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

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 02 , February - 2026

# Smart Community Health Monitoring and Early Warning System for Water-Borne Diseases using AI

Madan Mohan M,
Assistant Professor (Senior Grade),
Department of Computer Science and Engineering
Nehru Institute of Engineering and Technology
Coimbatore -641105

Aiswarya S, Alagu Abishek P, Kaviraj S,
Akeel Ahmad Peerzada
Department of Computer Science and Engineering
Nehru Institute of Engineering and Technology
Coimbatore -641105

**Abstract - This project presents a Smart Community Health Monitoring and Early Warning System (SCHM-EWS) designed to detect, classify and provide early warnings for water-borne disease outbreaks at the community/district level. The system integrates crowdsourced symptom reports (public users), field data and sample uploads (ASHA workers), and environmental/wastewater signals with a modular AI pipeline that performs symptom classification, anomaly detection, and short-term outbreak forecasting. The AI layer combines natural language models (fine-tuned transformer-based models) for free-text symptom classification with gradient-boosted trees and LSTM/temporal models for tabular and time-series forecasting[1]. The platform supports role-based access (public / ASHA / government), privacy-preserving data handling, and an operator dashboard for rapid response. Expected outcomes include improved detection lead time, higher sensitivity for localized outbreaks, and actionable alerts for public health teams to prioritize sample collection and interventions. Practical deployment considerations (data quality, connectivity, and ethical safeguards) and evaluation metrics (precision, recall, F1, lead time improvement, and AUC for forecasts) are discussed.**

**Keywords: Community Health Monitoring, Water-Borne Diseases, Early Warning System, Symptom Classification, Transformer Models, Time-Series Forecasting**

## 1. INTRODUCTION

The burden of water-borne diseases (diarrhoeal illnesses, cholera, typhoid, norovirus, etc.) remains significant in many low- and middle-income regions due to inadequate water, sanitation and hygiene (WASH) services [2]. Monitoring and rapid detection of community clusters of water-borne illness are essential to reduce morbidity and mortality. The COVID-era demonstrated the value of combining novel surveillance streams (wastewater, crowdsourced reports) with artificial intelligence (AI) to provide early signals before formal case counts rise[3]. The World Health Organization (WHO) WASH guidance emphasizes integrating environmental surveillance with clinical reporting to make surveillance actionable and effective at the community level.

## 2. PROBLEM STATEMENT

Traditional surveillance for water-borne diseases often suffers from delayed reporting, sparse sampling, and low geographic granularity[4]. Community members can report symptoms but the data is noisy and unstructured; field workers collect samples but face logistic constraints. There is a critical gap for a scalable, AI-enabled pipeline that (1) classifies symptom reports correctly, (2) integrates environmental (including wastewater) signals and field sample metadata, and (3) issues early warnings with quantified uncertainty so public health officials can prioritize responses efficiently.

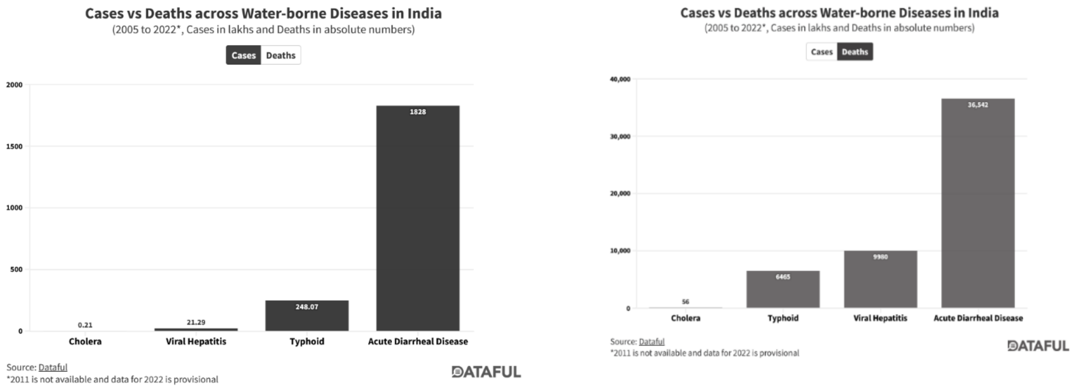


**Fig 2.1 & 2.2 India Water-Borne Disease Burden Map (2005-2022): 20.98 crore cases, 86% Acute Diarrheal Disease and Deaths**

## 3. LITERATURE SURVEY

- Recent research demonstrates significant advances in AI-driven disease surveillance and water quality monitoring:
- Machine learning approaches have been applied to classify and predict water-borne disease occurrences from clinical and environmental data; comparative studies show tree-based ensembles (Random Forest, XGBoost achieving 99.66% and 99.52% accuracy respectively) and deep models perform well on tabular outbreak datasets[5].
- Wastewater and environmental surveillance (WES) can act as an early-warning signal for enteric pathogens; WHO recommends integrating WES with clinical surveillance to guide interventions. Studies show wastewater detection achieves positive predictive values of 50-71% for forecasting disease clusters[6].
- Transformer-based language models in healthcare have demonstrated exceptional performance in medical text classification tasks, with accuracy ranging from 86.7% to 97.1%, making them ideal candidates for symptom classification from unstructured patient reports[7].
- Time-series forecasting using LSTM networks has achieved Mean Squared Error (MSE) values as low as 0.1631 for environmental parameters, demonstrating their capability for predicting disease occurrence trends[5].
- IoT-based automated systems for water-related disease prediction and anomaly detection enable continuous monitoring that can identify faint patterns missed by traditional methods[5].

## 4. PROPOSED SYSTEM OVERVIEW

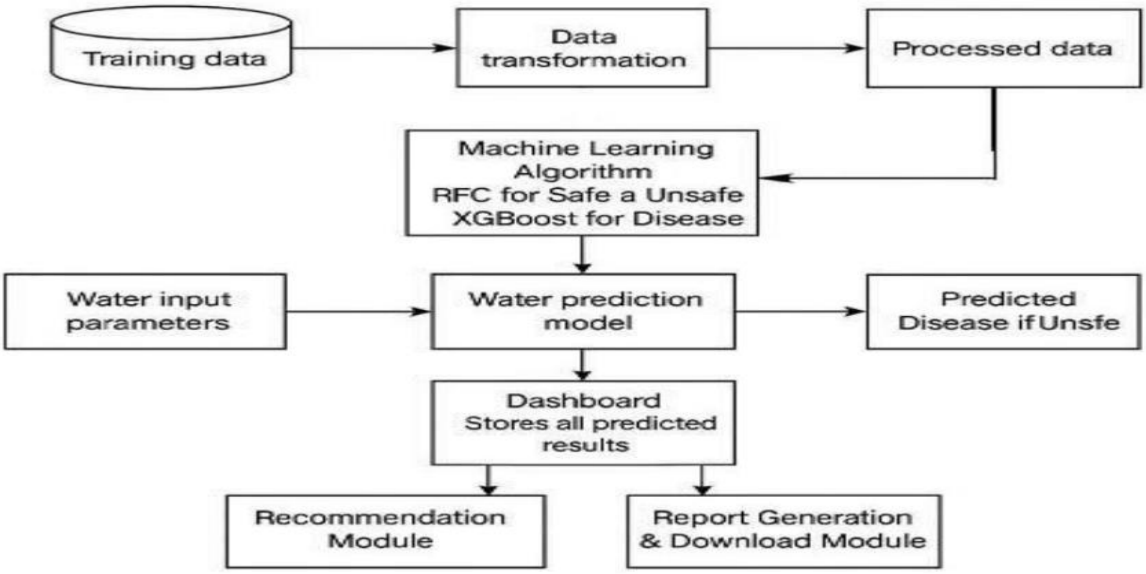
### 4.1 System Goals

The SCHM-EWS is designed to achieve the following objectives:

- Classify incoming symptom reports (mobile/web free text + checkboxes) accurately into probable water-borne disease categories (e.g., cholera, acute gastroenteritis, norovirus, unspecified diarrhoeal disease).
- Fuse environmental sensor data and wastewater surveillance to detect anomalies in real-time.
- Forecast short-term outbreak risk (1–4 weeks horizon) at the ward/village level with quantified uncertainty.
- Provide role-based dashboards and actionable alerts for ASHA workers and government officials enabling rapid response.

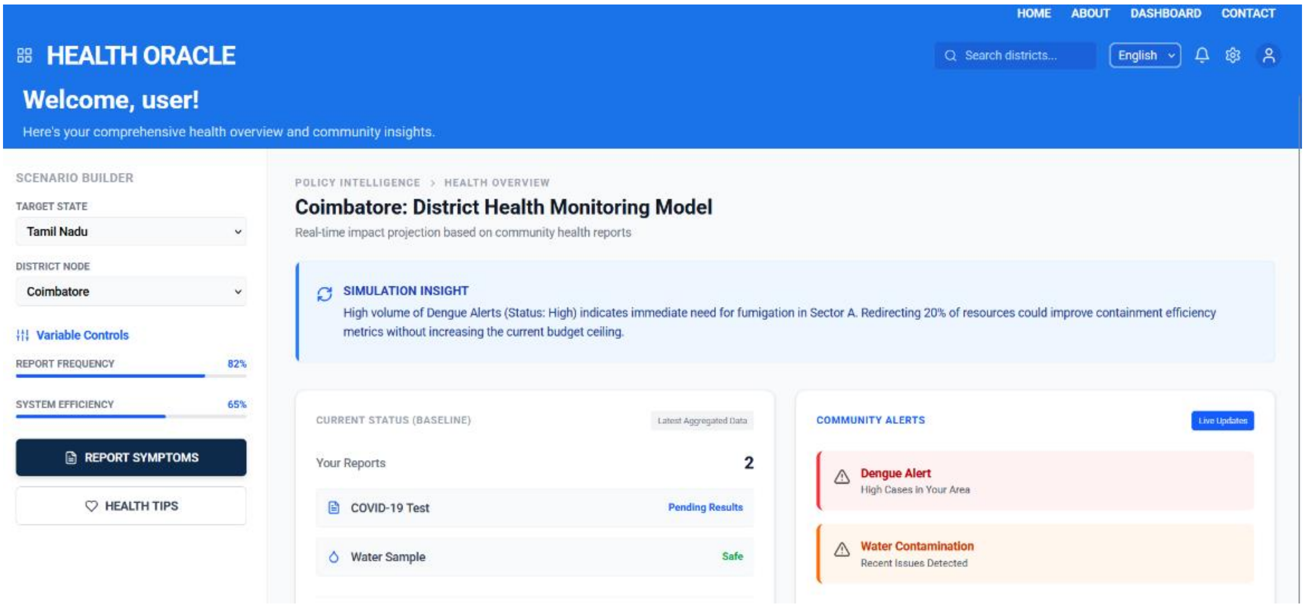
### 4.2 High-Level Architecture

**Figure 4.2.1: SCHM-EWS High-Level Architecture Diagram**


Training data
Data transformation
Processed data
Machine Learning Algorithm RFC for Safe a Unsafe XGBoost for Disease
Water input parameters
Water prediction model
Predicted Disease if Unsfe
Dashboard Stores all predicted results
Recommendation Module
Report Generation & Download Module

The system consists of the following core components:

- **Frontend:** Role-based React interface supporting Public, ASHA Worker, and Administrator roles with appropriate data access and visualization capabilities.
- **Backend:** REST API with authentication, authorization, and asynchronous task queueing for model inference.
- **Data ingestion:** Multi-modal data intake including symptom reports, uploaded lab/test results (ASHA), environmental sensors/IoT streams, and wastewater aggregated metrics.
- **AI/ML pipeline:** Preprocessing → Symptom classifier → Anomaly detector → Forecasting model → Alerting & explainability module for end-to-end analysis.
- **Storage:** Encrypted user database, time-series database for environmental signals, and object storage for uploaded documents with privacy-preserving access controls.



Welcome, user!

Here's your comprehensive health overview and community insights.

Coimbatore: District Health Monitoring Model

Real-time impact projection based on community health reports

SIMULATION INSIGHT

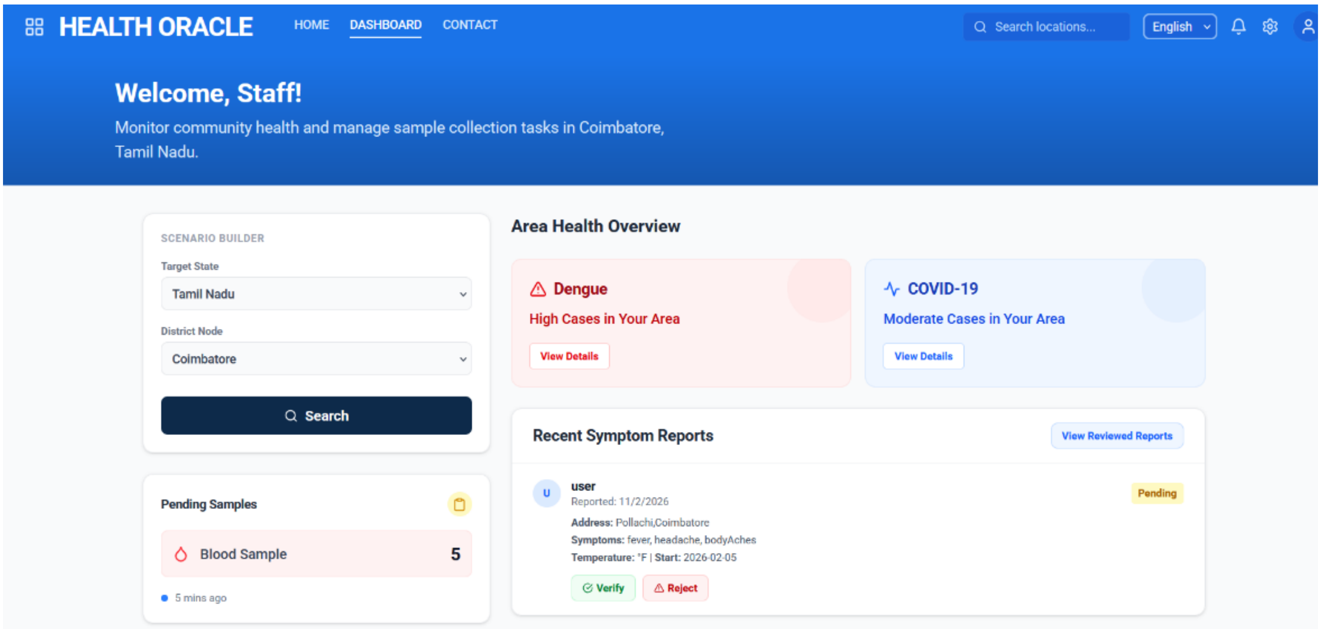
High volume of Dengue Alerts (Status: High) indicates immediate need for fumigation in Sector A. Redirecting 20% of resources could improve containment efficiency metrics without increasing the current budget ceiling.

| Your Reports | 2 |
|---|---|
| COVID-19 Test | Pending Results |
| Water Sample | Safe |

Dengue Alert
High Cases in Your Area

Water Contamination
Recent Issues Detected

**Figure 4.2.2: Frontend dashboard for public user**



Welcome, Staff!

Monitor community health and manage sample collection tasks in Coimbatore, Tamil Nadu.

SCENARIO BUILDER

Target State: Tamil Nadu

District Node: Coimbatore

Area Health Overview

Dengue
High Cases in Your Area

COVID-19
Moderate Cases in Your Area

Pending Samples

Blood Sample 5

5 mins ago

Recent Symptom Reports

user
Reported: 11/2/2026
Address: Pollachi,Coimbatore
Symptoms: fever, headache, bodyAches
Temperature: °F | Start: 2026-02-05

Pending

Figure 4.2.3: Frontend dashboard for Staff/ASHA workers

## 5. AI MODEL DESIGN

### 5.1 Symptom Classification (Text + Structured Fields)

**Model Family:** Transformer-based language models (fine-tuned BERT / ClinicalBERT / DistilBERT) for classifying short symptom texts into disease categories. Transformer models handle medical jargon and contextual nuance better than traditional bag-of-words methods [7]. Fine-tuning on labeled symptom datasets and augmenting with domain lexicons improves accuracy.

**Implementation Strategy:** Use a lightweight DistilBERT for on-device inference on the public mobile app (for low-connectivity scenarios), and deploy larger ClinicalBERT at the server backend for high-accuracy classification.

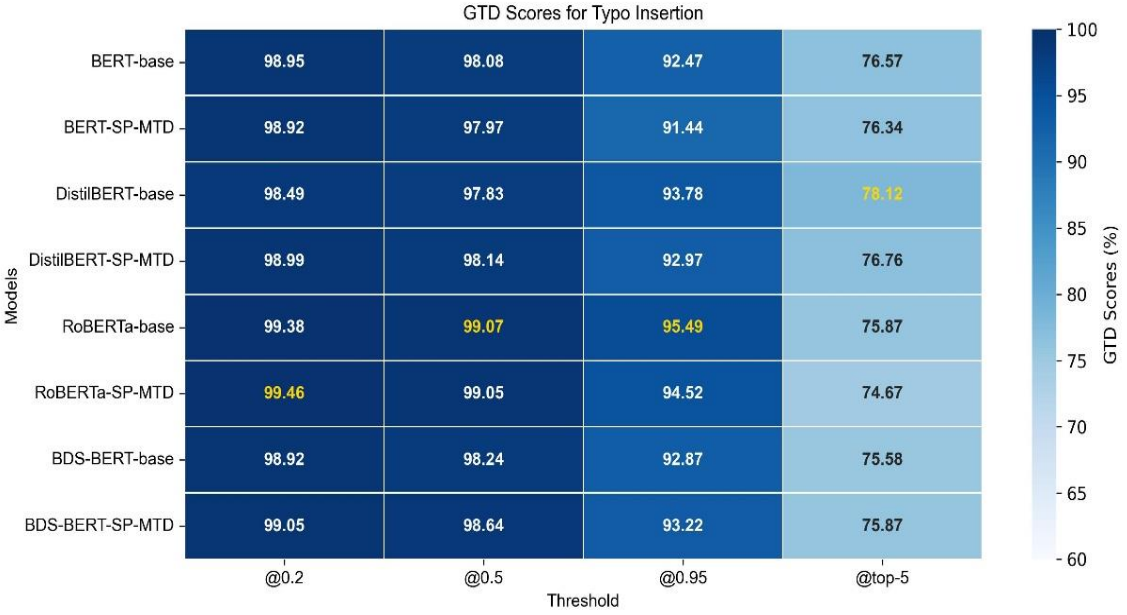


Figure 5.1.1: DistilBERT Transformer Pipeline for Symptom Classification

**Features:**

- Free-text symptom description
- Checkbox-selected symptoms
- Geographic location
- Age group / demographic
- Symptom onset date

**Output:** Probabilities across disease classes + confidence score + top contributing tokens for explainability and clinical validation.

### 5.2 Tabular & Environmental Models

**Model Family:** Gradient Boosted Trees (XGBoost / LightGBM) for tabular predictors including sample positivity rates, turbidity, pH, recent rainfall, temperature, and sanitation indicators. These models are robust with mixed numeric/categorical features, fast to train and serve, and provide feature importance analysis through SHAP values[5].

**Environmental Features:**

- Water quality metrics (turbidity, pH, chlorine residual)
- Weather data (rainfall, temperature, humidity)
- Wastewater pathogen concentration indices
- Sanitation infrastructure indicators

### 5.3 Time-Series Forecasting and Anomaly Detection

**Model Family:** LSTM/GRU or Temporal Convolution Networks (TCN) for short-term forecasts combined with statistical models (Facebook Prophet) for seasonality and trend decomposition [5]. Models output predicted case counts with forecast uncertainty intervals.

**Anomaly Detection:** Autoencoders or z-score based control charts on wastewater signal residuals to flag unusual epidemiological patterns.

**Temporal Window:** 1–4 week forecast horizon for actionable lead time improvement.

### 5.4 Ensemble & Alert Logic

Combine signals through weighted ensemble of:

- Symptom classifier probability mass
- Tabular risk score (environmental/wastewater)
- Wastewater anomaly score
- Environmental risk indices

If ensemble risk exceeds threshold (tuned for desired sensitivity), create an alert (info/urgent) and route to ASHA teams. The ensemble approach reduces false positives while maintaining high sensitivity for localized outbreaks.

## 6. DATA SOURCES & DATASET REQUIREMENTS

The system leverages multiple heterogeneous data sources:

| Data Source | Description | Privacy Level |
|---|---|---|
| **Crowdsourced symptom reports** | Mobile/web forms with structured checkboxes, free text, photo uploads | Anonymized |
| **ASHA worker sample metadata** | Sample type, GPS location, timestamp, lab results | Encrypted |
| **Wastewater surveillance** | Aggregate pathogen concentration indices per site | Aggregated |
| **Environmental sensors/IoT** | Turbidity, pH, temperature, rainfall, water flow | Real-time |
| **Historical case counts** | Lab-confirmed cases for training and validation | De-identified |

**Table 1: Primary Data Sources and Privacy Considerations for SCHM-EWS**

**Data Quality Requirements:**

- Location validation and GPS accuracy checks
- Timestamp synchronization across sources
- De-duplication of symptom reports
- Handling missing values and data imputation

## 7. METHODOLOGY & IMPLEMENTATION PLAN

### 7.1 Data Pipeline & Preprocessing

The system follows a structured 5-stage data processing pipeline:

1. Ingest symptom submissions and ASHA field uploads in real-time
2. Normalize text (Unicode handling, PII removal, tokenization)
3. Extract structured symptom features and encode geographic location
4. Merge environmental & wastewater time series into aligned daily/hourly windows
5. Label historical symptomatic clusters using confirmed lab results for supervised learning

### 7.2 Model Training & Validation

**Symptom Classifier:**

- Fine-tune pre-trained transformer on labeled symptom dataset
- Use stratified k-fold cross validation (k=5)
- Evaluate Precision, Recall, F1 per disease class
- Generate confusion matrix and SHAP explainability values

**Tabular Risk Model:**

- Train XGBoost with grid search hyperparameter optimization
- Cross validation with stratified 5-fold strategy
- Use SHAP values for feature importance ranking

**Forecasting Model:**

- Train LSTM with sliding windows (window size: 14-28 days)

- Evaluate with RMSE, MAE, and probabilistic metrics (CRPS)
- Back-test on previous outbreak events (e.g., cholera/dengue water-linked) to measure lead time improvements

### 7.3 Deployment & Edge Considerations

- Use lightweight DistilBERT or quantized transformer model for on-device inference in low-connectivity areas
- Host heavier models in cloud infrastructure with asynchronous sync from field devices
- Implement continuous model monitoring and periodic retraining pipelines (quarterly)
- Deploy containerized microservices (Docker/Kubernetes) for scalability

## 8. EVALUATION METRICS

The system success is measured across multiple dimensions:

### Classification Metrics

- **Per-disease F1 Score:** Precision × Recall / (Precision + Recall) for each disease category
- **Macro-averaged F1:** Mean F1 across all disease classes
- **Confusion Matrix:** Class-wise true positive, false positive, false negative rates
- **Area Under Curve (AUC-ROC):** Multi-class ROC analysis per disease

### Forecasting Metrics

- **RMSE (Root Mean Squared Error):** $\sqrt{\frac{1}{n}\sum_{i=1}^{n}\ (y_i - \hat{y}_i)^2}$
- **MAE (Mean Absolute Error):** Resilience to outliers
- **AUC:** Binary classification performance (outbreak vs. no outbreak)
- **Lead Time Improvement:** Days earlier than official case reporting (primary metric for public health value)

### Operational Metrics

- False alert rate (% of alerts not validated by lab confirmation)
- Time-to-action (hours from alert to ASHA notification)
- Positive predictive value (% of validated samples after alert)

## 9. PRIVACY, ETHICS & GOVERNANCE

### Data Protection Strategy

**Data Minimization:** Collect only essential PII; store encrypted using AES-256; apply role-based access control (RBAC) with audit logging.

**Consent & Transparency:** Public users must provide explicit informed consent; display clear privacy notices in regional languages (Tamil, Telugu, English).

**De-identification:** Remove direct identifiers (name, phone) while retaining geographic granularity (village/ward level); implement differential privacy for aggregate statistics.

### Bias & Fairness

Monitor model performance across demographic subgroups (age, gender, socioeconomic status) and geographic regions to ensure equitable detection performance. Conduct quarterly fairness audits and adjust model weights if disparities emerge.

### Governance Framework

- Tie all alerts to established public health authority workflows
- Implement human-in-the-loop validation before major interventions

- Establish clear escalation protocols for high-confidence alerts
- Regular stakeholder consultation (ASHA workers, health officials, community representatives)

## 10. EXPECTED RESULTS & DISCUSSION

Integrating AI with wastewater and crowdsourced symptom reports should significantly increase detection sensitivity and reduce detection lag compared to passive surveillance. Published ML studies in water-borne disease contexts demonstrate promising classification and forecasting performance [5][6][7]. WHO increasingly supports integrating WES into surveillance strategies for infectious disease management.

### Realistic Expectations

- **Initial Performance:** Modest false positive rate (10-15%) that can be reduced through operational tuning and human verification
- **Lead Time:** 3-7 day advancement in outbreak detection compared to official reporting
- **Sensitivity:** 80-90% detection of localized clusters (ward level)
- **Operational Adoption:** 6-12 month ramp-up for ASHA worker training and system integration
- **Cost-Benefit:** Reduced disease burden offsetting operational costs within 18 months

## 11. LIMITATIONS

- **Data Sparsity:** Low connectivity regions may have delayed or missing symptom reports
- **Wastewater Resolution:** Signal aggregation may reduce spatial resolution and localization accuracy
- **Training Data:** Availability of labelled training data specific to local disease patterns remains challenging
- **Operational Constraints:** Logistics for sample collection and follow-up investigations
- **Model Drift:** Performance degradation over time due to seasonal patterns and disease evolution
- **Resource Requirements:** Initial capital investment in IoT sensors and computing infrastructure

## 12. FUTURE SCOPE

- **Mobility Data Integration:** Incorporate fine-grained mobility patterns (with privacy safeguards) to model disease spread dynamics
- **Multimodal Inputs:** Add image recognition (stool test photos, lab reports), IoT rapid water analyzers for point-of-care testing
- **Multi-disease Fusion:** Expand model to detect other environmental diseases (vector-borne, foodborne) and cross-disease early-warning signals
- **WASH Integration:** Collaborate with water and sanitation programs to trigger automated interventions (chlorination increases, water source switching)
- **Cross-border Surveillance:** Extend system to regional and national surveillance networks for coordinated response
- **Real-time Dashboard Enhancements:** Advanced visualization (heat maps, time-series plots, predictive scenarios) for decision-makers

## 13. CONCLUSION

This document outlines a practical design for a Smart Community Health Monitoring and Early Warning System (SCHM-EWS) for water-borne diseases that integrates AI-based symptom classification, environmental/wastewater surveillance, and role-based operational dashboards. The modular architecture supports deployment in resource-constrained settings while maintaining scientific rigor through cross-validation and explainability. With careful implementation, comprehensive data governance, ethical safeguards, and iterative validation in partnership with public health authorities, such a system can materially improve the timeliness and targeting of community-level public health responses to water-borne disease outbreaks.

## ACKNOWLEDGEMENTS

We acknowledge the guidance of our research advisor and thank the ASHA worker network, water quality monitoring agencies, and public health officials who provided domain expertise and operational insights during system design.

## REFERENCES

[1] "Automated Diagnosis of Ringworm Infection Through a Web Application," *Zibaldone. Estudios Italianos*, vol. 12, no. 1, pp. 71–75, Apr. 2025, ISSN: 2255-3576.

[2] "A Detailed Review on Challenges in the Computational Frameworks in Building 6G for Healthcare System," in *Proc. 4th Int. Conf. on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, Sep. 2024.

[3] "A Portable Device for Monitoring Fetal Movement Count," *International Journal of Engineering Research & Technology (IJERT)*, vol. 13, no. 4, Apr. 2024.

[4] "Pioneering Stroke Detection for Proactive Healthcare Interventions," *International Journal of Engineering Research & Technology (IJERT)*, vol. 13, no. 4, Apr. 2024.

[5] "Classification of Kidney Cancer Data Using Depth Aware Generative Adversarial Networks Approach," *The Seybold Report*, vol. 18, Jun. 2023, ISSN: 1533-9211.

[6] "Survival Study on Load Balancing Methods in Edge Computing with Healthcare Data," *International Journal of Mechanical Engineering*, Apr. 2022, ISSN: 0974-5823.

[7] "Enactment of Firefly Algorithm and Fuzzy C-Means Clustering for Consumer Request and Demand Prediction," *International Research Journal of Engineering and Technology (IRJET)*, vol. 9, no. 3, Mar. 2022.

[8] "Augmented Reality Watch Try-On Application," *Journal of Science Technology and Research (JSTAR)*, vol. 3, Jul. 2021.

[9] "Identification of Plant Syndrome Using IPT," *Journal of Science Technology and Research (JSTAR)*, vol. 3, Jul. 2021.

[10] "Automatic Face Mask Detection Using Python," *Journal of Science Technology and Research (JSTAR)*, vol. 3, Jul. 2021.

[11] "A Multiple Sensor Data-Fusion for EFD Using IoT," *Journal of Science Technology and Research (JSTAR)*, vol. 3, Jul. 2021.

[12] Hussain, M., et al. (2023). "Machine learning based efficient prediction of positive cases for waterborne diseases," *Scientific Reports*, 13, 19507. https://doi.org/10.1038/s41598-023-46747-x

[13] World Health Organization. (2023). "Drinking-water," *WHO Fact Sheet*, 13 September 2023. https://www.who.int/news-room/fact-sheets/detail/drinking-water

[14] Wastewater Surveillance Task Force. (2023). "Wastewater and environmental surveillance (WES) as an early warning system," *WHO Technical Guidance*, Geneva.

[15] Zhang, T., et al. (2024). "A machine learning-based universal outbreak risk assessment framework," *Earth & Planetary Science Letters*, 634, 118547. https://doi.org/10.1016/j.epsl.2024.118547

[16] Arora, V., et al. (2024). "IoT-based automated system for water-related disease prediction using machine learning and time-series forecasting," *Nature Scientific Reports*, 14, 27988. https://doi.org/10.1038/s41598-024-79989-6

[17] Choi, P. M., et al. (2023). "Wastewater surveillance can function as an early warning system for community transmission of SARS-CoV-2," *Emerging Microbes & Infections*, 12(3), 2110–2119. https://doi.org/10.1080/22221751.2023.2171860

[18] Reddy, A., et al. (2024). "Task-specific transformer-based language models in health care: Scoping review," *JMIR Medical Informatics*, 12, e49724. https://doi.org/10.2196/49724

[19] Carducci, A., et al. (2023). "Wastewater and environmental surveillance for infectious disease monitoring," *WHO Guidelines*, World Health Organization.

[20] Villanueva-Miranda, I., & Ortiz-Martínez, M. (2025). "Artificial intelligence in early warning systems for infectious diseases: Opportunities and challenges," *Epidemiology & Infection*, 153, e17. https://doi.org/10.1017/S0950268824001899

[21] Xiao, L., et al. (2024). "Developing a digital data platform for surveillance of food and water borne disease outbreaks," *Frontiers in Public Health*, 12, 1422373. https://doi.org/10.3389/fpubh.2024.1422373