INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

BIG DATA FRAMEWORK FOR IDENTIFYING EMERGING DISEASES FROM WASTEWATER ANALYTICS

[1]Mr SIDDESH K T, [2]SUJATA P HAVERI
[1]Assistant Professor, Department of MCA, BIET, Davenagere
[2]Student, Department of MCA, BIET, Davanagere

ABSTRACT

Proactive Outbreak Defense Integrating Wastewater Testing with Big Data Effective disease management relies on speed. By the time a patient enters a clinic, the virus has often already established a foothold in the local population. To counter this, our initiative focuses on Wastewater- Based Epidemiology (WBE) a method that treats the city's sewage as a collective biological mirror. The Surveillance. Framework Unlike traditional testing, which is reactive, monitoring wastewater allows us to catch "silent" viral shedding from asymptomatic or pre-symptomatic individuals. Environmental Sampling We don't just "measure" water; we extract and sequence genetic material (RNA/DNA) from residential and industrial runoff. Using localized sensors, we can pinpoint which specific neighbourhoods or facilities. are showing rising concentrations of pathogens like Norovirus or SARS-CoV-2.

The Big Data Processing Layer Raw biological data is noisy and massive. We use distributed computing frameworks to ingest these streams, comparing current levels against historical baselines. Such a measure helps in differentiating. between a normal seasonal change and a real medical crisis. Intelligent Alerting What our platform delivers is not a spreadsheet; it provides intelligence. Upon confirmation of the existence of the spike, reports are automatically generated. This could giveway to a localized vaccine campaign or a warning for the respective ERs in the specific zip code in question.

Keywords *: Wastewater-Based Epidemiology (WBE), Big Data Analytics, Real-Time Pathogen Monitoring.Early Warning System, Predictive Healthcare, Intelligent Alerting, Public Health Intelligence..*

1.INTRODUCTION

Within the ever-changing international framework for health, speed is this also a primary factor in saving lives Our project uses. Wastewater-Based Epidemiology (WBE) - an active surveillance tool that considers athe community's sewage system as a communal biological census. Instead of waiting for patients who show symptoms to come into hospitals for medical attention, we examine the microscopic traces. of infection left behind when patients touch certain surfaces and thereby pathogens: viruses, bacteria, and chemical markers released into the wastewater system from society.

This facilitates a broad anonymized. health perspective on whole neighbourhoods without the practical burden of individual clinical testing. WBE is a verified concept in theory, but its major slowdown is in the amount and noise of data that is being generated. Every sample will represent a complex combination of viral RNA, dynamic levels of various chemicals, and environmental factors such as sunlight exposure that. can influence the variables such as temperature and flow rate. This processing would be impossible by hand. Our project bridges this gap by. incorporating a Big Data Framework, which is aimed at processing raw effluent data to actionable public health intelligence.

The venture is divided into two key operational segments. Pathogen Identification & Wastewater Analysis Weit will emphasize the systematic collection and. molecular characterization of samples of sewage. With the use of advanced

sensors and laboratory tests, we measure the pathogen level in the samples. By identifying certain genetic signatures of emergent diseases, it becomes possible. to detect the existence of a virus likeSARS-CoV-2 or new influenza strains two weeks prior to the onset of the surge in hospital admissions.

In order to cope with the enormous volume of data. pertaining to the environment that we are receiving daily, we are using. a distributed computing stack that uses Apache Spark and Hadoop. Such tools enable us to IngestMulti-SourceData Integrate laboratory data with weather patterns and demographicsin real time. Predictive Modeling Use. machine learning algorithms to identify abnormal spike patterns that are outliers tohistorical baselines. Data Visualization Synthesize complex analytics into intuitive dashboards.The Goal By the point at which our system alerts a trend, health authorities are enabled to launch specific medicalthrough resources or public. warning systems that effectively bring outbreaks to a halt.

2. LITERATURE REVIEW

Wastewater-Based Epidemiology (WBE) is widely used to monitor community health by detecting pathogens in wastewater. Existing systems mainly rely on manual sample collection and PCR-based laboratory analysis, which, although accurate, are reactive and time-consuming. Detection often occurs only after widespread community transmission, reducing the effectiveness of early intervention.

Previous studies highlight major limitations such as high dependency on skilled manpower, delayed reporting, and threshold-based analytics that fail to adapt to variations caused by rainfall, industrial discharge, and population behavior. Most current systems are PCR-centric and pathogen-specific, making them ineffective in detecting emerging or mutated diseases.

Literature also points out that centralized monitoring provides only city-level insights and lacks spatial resolution. Conventional Big Data tools are not optimized for noisy and unstructured wastewater data, and existing dashboards focus mainly on descriptive analysis rather than prediction.

Recent research emphasizes the need for IoT-based sensing, distributed Big Data platforms, and machine learning models to enable real-time monitoring, predictive analytics, and early warning systems. These findings support the development of an advanced, scalable, and proactive wastewater analytics framework for emerging disease detection.

3. EXISTING SYSTEM

Current environmental pathogen surveillance systems are largely reactive rather than predictive, detecting outbreaks only after widespread transmission has occurred. Traditional methods depend on manual wastewater sampling and laboratory-based PCR analysis, which, despite high accuracy, involve delays in collection, transport, and processing. These delays reduce real-time responsiveness and often lead to detection after infection peaks, while the reliance on skilled manpower makes continuous monitoring costly and impractical for most municipalities.

Most existing systems employ heuristic and threshold-based analytics, triggering alerts only when pathogen concentrations exceed predefined limits. However, wastewater composition varies substantially due to rainfall, industrial effluents, population behavior, and storm events, making static thresholds unreliable and leading to false positives or missed outbreak signals. Furthermore, current surveillance infrastructures are largely. PCR-centric and focused on known pathogens such as Polio or SARS-CoV, using specific primers that detect only predefined targets. As a result, these systems remain blind to emerging or mutated pathogens, while the high cost of specialized reagents and the requirement for trained personnel limit large scale and continuous screening across multiple pathogens.

Current environmental monitoring systems are centralized and low-resolution, providing only city-level insights and lacking neighborhood-level detail. Wastewater data is highly noisy due to factors like rainfall, industrial interference, and temperature effects, while traditional Big Data

tools are not designed to handle such unstructured and dynamic data. Existing architectures also lack proper preprocessing and normalization layers, reducing decision-making accuracy. Additionally, most public health dashboards offer only descriptive analytics, with limited predictive capabilities, hindering the development of scalable, early-warning surveillance systems.

4. PROPOSED SYSTEM

The proposed Emerging Diseases from Wastewater Analytics overcomes the limitations of traditional wastewater surveillance by combining big IoT-based sensors, and predictive analytics in a unified architecture.

4.1.1 Proposed System: Advanced Big Data Architecture for WBE Our framework represents a fundamental shift from traditional, reactive laboratory methods toward a proactive, real-time surveillance ecosystem. By integrating automated data pipelines with continuous environmental monitoring, the system provides a comprehensive "early warning" shield for public health.

4.1.2 High-Throughput Data Ingestion and Processing The system uses a distributed architecture built on Apache Hadoop and Apache Spark, with Apache Kafka enabling real-time data streaming from thousands of sensors. Edge intelligence is applied at sampling sites to perform initial pathogen analysis locally, reducing latency and cloud workload for faster detection. To improve accuracy, the platform uses multi-source data fusion, combining wastewater signals with climate data, population density, and hospital trends to reduce environmental noise and enhance precision.

4.1.3 Advanced Predictive Analytics The platform uses advanced analytics to detect meaningful patterns in complex wastewater data through three layers: anomaly detection to identify deviations from normal community baselines, pathogen classification using deep learning to recognize viral and microbial signatures, and time-series forecasting to predict outbreaks before clinical cases appear. The system supports adaptive learning, allowing models to evolve with new pathogen strains or chemical changes without restarting the system. A secure cloud layer enables inter-agency collaboration through anonymized and aggregated data sharing, ensuring strong privacy protection. Overall, this scalable, data-driven solution reduces reliance on slow laboratory testing and enables a resilient, proactive public-health defense system.

System Architecture

Fig 1. System Architecture

5. METHODOLOGY

The proposed methodology for the Emerging Disease from Wastewater Analytics integrates environmental sampling, big data engineering, machine learning, and intelligent decision-making techniques to achieve accurate and timely disease surveillance.

5.1.1 System Workflow and Data Orchestration The operational lifecycle of the framework is a multi-stage transformation that converts raw, chaotic environmental samples into refined, actionable public health intelligence. This process is engineered to handle the high volatility of biological data while maintaining sub-second analytical speeds.

5.1.2 Ingestion and Signal Refinement The workflow begins at municipal treatment plants or community hubs, where raw biological and chemical wastewater data is collected and converted into a processable format. Noise caused by factors such as chemical runoff or rainfall dilution is removed to ensure reliable analysis. The cleaned data is then normalized and key features—such as pathogen concentration and chemical biomarkers—are extracted and mapped into epidemiological embeddings. These biological signatures help identify mutation patterns and spread trends. Finally, the intelligence core analyzes the data in parallel to detect anomalies and predict potential disease outbreaks.

5.1.3 Strategic Visualization and Decision Support The final stage of the workflow bridges the gap between raw science and strategic action. Unified Dashboards: Processed outputs are synthesized into a GIS-integrated webinterface. This dashboard acts as a "command center" for health administrators, offering a granular view of community-level health. Proactive Alerting: Rather than just showing graphs, the system provides hotspot identification and real-time risk scoring. This allows public health officials to bypass the delays of clinical reporting, enabling a proactive defense that can begin weeks before the first patient enters an emergency room.

5.1.4 System Logging and Features The framework maintains a detailed log file that records sampling data, analysis results, and predictive alerts with timestamps for future auditing, reporting, and epidemiological research.

5.1.5 Data Security and System Architecture The surveillance framework follows a security-first approach using a multi-layer defense architecture that separates raw data ingestion from reporting systems to prevent misuse. Sensitive telemetry data is protected through strong encryption and architectural decoupling, reducing the risk of breaches and ensuring data integrity with tamper-proof audit logs. The system is highly scalable and resilient, supporting both small-scale and nationwide monitoring through container-based modular scaling. It also adapts to environmental changes such as rainfall dilution and demographic variations. Beyond public health, the framework's flexible design allows it to be applied to urban planning and research, ensuring long-term value for smart city ecosystems.

6. RESULTS

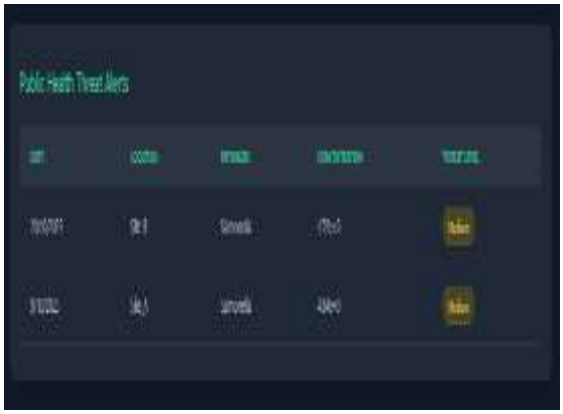
The first figure demonstrate the "Wastewater Analytics Dashboard" showing various metrics and visualizations related to wastewater analysis. The dashboard provides insights into pathogen concentration over time and top hotspot locations, with key statistics such as total samples, unique locations, pathogens detected, and average concentration.

The second image demonstrates ,The graph shows the concentration of three pathogens - E. coli, Norovirus, and Salmonella - over time, from October 5, 2025, to October 30, 2025. The pathogen concentrations fluctuate between 0 and 4 units.

The third image presents a bar graph showing the top 5 hotspot locations based on their average concentration. The graph indicates that Site_C has the highest average concentration, followed by Site_A and Site_B. However, the exact values and the remaining two locations are not visible in the provided image.

The fourth image demonstrates "Public Health Threat Alerts" with information on pathogen outbreaks.The table has five columns: Date, Location, Pathogen, Concentration, and Threat Level.Two rows of data are visible, both indicating Salmonella outbreaks at different sites with medium threat levels.The laptop is open, showing the table on its screen, and the keyboard is visible

below.



7. CONCLUSION

The Emerging Diseases from Wastewater Analytics highlights the powerful role of data-driven surveillance in supporting public health and early outbreak detection Project Synthesis.The integration of Wastewater-Based Epidemiology (WBE) with a high-throughput Big Data architecture represents a paradigm shift in proactive public health. By treating municipal sewage as a collective biological sensor, this framework effectively bypasses the "detection lag" inherent in traditional clinical diagnostics. Our system provides a high-resolution, non invasive window into community health, capturing asymptomatic viral shedding and pathogen trends long before they manifest as a surge in hospital admissions. Technical Contributions project successfully addresses the "Data Noise" challenge of environmental surveillance through a robust, three-tiered methodology Automated Data Pipelines By utilizing Apache Spark and Kafka, the system maintains the velocity needed for real-time monitoring across thousands of sampling points.
Intelligent Forecasting The application of deep-learning anomaly detection ensures that viral spikes are distinguished from routine seasonal fluctuations, providing health officials with a reliable predictive lead time of up to 14 days. Scalable ResilienceWhether deployed across university campuses or integrated into a nationwide Smart City grid, the modular design ensures the framework can scale horizontally to meet the demands of a growing biological threat landscape. Final Remarks Ultimately, this project delivers a resilient and privacy conscious shield for modern society. By eliminating the need for invasive individual testing and replacing it with aggregated biogenetic telemetry, we have developed a cost-effective alternative for pandemic preparedness. This framework does not just track diseases; it empowers governments with the foresight needed to initiate localized, precision-based interventions, potentially saving lives and preserving economic stability through the power of data-driven intelligence.

REFERENCES

1. Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The Hadoop Distributed File System. In Mass Storage Systems and Technologies (MSST), (pp. 1-10). (Focuses on HDFS and Big Data Storage).
2. Kreps, J., Narkhede, S., & Rao, J. (2011). Kafka: A Distributed Messaging System for Log Processing. SIGMOD Workshop on Algorithms and Systems for Database Management (DASFAA). (Focuses on Apache Kafka for Real-Time Data Ingestion).
3. Polo, D., et al. (2020). Wastewater-Based Epidemiology for Community-Wide Monitoring of 2 Circulation: A Multicountry Survey. Water Research, 186, 116349. (Provides domain-specific context and methodology).
4. Hata, A., et al. (2021). Bioinformatics Framework for the Analysis of SARS-CoV-2 in Wastewater Samples. Frontiers in Microbiology, 12, 638927. (Focuses on Genomic Data Preprocessing and Bioinformatics).
5. G., et al. (2021). Wastewater Monitoring and Data Analysis Strategies to 2 Tracking and Variant Detection. Viruses, 13(8), 1621. (Covers Data Analysis Strategies and Variant Identification in).
6. Shmueli, G., et al. (2016). Practical Time Series Forecasting with: A Hands-On Guide (Relevant for Time-Series Analysis and Prediction of disease trends).
7. Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly Detection: A Survey. Computing Surveys (), 41(3), 1-58. (Relevant for the Anomaly Detection component in the model).
8. Lagerqvist, C. (2020). Big Data and Artificial Intelligence in Public Health Informatics. Scandinavian Journal of Public Health, 48(2), 119-122. (Discusses the broader use of Big Data and in Public Health).