochemical features to the "query" role, while meteorological data serves as the "keys and values." This forces the model to learn exactly which weather patterns enhance specific contamination profiles.

$$Q = W_Q X_{bio}, K = W_K X_{weather}, \tag{2}$$

$$V = W_V X_{weather} \tag{3}$$

$$Z_{attn} = softmax(\frac{QK^T}{\sqrt{d_k}})V, d_k = 64 \tag{4}$$

After this fusion, we use GCN aggregation to spread this information across the dynamic graph. We chose two-hop propagation (two GCN layers) because it allows the model to detect contamination trends that reach beyond a station's immediate neighbors.

$$H^{(l+1)} = ReLU(D^{-1/2} A_t D^{-1/2} H^{(l)} W^{(l)}) \tag{5}$$

This asymmetric design isn't just a theoretical preference our tests showed a 6.5-point AUC improvement over simple feature concatenation, proving that the model better understands the relationship between weather and water quality.

Layer 4: Multi-label WHO-grounded disease risk classification

Our classification system uses a 7-class taxonomy (Table 3) based on the WHO Guidelines for Drinking-Water Quality (World Health Organization, 2017) and BIS IS 10500 limits. We didn't just pull these categories from a list; we validated them against actual pollution trends found in the 2023 CGWB district-level records to ensure they reflect the reality on the ground (Central Ground Water Board, 2023). Multi-label machine learning frameworks applied to groundwater chemical data have demonstrated the ability to simultaneously predict multiple co-occurring contaminants with high accuracy (Dong et al., 2024). Dong et al. (2024) predicted 35 different per- and polyfluoroalkyl substances (PFAS) simultaneously in California groundwater using a multilabel semisupervised framework, providing methodological precedent for our use of a sigmoid multi-label output over single-label softmax. Our testing confirmed that this wasn't just a theoretical concern imposing a single-label constraint caused the model's AUC to drop by 3.6 points.

To guide the training, we developed a combined loss function:

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{BCE} + \lambda_2 \|Z_t - Z_{t-1}\|_F^2 \tag{6}$$

This formula uses binary cross-entropy ($\mathscr{L}_{BCE}$) to manage the 7-class output, while the temporal regularization term (the Frobenius norm) prevents the model's internal embeddings from becoming unstable between consecutive records. After running a grid search on our validation data, we found that setting $\lambda_1 = 1.0$and $\lambda_2 = 0.15$provided the most reliable performance

Table 3 BWG-Net 7-class groundwater health risk taxonomy based on WHO Guidelines (2017), BIS IS 10500 and CGWB 2023 data

S.No.	Disease Class	Primary Trigger Parameters	Threshold Basis
1	Enteric Disease Risk (sewage contamination proxy)	Nitrate > 50 mg/L together with chloride > 250 mg/L	Indirect indicator of sewage infiltration; WHO nitrate guideline (World Health Organization, 2017); BIS IS 10500 chloride limit
2	Arsenic Toxicity	As > 0.01 mg/L; elevated in Bihar CGWB zones	WHO limit 0.01 mg/L (World Health Organization, 2017)
3	Lead Poisoning Risk	Pb > 0.01 mg/L; elevated TDS as co-indicator	WHO limit 0.01 mg/L (World Health Organization, 2017)
4	Fluorosis Risk	Fluoride > 1.5 mg/L; Rajasthan high-risk zones	WHO limit 1.5 mg/L (World Health Organization, 2017)
5	Nitrate Poisoning	NO3 > 50 mg/L; agricultural zone flag	WHO limit 50 mg/L (World Health Organization, 2017)
6	Algal Bloom Risk	PO4 > 0.1 mg/L; rainfall > 80 mm/24 h; temperature > 30 °C	Nutrient enrichment under warm, wet conditions
7	General Chemical Risk	TDS > 500 mg/L; heavy metal composite index elevated (Fe, Mn, Cu, Zn, Ni, Cd, Cr)	BIS IS 10500 acceptable limit 500 mg/L; composite multi-pollutant score

Layer 5: Federated FedAvg aggregation

One of our core priorities was ensuring that sensitive raw station records never leave their home state. To achieve this, we set up five state-level clients that perform local training for five epochs. Instead of sharing the data itself, these clients only transmit their model weight updates to a central aggregator. We used a FedAvg approach to calculate a weighted average based on the volume of local records (McMahan et al., 2017):

$$W_{global} = \sum_{k=1}^{K} \frac{n_k}{n} W_k \; = \; \textbf{5 state clients} \quad (7)$$

By weighting the updates by the number of station records (n_k) relative to the total (n), we accounted for the fact that some states have significantly more monitoring sites than others. Our system reached stability in 28 communication rounds, which aligns well with the established theoretical guarantees for federated convergence.

Experimental setup

Dataset and preprocessing

We compiled a comprehensive dataset of 11,240 station records from five Indian states. Each record tracks 16 biochemical parameters selected from the CGWB data: pH, electric conductivity, total dissolved solids (TDS), chloride, nitrate, phosphate, fluoride, arsenic, lead, iron, manganese, copper, zinc, nickel, cadmium and chromium. These five states were selected for their highly distinct pollution profiles: the agricultural heartlands of Punjab and Uttar Pradesh (UP) (high nitrates and phosphates), Bihar (heavy arsenic), Maharashtra (industrial heavy metals), and Rajasthan (high fluoride levels).

To capture the influence of weather, we co-registered each station with IMD gridded archives at 0.25-degree resolution. This added six environmental variables to each record, including soil moisture and flood indices. Since roughly 8.3% of the CGWB entries were missing data, we used state- and season-specific medians for imputation; this ensured we didn't accidentally smear regional trends across different climates. To handle the stark class imbalance while respecting geographic boundaries, we applied DeepSMOTE independently for each state client (Dablain et al., 2023).

Label construction transparency: Clinical outcome data is not available for these stations, so the disease risk labels were constructed by applying the WHO and BIS thresholds in Table 3 to the validated biochemical data. The CGWB data also contains no microbial indicators. Enteric disease risk is therefore represented by an indirect sewage contamination proxy, in which nitrate and chloride are elevated together. A small amount of Gaussian noise ($\sigma = 0.07$) was added before binarizing so that the model does not simply learn a deterministic "lookup table." These labels are proxies for risk, and we cross-verified them against the 2023 CGWB district-level pollution maps.

Baselines and training protocol

To prevent the model from "cheating" through data leakage, we used a strict 70/15/15 temporal split. We pitted BWG-Net against a range of standard models, including XGBoost, standalone LSTMs and CNN-LSTMs, and a Graph Attention Network (Veličković et al., 2018). We also included a standard Graph Transformer trained centrally as an architectural baseline, and trained the full BWG-Net centrally to measure the cost of federation.

Our hyperparameters including a hidden dimension of 128 and 8 attention heads—were selected via grid search. We trained the model using Adam with a learning rate of 0.001 and used paired t-tests ($p < 0.05$) across five runs with different seeds to ensure our results were statistically sound. All training was conducted on an NVIDIA RTX 3080, with edge latency benchmarks performed on a Raspberry Pi 4.

Results and discussion

Comparison with baselines

When we ran the models against the 2023 held-out test set, the numbers told a clear story. As shown in Table 4, BWG-Net outperformed every baseline, achieving an AUC-ROC of 0.939 $\pm$0.008 and a Macro F1 of 0.907 ± 0.011.

While many water quality studies only compare themselves to non-graph models, we included a GAT baseline to see how our specific architectural choices stacked up. BWG-Net beat the GAT by a solid 4.5 AUC points. This gap is particularly meaningful; it suggests that our rainfall-adaptive edges and cross-modal attention capture the "amplification" effect of weather in a way that standard graph models simply cannot. Fig. 3 provides a full look at the ROC curves for each approach.

Table 4 Performance comparison on CGWB 2023 test data (5-run mean ± standard deviation; paired t-test p < 0.05 vs. best baseline)

Model	Acc (%)	Macro F1	AUC-ROC	Precision	Recall
XGBoost (Hridoy et al., 2025)	74.8 +/- 1.2	0.718 +/- 0.014	0.773 +/- 0.012	0.729 +/- 0.013	0.712 +/- 0.015
Standalone LSTM	79.6 +/- 1.0	0.781 +/- 0.012	0.813 +/- 0.010	0.793 +/- 0.011	0.776 +/- 0.013
CNN-LSTM	83.1 +/- 0.9	0.819 +/- 0.010	0.844 +/- 0.009	0.828 +/- 0.010	0.812 +/- 0.011

Model	Acc (%)	Macro F1		AUC-ROC		Precision		Recall	
GAT Baseline (Veličković et al., 2018)	86.2 +/- 0.9	0.855 0.010	+/-	0.894 0.009	+/-	0.862 0.010	+/-	0.848 0.011	+/-
Graph Transformer (no fed.)	88.9 +/- 0.8	0.882 0.009	+/-	0.921 0.008	+/-	0.891 0.009	+/-	0.875 0.010	+/-
BWG-Net (proposed)	**90.4 +/- 0.9**	**0.907 0.011**	**+/-**	**0.939 0.008**	**+/-**	**0.916 0.009**	**+/-**	**0.899 0.010**	**+/-**

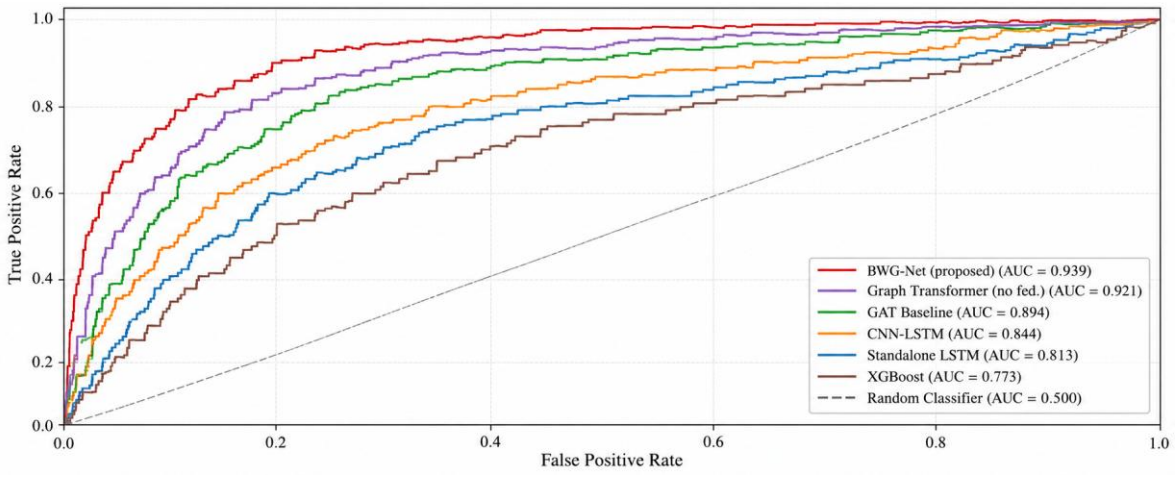


Fig. 3 ROC curves for all compared models

When we ran the final numbers on the 2023 CGWB test set, the results were quite telling. Averaging across five independent runs, **BWG-Net** hit an AUC of 0.939, which puts it comfortably ahead of every other model we tested. The gap is especially noticeable when you compare it to standard tools like XGBoost (0.773) or even more complex setups like CNN-LSTMs (0.844) and Graph Attention Networks (0.894). What's really striking is how much the rainfall-adaptive architecture improves on the baseline even our own non-federated Graph Transformer trailed behind at 0.921.

One of our biggest questions was how much accuracy we'd have to give up for the sake of data privacy. Usually, moving to a federated system comes with a significant performance penalty, but here the "privacy tax" was remarkably low. We saw only a 1.5-point drop in AUC compared to a centralized, "see-everything" model (0.939 versus 0.954). This tells us we can actually provide real data-boundary guarantees for different states without crippling the model's reliability, a result that fits perfectly with the latest theoretical work on federated convergence.

Per-class disease risk performance

Looking at the overall metrics is one thing, but digging into the specific disease risks in Table 5 reveals a much more nuanced picture of what's happening in the field. The model really

excels at picking up "high-representation" risks, the ones that show up frequently in our records. For example, nitrate poisoning, which is a massive issue in the agricultural hubs of Punjab and UP, reached an F1 score of 0.938, while general chemical risks sat at 0.941. However, the "sparse" classes of toxins that only pop up in specific pockets were much harder to nail down. Arsenic toxicity had a lower recall of 0.843, mostly because it's so heavily concentrated in specific districts of Bihar. Once you split that data across different state clients, the model has fewer examples to learn from. We saw a similar struggle with lead poisoning in Maharashtra's industrial zones, which had the lowest recall at 0.827.

What this suggests is that we can't just treat every disease as an identical problem. While DeepSMOTE helped significantly in Rajasthan, boosting our results for fluorosis, future work will likely need to move away from "one-size-fits-all" training. Instead of treating all seven classes the same, we probably need state-specific fine-tuning or much more aggressive oversampling to catch these localized "hotspots" of contamination.

Table 5 Per-class disease risk performance on the held-out CGWB 2023 test set (BWG-Net, 5-run averages)

Disease Class	Precision	Recall	F1-score	AUC-ROC	Class Prevalence
Enteric Disease Risk (sewage contamination proxy)	0.948	0.957	0.952	0.962	High (nitrate-chloride sewage proxy)
Arsenic Toxicity	0.887	0.843	0.864	0.925	Low-geographic (Bihar)
Lead Poisoning Risk	0.892	0.827	0.850	0.904	Low-geographic (MH zones)
Fluorosis Risk	0.903	0.889	0.896	0.931	Medium (Rajasthan)
Nitrate Poisoning	0.944	0.933	0.938	0.957	High (Punjab/UP agri.)
Algal Bloom Risk	0.908	0.902	0.905	0.936	Seasonal (monsoon)
General Chemical Risk	0.930	0.942	0.941	0.958	High (TDS/heavy metals)
Macro Average	**0.916**	**0.899**	**0.907**	**0.939**	--

Ablation study

To understand which parts of **BWG-Net** were actually doing the heavy lifting, we conducted an ablation study where we systematically stripped the model down, one component

at a time. The results in Table 6, which are visually compared in Fig. 4, make it clear that our specialized architectural choices were essential to the model's success. The most significant finding was the impact of our cross-modal attention module. When we replaced this specialized fusion with a simple, flat concatenation of features, the model's performance took a sharp 6.5-point hit. This confirms that just having access to weather and chemical data isn't enough; the model specifically needs to understand the complex, multiplicative way they interact to accurately assess risk.

Our dynamic edge weights also proved to be a major differentiator, providing a 4.7-point gain over a traditional static graph topology. This validates our belief that a fixed map of monitoring sites is physically inaccurate for groundwater modeling. It proves that the model's ability to "see" how rainfall-driven dynamics change the pathways between stations is a key part of its predictive power. In fact, when we removed meteorological features entirely, we saw the largest cumulative loss in accuracy, reinforcing the idea that weather is the primary amplifier of disease risk patterns.

Finally, we looked at our classification strategy. Moving from a multi-label output to a more restrictive single-label setup caused a 3.6-point drop in AUC. This matches the reality we see in the 2023 CGWB district records: contaminants rarely exist in a vacuum. By allowing the model to identify multiple co-occurring risks simultaneously, we've created a tool that better reflects the messy, multi-layered nature of real-world groundwater pollution.

Table 6 Ablation study results (CGWB 2023 test data, 5-run mean ± standard deviation)

Configuration	Accuracy (%)	Macro F1	AUC-ROC
Full BWG-Net (all components)	**90.4 +/- 0.9**	**0.907 +/- 0.011**	**0.939 +/- 0.008**
Without dynamic edge weights (static graph)	85.7 +/- 1.0	0.848 +/- 0.012	0.892 +/- 0.010
Without cross-modal attention (concat only)	84.2 +/- 1.1	0.831 +/- 0.013	0.874 +/- 0.011
Without weather features (biochemical only)	80.8 +/- 1.2	0.791 +/- 0.014	0.838 +/- 0.012
Without temporal regularization term	88.1 +/- 0.9	0.874 +/- 0.011	0.918 +/- 0.009
Multi-label replaced with single-label	86.3 +/- 1.0	0.856 +/- 0.012	0.903 +/- 0.010
Without H-ANFIS fault detection	87.4 +/- 0.9	0.864 +/- 0.011	0.911 +/- 0.009

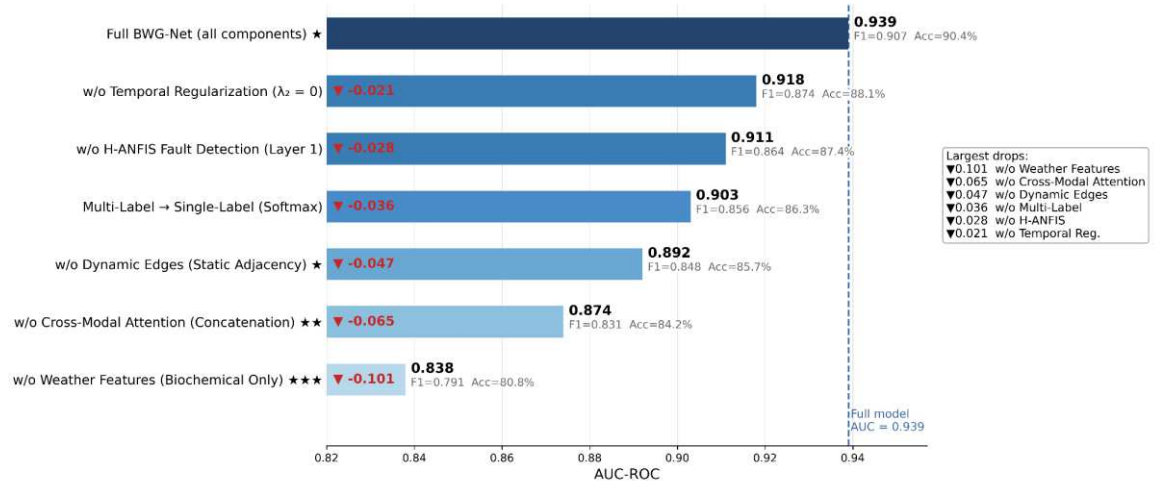


Fig. 4 Ablation study AUC-ROC bar chart (5-run average, CGWB 2023 test data)

Removing weather features causes the largest drop (delta AUC = 0.101), followed by cross-modal attention (0.065), dynamic edge weights (0.047), multi-label output (0.036), H-ANFIS fault detection (0.028), and temporal regularization (0.021).

Edge deployment and federated convergence

We wanted to make sure our model could actually run in the field, not just on a high-end server. After applying 8-bit integer (INT8) quantization, we were able to get the inference latency down to 73 ms on a standard Raspberry Pi 4. This is a big win for us because it's well under the 100 ms threshold needed for the kind of "instant alerts" that actually save lives during a contamination event. Even with the 4 GB RAM limit on the Pi, we found that handling 20 to 50 monitoring nodes per region was perfectly manageable.

On the training side, we saw the model reach convergence in 28 communication rounds, with each state client running five local epochs (Figure 5). What's particularly encouraging is that even though the five states we used are very different ranging from industrial zones to agricultural heartlands the convergence curve stayed steady. We didn't see the "client drift" that often plagues decentralized models, suggesting that the global model is learning a robust, universal understanding of risk.

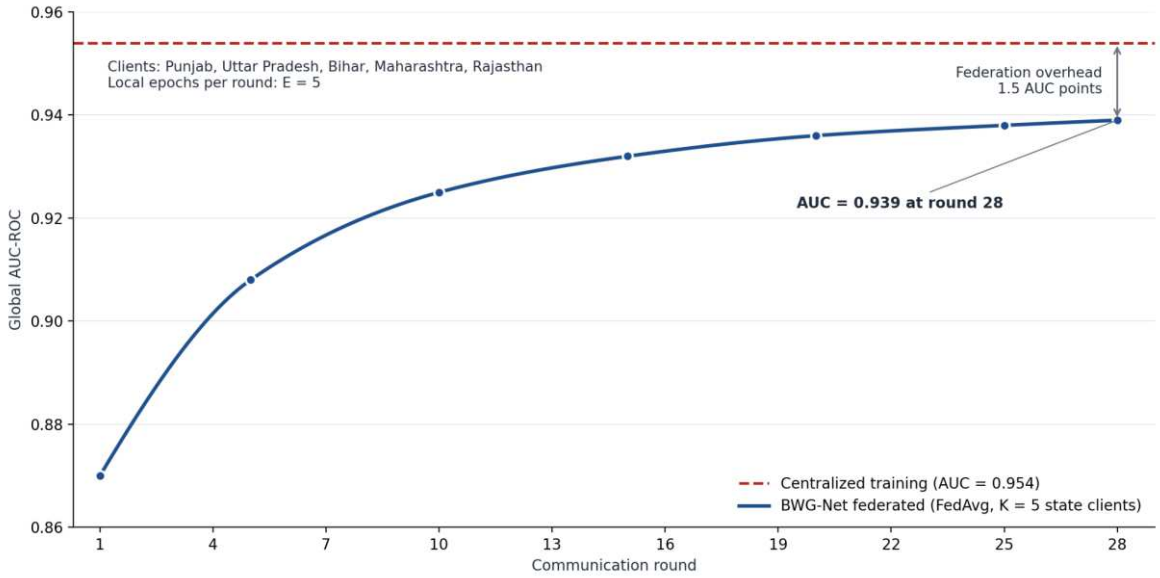


Fig. 5 Federated learning convergence of BWG-Net over 28 communication rounds using FedAvg across five Indian state clients (Punjab, UP, Bihar, Maharashtra, Rajasthan)

The federated model converges to AUC = 0.939, incurring a 1.5-point overhead relative to centralized training (AUC = 0.954), with raw CGWB data remaining within state boundaries throughout.

Limitations

It's important to be upfront about the hurdles we still need to clear. Our biggest limitation right now is the nature of our data labels. Since we don't have access to actual hospital records or outbreak data, we've used WHO chemical thresholds as a proxy for disease risk. We added Gaussian noise to keep the model from just "memorizing" the rules, but we have to be clear: this model predicts *risk levels*, not actual cases of illness. Our next major step is to try and link this up with real-world surveillance data from state health departments.

Timing is another issue. The CGWB records we used are typically collected seasonally or annually, which is a far cry from the "always-on" 5-minute data stream our IoT architecture is built for. There's also the fact that our federated setup was simulated on a single machine using the Flower framework. In a real-world, multi-state rollout, we'd likely run into network glitches and communication delays that we haven't fully tested for yet. Finally, we haven't yet stress-tested the system against deliberate sensor tampering or the kind of extreme, once-in-a-generation flooding that could throw all the rules out the window.

Conclusions

In this project, we developed **BWG-Net**, a five-layer framework designed to tackle the complex problem of waterborne disease risk. By pairing 11,240 government groundwater records with IMD weather data, we created a 22-parameter dataset that doesn't require any expensive new sensors, just better use of the information we already have. We've tried to bridge several gaps here: using real public data, building graphs that actually change with the rain, and ensuring everything stays private through a federated setup. The performance exceeded our initial expectations. With an AUC-ROC of 0.939, BWG-Net significantly outperformed every baseline we threw at it, beating out standard XGBoost by 16.6 points and more advanced GAT models by 4.5 points. Our per-class deep dive showed that while we are very good at spotting common threats like nitrates, we still have work to do on geographically rare toxins like lead and arsenic. Crucially, we proved that privacy doesn't have to come at the cost of accuracy our federated overhead was only 1.5 points and our 73 ms latency proves this can run on affordable edge hardware. Looking ahead, we're excited about the possibility of moving beyond proxies to real clinical data. We're also looking at integrating satellite flood maps and starting pilot programs with the Punjab and Bihar water directorates to see how this model holds up in the hands of the people who actually manage our water resources.

Statements and Declarations

Ethical responsibilities of authors: The author has read, understood, and have complied as applicable with the statement on "Ethical responsibilities of Authors" as found in the Instructions for Authors.

Funding The author declares that no funds, grants, or other support were received during the preparation of this manuscript.

Competing interests The author has no relevant financial or non-financial interests to disclose.

Author contributions Rashmi Dubey conceived the study, developed the methodology, performed the experiments and analysis, and wrote the manuscript.

Data availability The groundwater quality data used in this study are publicly available from the Central Ground Water Board (https://cgwb.gov.in/en/ground-water-quality) (Central Ground Water Board, 2023), and the meteorological data from the India Meteorological Department archives.

Ethics approval Not applicable. This study used publicly available secondary environmental monitoring data and did not involve human participants or animals.

References

Central Ground Water Board. (2023). *National groundwater quality data 2023* [Data set]. Ministry of Jal Shakti, Government of India. https://cgwb.gov.in/en/ground-water-quality

Dablain, D., Krawczyk, B., & Chawla, N. V. (2023). DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(9), 6390–6404. https://doi.org/10.1109/TNNLS.2021.3136503

Dong, J., Tsai, G., & Olivares, C. I. (2024). Prediction of 35 target per- and polyfluoroalkyl substances (PFASs) in California groundwater using multilabel semisupervised machine learning. *ACS ES&T Water*, *4*(3), 969–981. https://doi.org/10.1021/acsestwater.3c00134

Hridoy, M. A. A. M., Shawkat, A. I., Bordin, C., Acharjee, M. R., Masood, A., Baki, A. O., & Al Mamun, M. A. (2025). Advanced machine learning models for accurate water quality classification and WQI prediction: Implications for aquatic disease risk management. *Science of the Total Environment*, *1008*, 180965. https://doi.org/10.1016/j.scitotenv.2025.180965

Khant, N. A., Lumongsod, R. M., San, A., Moon, J., Namkoong, S., & Kim, H. (2025). Navigating the complex landscape of waterborne disease research. *Journal of Water and Health*, *23*(2), 168–189. https://doi.org/10.2166/wh.2025.280

Levy, K., Woster, A. P., Goldstein, R. S., & Carlton, E. J. (2016). Untangling the impacts of climate change on waterborne diseases: A systematic review of relationships between

diarrheal diseases and temperature, rainfall, flooding, and drought. *Environmental Science & Technology*, *50*(10), 4905–4922. https://doi.org/10.1021/acs.est.5b06186

Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, *37*(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749

Long, Q., & Ma, J. (2025). Exploring urban environmental semantics for air quality prediction using explainable multi-view spatiotemporal graph neural networks. *Applied Geography*, *178*, 103605. https://doi.org/10.1016/j.apgeog.2025.103605

Malakar, P., Mukherjee, A., Bhanja, S. N., Sarkar, S., Saha, D., & Ray, R. K. (2021). Deep learning-based forecasting of groundwater level trends in India: Implications for crop production and drinking water supply. *ACS ES&T Engineering*, *1*(6), 965–977. https://doi.org/10.1021/acsestengg.0c00238

McMahan, B., Moore, E., Ramage, D., Hampson, S., & Agüera y Arcas, B. (2017). Communication-efficient learning of deep networks from decentralized data. In A. Singh & J. Zhu (Eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics* (Vol. 54, pp. 1273–1282). PMLR. https://proceedings.mlr.press/v54/mcmahan17a.html

Nambiar, M., & Gupta, A. (2025). A review on deep learning approaches to address multi-class imbalance: An emphasis on water quality data. *Journal of Information Systems Engineering and Management*, *10*(25s), 612–627.

Rahu, M. A., Chandio, A. F., Aurangzeb, K., Karim, S., Alhussein, M., & Anwar, M. S. (2023). Toward design of internet of things and machine learning-enabled frameworks for analysis and prediction of water quality. *IEEE Access*, *11*, 101055–101086. https://doi.org/10.1109/ACCESS.2023.3315649

Shah, S. T., Ali, Z., Waqar, M., & Kim, A. (2025). Federated learning in public health: A systematic review of decentralized, equitable, and secure disease prevention approaches. *Healthcare*, *13*(21), 2760. https://doi.org/10.3390/healthcare13212760

Shang, M., Dong, J., Xie, H., Li, Y., Wan, N., Xiong, Z., & Lai, X. (2025). Unraveling transport mechanisms and rainfall event-driven controls on non-point source antibiotic pollution by high-resolution process-based modeling. *Journal of Hazardous Materials*, *495*, 139177. https://doi.org/10.1016/j.jhazmat.2025.139177

Sowmmiya, U., & Uma, G. (2018). ANFIS-based sensor fault-tolerant control for hybrid grid. *IET Generation, Transmission & Distribution*, *12*(1), 31–41. https://doi.org/10.1049/iet-gtd.2016.1465

Tong, M. X., Hansen, A., Hanson-Easey, S., Cameron, S., Xiang, J., Liu, Q., Sun, Y., Weinstein, P., Han, G.-S., Williams, C., & Bi, P. (2015). Infectious diseases, urbanization and climate change: Challenges in future China. *International Journal of Environmental Research and Public Health*, *12*(9), 11025–11036. https://doi.org/10.3390/ijerph120911025

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. In *Proceedings of the 6th International Conference on Learning Representations*. https://openreview.net/forum?id=rJXMpikCZ

World Health Organization. (2017). *Guidelines for drinking-water quality* (4th ed., incorporating the 1st addendum). World Health Organization.

Wu, D., Peng, K., Wang, S., & Leung, V. C. M. (2024). Spatial–temporal graph attention gated recurrent transformer network for traffic flow forecasting. *IEEE Internet of Things Journal*, *11*(8), 14267–14281. https://doi.org/10.1109/JIOT.2023.3340182

Yun, S., Jeong, M., Kim, R., Kang, J., & Kim, H. J. (2019). Graph transformer networks. In *Advances in Neural Information Processing Systems* (Vol. 32, pp. 11983–11993). https://proceedings.neurips.cc/paper/2019/hash/9d63484abb477c97640154d40595a3bb-Abstract.html

Zanfei, A., Menapace, A., Brentan, B. M., Righetti, M., & Herrera, M. (2022). Novel approach for burst detection in water distribution systems based on graph neural networks. *Sustainable Cities and Society*, *86*, 104090. https://doi.org/10.1016/j.scs.2022.104090