MSI Journal of
AI and Technology

Volume: 2, Issue: 3 (July-September 2026)
Published by *MSI Publishers*
ISSN 3107-6181 (Online)

Received: 15-07-2026
Accepted: 30-08-2026
Published: 03-09-2026

A Five-Layer IoT-Big Data Architecture for Smart Wastewater Monitoring and Treatment: Design, Integration of Machine Learning, and Deployment Feasibility for Vietnam

Authors Details

Đỗ Gia Bảo[1*]
Faculty of Information Technology, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam.

Nguyễn Tài Tiệp[2]
Faculty of Information Technology, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam.

Corresponding Author: *Đỗ Gia Bảo*

ABSTRACT: Industrial and urban wastewater pollution constitutes a critical environmental challenge in Vietnam, where only approximately 12.5% of municipal wastewater currently undergoes adequate treatment before discharge. This paper proposes the Smart Wastewater Monitoring and Treatment (SWMT) framework, a five-layer IoT–Big Data architecture designed to enable real-time, data-driven management of wastewater quality and treatment processes. The framework integrates: (i) a multi-parameter wireless sensor network (WSN) measuring pH, dissolved oxygen (DO), biochemical oxygen demand (BOD), chemical oxygen demand (COD), total suspended solids (TSS), ammonium (NH_4^+), temperature, and flow rate; (ii) LoRaWAN and MQTT-based communication protocols; (iii) edge computing for on-site preprocessing and rapid rule-based alerting; (iv) a cloud-hosted Lambda Architecture employing Apache Kafka and Spark Streaming for high-throughput real-time analytics; and (v) machine learning models - Long Short-Term Memory (LSTM) networks, Isolation Forest, and Random Forest - for

 DOI: ***10.5281/zenodo.22270023***

predictive water quality forecasting, unsupervised anomaly detection, and pollution severity classification. Simulation results demonstrate an anomaly detection latency of 3.2 minutes (versus ~8 hours with conventional manual sampling, representing a 99.3% reduction), a true positive rate of 94.2% (FPR = 2.8%), an LSTM forecast RMSE of 0.12 mg/L for COD, and an estimated 18.5% improvement in treatment efficiency through ML - optimised chemical dosing. The paper further analyses regulatory compliance with QCVN 40:2011/BTNMT and QCVN 14:2008/BTNMT, and presents deployment feasibility assessments for three target contexts: industrial zones (IZs), urban wastewater networks, and rural craft villages. The SWMT framework contributes a comprehensive, scalable, and cost-effective blueprint for intelligent wastewater governance in developing-country contexts.

Keywords: *Internet of Things (IoT); Big Data; wastewater monitoring; LSTM; anomaly detection; LoRaWAN; smart water management; Vietnam; QCVN; machine learning.*

1. Introduction

Rapid industrialisation and urbanisation in Southeast Asia have dramatically intensified pressures on freshwater ecosystems. Vietnam presents a particularly acute case: more than 10 million m^3 of wastewater are discharged daily, of which only about 12.5% receives adequate treatment before entering receiving water bodies (Center for Environment and Community Research [CECR], 2023). The 295 active industrial zones (IZs) collectively generate roughly 1.0 million m^3/day of industrial effluent, while the approximately 1,300 identified craft villages contribute a further 50,000 m^3/day of largely untreated discharges (CECR, 2023; Vietnam Government, 2022). Despite legislative advances - including QCVN 40:2011/BTNMT (industrial wastewater) (Ministry of Natural Resources and Environment [MONRE], 2011) and QCVN 14:2008/BTNMT (domestic wastewater) (MONRE, 2008) - enforcement remains constrained by infrequent manual grab sampling, limited online monitoring infrastructure, and high operational costs for regulatory agencies.

Conventional wastewater monitoring relies on periodic manual sampling followed by off-site laboratory analysis, a workflow with inherent latencies of 24–72 hours (El-Shafeiy et al., 2023). Such delays preclude timely intervention during sudden discharge spikes, toxic chemical flushes, or equipment failures. The deployment of Internet of Things (IoT) sensor networks can bridge this gap by enabling continuous, multi-parameter acquisition at sub-minute resolution (Ahmad et al., 2025; Essamlali et al., 2024). When combined with Big Data streaming pipelines and machine learning (ML) analytics, these networks can transition wastewater management from reactive policing to predictive control, supporting both operational optimisation and regulatory compliance (Alzahrani et al., 2023; Dada et al., 2024; Lowe et al., 2022).

Despite a growing body of literature on smart water systems (Dada et al., 2024; Essamlali et al., 2024; Lowe et al., 2022; Taloma et al., 2025), the majority of proposed architectures have been designed for well-resourced contexts featuring reliable grid power, ubiquitous high-speed cellular coverage, and substantial municipal budgets. Vietnam's heterogeneous infrastructure - characterised by variable network coverage, diverse industrial effluent profiles, and budget - constrained local administrative bodies-demands a purpose-built, adaptable architecture. Furthermore, no prior study has systematically mapped an end-to-end IoT–Big Data wastewater framework to Vietnam's QCVN regulatory standards with deployment-specific techno-economic feasibility analyses.

This paper addresses this gap by proposing the SWMT (Smart Wastewater Monitoring and Treatment) framework. The specific contributions are: (1) a five-layer IoT architecture (Perception, Network, Edge, Cloud, Application) tailored to infrastructure constraints in developing economies; (2) a QCVN-aligned eight-parameter sensor node specification with resilient local buffering and automated in-situ recalibration; (3) a Lambda Architecture Big Data pipeline benchmarked at a sustained processing throughput of 850,000 data points per minute; (4) an integrated ML ensemble (LSTM + Isolation Forest + Random Forest) evaluated on a rigorously synthesized dataset reflecting real-world industrial effluent dynamics; and (5) techno-economic feasibility analyses and modular deployment roadmaps across industrial zones, municipal networks, and craft villages in Vietnam.

The remainder of the paper is structured as follows. Section 2 reviews related work. Section 3 describes the research methodology. Section 4 presents the SWMT architecture. Section 5 details the ML integration. Section 6 discusses deployment feasibility. Section 7 reports simulation results. Section 8 presents discussion, and Section 9 concludes with future directions.

DOI: ***10.5281/zenodo.22270023***

2. Related Work

2.1 IoT-Based Water Quality Monitoring

The integration of IoT sensor networks and cloud platforms for environmental monitoring has expanded rapidly. Essamlali et al. (2024) demonstrated that combining IoT sensor arrays with machine learning enables proactive public health protection and early anomaly isolation, replacing manual intervention with low-power wireless telemetry. For real-time anomaly detection in water sensor streams, El-Shafeiy et al. (2023) developed a multivariate deep learning model achieving high detection accuracy across multi-parameter field sensors, demonstrating that deep architectures substantially outperform single-threshold triggers when co-dependent water parameters are monitored simultaneously. Liu et al. (2025) proposed a hybrid VAE-LSTM model for wastewater treatment plants, integrating reconstruction error with temporal sequence forecasting to identify anomalous influent shocks with an AUC of 0.97.

In terms of communication protocols for resource-constrained environments, Jabbar et al. (2024) validated a LoRaWAN IoT architecture for rural water quality monitoring, proving transmission ranges exceeding 10 km under battery-powered conditions. Ramos Villalta et al. (2024) implemented an ESP32 and Raspberry Pi-based LoRaWAN setup for multi-parameter water telemetry, confirming the protocol's stability and affordability for continuous environmental tracking.

2.2 Big Data and Machine Learning in Wastewater Management

Big Data streaming platforms have become foundational for environmental cyber-physical systems. Dada et al. (2024) showed that AI-driven smart water infrastructure significantly optimizes chemical coagulant dosing while mitigating system downtime. Lowe et al. (2022) reviewed predictive analytics across wastewater treatment life cycles, establishing that LSTM recurrent networks consistently outperform classical autoregressive models for non-linear time-series tracking of chemical oxygen demand (COD) and biochemical oxygen demand (BOD).

In distribution and collection network monitoring, Taloma et al. (2025) surveyed machine learning methods for anomaly and burst detection, highlighting deep learning as state-of-the-art for high-dimensional sensory streams. For unsupervised anomaly identification in

DOI: 10.5281/zenodo.22270023

continuous metering streams, Kanyama et al. (2024) identified Isolation Forest as particularly well-suited for IoT environmental feeds due to its sub-sampling efficiency and robust isolation of multidimensional outliers without requiring pre-labeled training datasets. Ahmad et al. (2025) further emphasized that combining low-power edge gateways with scalable cloud telemetry is imperative for modern water resource governance.

2.3 Gaps in the Literature and Positioning of This Work

Despite these developments, three critical gaps remain: (i) existing architectures rarely account for the intermittent connectivity, heavy sensor fouling, and severe capital budget constraints typical of industrial parks and craft villages in developing regions; (ii) no existing IoT–ML pipeline is explicitly calibrated against Vietnamese National Technical Regulations (QCVN 40:2011/BTNMT and QCVN 14:2008/BTNMT); and (iii) published frameworks lack detailed deployment-level cost-benefit models for local governance adoption. The SWMT architecture resolves these gaps directly.

3. Research Methodology

3.1 Architecture-Driven Design Approach

The SWMT architecture follows the ITU-T Y.2060 IoT reference model, augmented with an intermediate Edge Computing tier aligned with the OpenFog Reference Architecture (IEEE 1934). System requirements were synthesized from: (i) a systematic literature review following PRISMA guidelines; and (ii) field assessments across five industrial zones in Ho Chi Minh City, Binh Duong, and Dong Nai provinces, evaluating existing SCADA availability, grid stability, telecommunications signal quality, and local operator skill levels.

3.2 Data Model and Pipeline Design

The data schema adheres to the OGC SensorThings API (OGC 15-078r6) specification, ensuring native interoperability with Vietnam's National Environmental Monitoring System (EMON). Data flows are structured into a Lambda Architecture:

- Speed Layer: Kafka ingestion → Spark Streaming → Redis cache (target end-to-end processing latency < 2 seconds);

DOI: *10.5281/zenodo.22270023*

- Batch Layer: Apache HDFS cold storage → Spark MLlib batch re-training → InfluxDB/HBase analytical views (scheduled on hourly and daily batches).

3.3 Machine Learning Evaluation and Synthetic Data Protocol

To evaluate model performance prior to physical commissioning, a comprehensive synthetic telemetry dataset was generated. The dataset simulates 30 continuous operational days across 50 virtual sensor nodes (43,200 time-steps per parameter per node at 1-minute sampling intervals).

The data generation protocol was rigorously modeled after historical telemetry profiles obtained from municipal and industrial wastewater facilities in Binh Duong Province:

1. Diurnal Baseline Modeling: Base temporal variations of COD, BOD, TSS, DO, and pH were synthesized using coupled sinusoidal and Fourier series capturing 24-hour industrial production and domestic wash cycles.
2. Stochastic Environmental Perturbations: Gaussian noise ($\mu = 0$, $\sigma = 0.05 \cdot$ baseline) and diurnal thermal fluctuations were injected to reflect sensor noise and ambient weather impacts.
3. Anomaly Injection: Five realistic anomaly types (representing 5% of total samples) were programmatically introduced: (a) acute industrial chemical spikes; (b) gradual sensor fouling and calibration drift; (c) sensor freezing/stuck-value faults; (d) intermittent data dropouts; and (e) illegal off-hour discharge surges breaching QCVN Column A/B standards.

4. The SWMT Framework: A Five-Layer IoT–Big Data Architecture

4.1 Architectural Overview

The SWMT conceptual architecture bridges physical sensors with automated actuation and cloud analytics. Data flows from physical sensor nodes (Layer 1) through wireless networks (Layer 2) to edge gateways (Layer 3), then to a cloud Big Data platform (Layer 4), where processed intelligence is surfaced via dashboards and APIs (Layer 5). A feedback control path enables automated actuation of treatment equipment (pumps, chemical dosing

 DOI: 10.5281/zenodo.22270023

systems, valves) based on ML-generated commands, closing the monitoring–control loop. Table 1 summarises each layer's components, technologies, and primary functions.

Table 1. Five-layer architecture of the SWMT framework.

Layer	Key Components	Technologies	Primary Function
Perception	Multi-parameter sensor arrays (pH, DO, BOD, COD, TSS, NH_4^+, Temp, Flow)	Electrochemical & optical probes; RS-485/Modbus; ZigBee	Continuous acquisition of effluent physico-chemical metrics
Network	IoT gateways; LoRaWAN concentrators; 4G/NB-IoT modems	LoRaWAN (Class A/C), MQTT (QoS-2), CoAP, TLS 1.3	Secure, low-power telemetry transmission to cloud/edge
Edge Computing	Industrial Edge Gateways (Raspberry Pi 4 / Jetson Nano)	TensorFlow Lite, ONNX Runtime, Node-RED, SQLite	Local filtering, QCVN rule evaluation, <5s local alerting
Cloud / Big Data	Ingestion & Analytics Clusters (Kafka, Spark, HDFS, InfluxDB)	AWS / Azure / Private Cloud; Lambda Architecture	High-throughput stream processing; batch training & storage
Application	Operator Dashboards; Regulator Portals; REST APIs	Grafana, React.js, Python Flask, Webhooks	Visualization, QCVN compliance reporting, alert dispatch

4.2 Layer 1 – Perception: Multi-Parameter Sensor Nodes

Each field sensing node is driven by an ultra-low-power ARM Cortex-M4 microcontroller (< 10 mA in active acquisition state) paired with a 24-bit precision ADC. To guarantee zero data loss during network disruptions, each node incorporates a 512 MB external SPI Flash / MicroSD storage module operating in a FIFO store-and-forward mode. Power is supplied by a 20 W monocrystalline solar panel backed by a 10 Ah $LiFePO_4$ battery pack.

Eight physico-chemical parameters are measured simultaneously (Table 2). Thresholds are aligned with QCVN 40:2011/BTNMT (industrial wastewater, Column A) and QCVN 14:2008/BTNMT (domestic wastewater). Internal RS-485 bus (Modbus RTU) connects wired probes; ZigBee (IEEE 802.15.4) provides short-range wireless connectivity for

 DOI: 10.5281/zenodo.22270023

mobile or submerged nodes. Sensor calibration is performed automatically every 24 hours using on-board standard solutions, with drift alerts issued if calibration offset exceeds 5%.

Table 2. Sensor node specification and QCVN regulatory thresholds.

Parameter	Measurement Range	Target Accuracy (Nominal)	Sensing Technology	QCVN Regulatory Threshold
pH	0 – 14	± 0.02 pH	Combined glass electrode	6.0 – 9.0 (QCVN 40/14)
DO	0 – 20 mg/L	± 0.1 mg/L	Optical luminescence	≥ 2.0 mg/L (Receiving water)
BOD_5	0 – 500 mg/L	± 5% FS	Bio-sensor / Respirometric	< 30 mg/L (Col. A)
COD	0 – 2,000 mg/L	± 3% FS	UV/Vis Spectrophotometry	< 75 mg/L (Col. A)
TSS	0 – 2,000 mg/L	± 2% FS	Optical Nephelometry (NTU)	< 100 mg/L (Col. A)
Temperature	0 – 80 °C	± 0.1 °C	Pt100 RTD	< 40 °C
Flow Rate	0 – 10 m^3/s	± 0.5% FS	Ultrasonic / Electromagnetic	Continuous monitoring
Ammonium (NH_4^+)	0 – 100 mg/L	± 2% FS	Ion-Selective Electrode (ISE)	< 10 mg/L (Col. A)

Note: Accuracies represent nominal target specifications maintained via automated 24-hour buffer-solution rinsing and recalibration cycles.

4.3 Layer 2 – Network: Hybrid Low-Power Telemetry

Network technology is selected based on deployment context. LoRaWAN (Class A/C, SF7–SF12, AS923 MHz band) is the primary protocol for dispersed IZ deployments, supporting communication ranges of 2–15 km at 0.3–50 kbps and a node lifetime of up to 10 years on a single battery pack (Maurya et al., 2024). Each LoRaWAN concentrator handles up to 10,000 end devices per gateway, making it economically viable for city-scale deployments. In urban zones with dense infrastructure, 4G LTE CAT-M1 or NB-IoT modules provide higher bandwidth for video-augmented monitoring. Wi-Fi 6 or Ethernet is used within centralised WWTP control buildings.

DOI: 10.5281/zenodo.22270023

Application-layer messaging uses MQTT with QoS Level 2 (exactly-once delivery) for critical parameter streams and CoAP for resource-constrained nodes. All communications are secured with TLS 1.3 and X.509 certificate-based device authentication, preventing data injection and man-in-the-middle attacks.

4.4 Layer 3 – Edge Computing: On-Site Intelligence

Edge gateways (minimum specification: ARM Cortex-A72, 4 GB RAM, 128 GB NVMe, optional 1 TOPS NPU) serve clusters of 10–20 sensor nodes, performing: (i) Kalman filtering and missing-value interpolation; (ii) QCVN threshold-based rule evaluation with sub-5-second alert generation; (iii) delta-encoding compression (average 7:1 compression ratio) before cloud upload; and (iv) actuator control for local chemical dosing and bypass valves. TensorFlow Lite deployments of compact LSTM models (< 2 MB) support sub-100 ms on-site inference, enabling treatment responses independent of cloud connectivity.

4.5 Layer 4 – Cloud / Big Data: Lambda Architecture

The cloud layer implements the Lambda Architecture with dual processing paths:

- Speed Layer: Designed with an architectural capacity ceiling exceeding 100,000 raw telemetry events/second via distributed Kafka brokers. Apache Spark Streaming executes micro-batch processing (500 ms windows), detecting QCVN exceedances and running inference on the deployed LSTM model. Results are written to Redis Cache (< 1 s read latency) for dashboard consumption.

- Batch Layer: Raw data is persisted to Apache HDFS (cold storage, replication factor 3). Nightly Spark MLlib jobs retrain ML models on extended historical windows, updating model artefacts deployed to both cloud and edge nodes. Query-optimised results are stored in Apache HBase (operational data) and InfluxDB (time-series visualisation).

The Serving Layer aggregates outputs from both paths into a unified REST API, providing sub-second query responses for the Application Layer regardless of data volume.

DOI: ***10.5281/zenodo.22270023***

4.6 Layer 5 – Application: Decision Support and Governance

The Application Layer delivers intelligence to three user classes: (i) Plant operators via a Grafana-based real-time dashboard displaying GIS sensor maps, trend charts, and 4-hour LSTM forecasts; (ii) Environmental regulators via automated QCVN compliance reports pushed to the EMON national portal; and (iii) Data scientists via a Python/Jupyter interface for ad-hoc ML experimentation. A three-tier automated alert cascade-Yellow (QCVN Column B breach), Orange (Column A breach), Red (2× Column A)-triggers SMS, email, and mobile notifications, with Red-tier alerts simultaneously issuing automated valve-close commands.

5. Machine Learning Integration

5.1 LSTM-Based Water Quality Forecasting

The LSTM forecasting module predicts four-hour-ahead values of COD, pH, and DO from a 24-hour historical input window (96 time steps at 15-minute resolution). The three-layer stacked LSTM architecture (128–64–32 units, Dropout=0.2, Dense output) was selected over ARIMA and GRU based on superior RMSE performance on non-linear, non-stationary effluent data from wastewater treatment plants (Dada et al., 2024; Lowe et al., 2022). Training uses the Adam optimiser (learning rate = 0.001, batch size = 64, 100 epochs) with early stopping (patience = 10). Forecasts are used to pre-adjust chemical dosing 30–60 minutes ahead of predicted exceedances, yielding an 18.5% treatment efficiency gain.

5.2 Isolation Forest for Unsupervised Anomaly Detection

Isolation Forest (contamination rate = 0.05, 100 trees, sub-sampling size $\psi = 256$) detects multivariate anomalies across all eight parameters without requiring labelled training data - a critical advantage given the scarcity of annotated exceedance events in Vietnamese IZ monitoring records (Kanyama et al., 2024; Liu et al., 2008). The model is initialised on three months of normal-operation data and updated weekly via online learning. Its algorithmic training complexity of $O(t \cdot \psi \log \psi)$ and evaluation complexity of $O(n \log \psi)$ ensure seamless scalability across the full network without computational bottlenecks (Liu et al., 2008). In simulation, it achieved TPR = 94.2% and FPR = 2.8%, comparing favourably with benchmark environmental IoT deployments (El-Shafeiy et al., 2023).

 DOI: 10.5281/zenodo.22270023

5.3 Random Forest for Pollution Severity Classification

Random Forest (100 trees, max_depth = 15, class_weight = 'balanced') classifies concurrent readings into four severity levels: Safe, Advisory, Dangerous, and Critical, mapped to the QCVN threshold bands. Feature importance scores provide regulators with interpretable evidence identifying which parameters most contributed to each pollution event - supporting source attribution and enforcement actions (Lowe et al., 2022). The model achieves overall accuracy of 96.7% (F1-macro = 0.965) on the simulation dataset. Table 3 summarises the integrated algorithms.

Table 3. Machine learning algorithms integrated in the SWMT framework.

Algorithm	Dedicated Task	Performance (Simulated)	Technical Justification
Stacked LSTM (128-64-32)	4- hour forward forecasting (COD, pH, DO)	RMSE(COD) = 0.12 mg/L; RMSE (pH) = 0.08; MAE(DO) = 0.09 mg/L	Robust modeling of non-linear diurnal wastewater dynamics (Dada et al., 2024)
Isolation Forest (t = 100)	Unsupervised multivariate outlier detection	TPR = 94.2%; FPR = 2.8%; Latency = 3.2 min	Label-free training; efficient O(n log n) scalability (Liu et al., 2008)
Random Forest (100 trees)	4-tier QCVN severity classification	Overall Accuracy = 96.7%; Macro-F1 = 0.965	High interpretability; resilience against feature collinearity (Lowe et al., 2022)
VAE-LSTM (Hybrid, optional)	Sensor fault & cyber-intrusion detection	Benchmark AUC = 0.97	Joint reconstruction and temporal error profiling (Liu et al., 2025)

6. Deployment Feasibility for Vietnam

6.1 Industrial Zone Scenario

For a medium-scale IZ (200–500 ha, 50–150 enterprises), the recommended SWMT deployment includes 30–60 sensor nodes at enterprise discharge points and 5 nodes at the centralised WWTP. Hardware costs are estimated at VND 2–4 billion (~USD 80,000–160,000), with annual operating costs of VND 400–600 million. Under Decree

 DOI: 10.5281/zenodo.22270023

08/2022/ND-CP (Vietnam Government, 2022), which mandates automatic continuous monitoring for IZs with discharge volume > 500 m^3/day, SWMT provides a regulatory-compliant solution at an estimated 40–50% lower lifecycle cost than equivalent SCADA-based systems, through the use of open-source software and LoRaWAN's shared infrastructure model.

6.2 Urban Wastewater Network Scenario

In Ho Chi Minh City, where the centralised wastewater treatment rate is targeted to reach 71.3% by 2025 (CECR, 2023), SWMT can be integrated into existing pumping stations and treatment plants as a low-cost monitoring overlay. Priority deployment corridors include the Tan Hoa – Lo Gom and Te Canal catchments, where combined sewer overflows are most frequent. Cloud analytics can correlate sensor data with rainfall events (via meteorological API) to predict overflow risk and trigger pre-emptive operational adjustments.

6.3 Craft Village Scenario

Only 16% of Vietnam's approximately 1,300 craft villages possess standard-compliant treatment systems (CECR, 2023). The SWMT Lite variant (3–5 sensors/node, LoRaWAN-only, simplified edge gateway at < VND 3 million/node) reduces total deployment cost to VND 150–400 million per village, aligning with district-level environmental management budgets. Shared LoRaWAN gateway infrastructure across multiple villages further reduces per-village costs by 30–40%.

6.4 Key Challenges and Mitigation Strategies

Four principal implementation challenges are identified:

- Heterogeneous network coverage: Mitigated by LoRaWAN for rural/semi-urban areas and NB-IoT for dense urban zones; satellite IoT (e.g., Swarm/Globalstar) as fallback for remote sites.

- Sensor fouling and calibration drift: Addressed by automated 24-hour in-situ calibration, alert-triggered manual maintenance scheduling, and modular probe replacement design (< 15 min field swap).

 DOI: *10.5281/zenodo.22270023*

- Operator capacity: Addressed through progressive training programmes, AI-driven contextual guidance within the dashboard, and Vietnamese-language documentation.

- Data security and regulatory integrity: End-to-end TLS 1.3 encryption, hardware security modules (HSM) at edge gateways, and immutable append-only logging for regulatory reports.

7. Simulation Results

Table 4 summarises SWMT performance from 30-day simulations across a 50-node virtual network, compared to a conventional periodic-sampling baseline. All simulations were executed on a 3-node Apache Spark cluster (Ubuntu 22.04; 3 × 32-core AMD EPYC, 128 GB RAM per node; Apache Spark 3.5; TensorFlow 2.15).

Table 4. Performance benchmark of SWMT vs. conventional monitoring.

KPI / Performance Metric	Proposed SWMT System	Conventional Baseline	Improvement / Note
Mean Anomaly Detection Latency	3.2 minutes	~480 minutes (8h grab sampling)	99.33% reduction in response time
True Positive Detection Rate (TPR)	94.2%	N/A (Reactive only)	Evaluated on 50-node synthetic dataset
False Positive Rate (FPR)	2.8%	N/A	Filtered via multi-tier alert engine
COD Forecasting RMSE	0.12 mg/L	N/A	4-hour forward window
Treatment Efficiency Gain	+18.5%	Baseline	Via ML-optimized chemical dosing
End-to-End Data Latency	1.8 seconds (avg)	24 – 72 hours (Laboratory analysis)	Sensor probe → Cloud Dashboard
Sustained Stream Throughput	850,000 points/min	N/A	Benchmark load under 50-node setup

DOI: *10.5281/zenodo.22270023*

The 99.3% reduction in anomaly detection latency (3.2 min vs. ~480 min) is the most operationally significant outcome, as it enables treatment interventions before non-compliant effluent reaches receiving water bodies. The 18.5% treatment efficiency gain arises primarily from LSTM-optimised dosing of coagulants and disinfectants, which reduces both chemical overconsumption and under-treatment events. The sustained throughput of 850,000 data points/minute (~14,167 events/second) on a 3-node cluster confirms the pipeline can scale to city-wide deployments of 5,000+ nodes without architectural changes.

The LSTM model's RMSE of 0.12 mg/L for COD compares favourably with the RMSE of 0.28 mg/L reported for standard LSTM in a comparable coastal WWTP study (Dada et al., 2024), suggesting that the 24-hour sliding-window input design effectively captures diurnal discharge cycles characteristic of Vietnamese IZ operations. Isolation Forest's FPR of 2.8% is consistent with the 2–4% false alarm rates documented in smart metering anomaly detection reviews (Kanyama et al., 2024) and is operationally acceptable given the three-tier alert structure which filters rule-based confirmations before operator notifications.

8. Discussion

8.1 Comparative Analysis with Prior Work

Compared to Alzahrani et al.'s (2023) IoT–blockchain smart city wastewater system, the SWMT framework achieves comparable detection rates while reducing estimated deployment cost by approximately 35% through the substitution of cellular connectivity with LoRaWAN and the use of open-source Big Data components. Unlike the system described by Jabbar et al. (2024), which focused on single-site industrial monitoring, SWMT's Lambda Architecture supports simultaneous multi-site federation - essential for IZ-level regulatory oversight. The VAE-LSTM hybrid of Liu et al. (2025) achieves a higher AUC (0.97) for sensor fault detection; integrating this model as a fourth ML component represents a natural extension of the proposed framework.

8.2 Limitations

The current evaluation is based entirely on simulation data. While the synthetic dataset was calibrated to historical IZ records, real-world factors - including biofouling of DO sensors,

DOI: *10.5281/zenodo.22270023*

temperature-induced pH drift, and RF interference in metal-clad factories - may degrade performance. A pilot deployment at an IZ in Binh Duong Province is planned for late 2026, which will provide ground-truth validation. Additionally, the Lambda Architecture introduces operational complexity; organisations with limited DevOps capacity may benefit from a simplified Kappa Architecture (stream-only) variant as an interim deployment option.

8.3 Broader Implications

The SWMT framework's modular design permits incremental deployment - beginning with 3–5 nodes at the most critical discharge points and expanding as budgets allow - making it practically adoptable by Vietnam's resource-constrained local governments. The QCVN-aligned alert thresholds and automated regulatory reporting directly reduce the administrative burden on both operators and regulators, potentially accelerating enforcement. As Vietnam progresses toward its National Green Growth Strategy 2021–2030 targets, architectures such as SWMT that simultaneously reduce operational costs and strengthen environmental compliance provide a pathway toward data-driven, sustainable industrial water governance.

9. Conclusion

This paper presented the SWMT framework, a five-layer IoT–Big Data architecture for smart wastewater monitoring and treatment tailored to Vietnam's regulatory, infrastructural, and economic context. The framework integrates multi-parameter IoT sensor nodes, LoRaWAN networking, edge computing, a cloud-hosted Lambda Architecture pipeline, and an ensemble of ML algorithms (LSTM, Isolation Forest, Random Forest) within a coherent, scalable design. Simulation results demonstrate a 99.3% reduction in anomaly detection latency (3.2 minutes vs. 8 hours), a 94.2% true positive detection rate, and an estimated 18.5% improvement in treatment efficiency compared to conventional systems.

The principal scientific contributions are: (1) the first comprehensive IoT–Big Data wastewater architecture explicitly aligned to Vietnam's QCVN regulatory framework; (2) a QCVN-calibrated eight-parameter sensor node specification with in-situ auto-calibration; (3) a benchmarked Lambda Architecture pipeline processing 850,000 data points/minute;

 DOI: ***10.5281/zenodo.22270023***

(4) an integrated, simulation-evaluated ML ensemble; and (5) techno-economic feasibility analyses for three deployment contexts.

Acknowledgment

The authors would like to express their sincere gratitude to the Faculty of Information Technology, Nguyen Tat Thanh University, for providing the necessary facilities, computing resources, and academic support to carry out this research.

References

1. Addas, A., Khan, M. N., & Naseer, F. (2024). Waste management 2.0 leveraging Internet of Things for an efficient and eco-friendly smart city solution. PLOS ONE, 19(7), e0307608. https://doi.org/10.1371/journal.pone.0307608

2. Ahmad, S., Noor, S., Ali, F., Iqbal, J., & Ahmad, I. (2025). Use of Internet of Things in water resources applications: Challenges and future directions. Discover Internet of Things, 5, Article 112. https://doi.org/10.1007/s43926-025-00193-7

3. Al-Mansoori, N., & Almujibah, H. (2025). LoRaWAN-based smart water management IoT applications: A review. Journal of Information and Telecommunication, 9(4), 420–446. https://doi.org/10.1080/24751839.2025.2458889

4. Alzahrani, A. I. A., Chauhdary, S. H., & Alshdadi, A. A. (2023). Internet of Things (IoT)-based wastewater management in smart cities. Electronics, 12(12), 2590. https://doi.org/10.3390/electronics12122590

5. Center for Environment and Community Research. (2023). Current status of waste pollution and water pollution in Vietnam. CECR. https://cecr.vn/en/current-status-of-waste-pollution-and-water-pollution-in-vietnam/

6. Dada, M. A., Majemite, M. T., Obaigbena, A., Daraojimba, O. H., Oliha, J. S., & Nwokediegwu, Z. Q. S. (2024). Review of smart water management: IoT and AI in water and wastewater treatment. World Journal of Advanced Research and Reviews, 21(1), 1373–1382. https://doi.org/10.30574/wjarr.2024.21.1.0171

 DOI: *10.5281/zenodo.22270023*

7. El-Shafeiy, M. A., Alsabaan, M., Ibrahem, M. I., & Elwahsh, H. (2023). Real-time anomaly detection for water quality sensor monitoring based on multivariate deep learning technique. Sensors, 23(20), 8613. https://doi.org/10.3390/s23208613

8. Essamlali, I., Nhaila, H., & El Khaili, M. (2024). Advances in machine learning and IoT for water quality monitoring: A comprehensive review. Heliyon, 10(6), e27920. https://doi.org/10.1016/j.heliyon.2024.e27920

9. Jabbar, W. A., Ting, T. M., Hamidun, M. F. I., Kamarudin, A. H. C., Wu, W., Sultan, J., Alsewari, A. A., & Ali, M. A. (2024). Development of LoRaWAN-based IoT system for water quality monitoring in rural areas. Expert Systems with Applications, 242, 122761. https://doi.org/10.1016/j.eswa.2023.122761

10. Kanyama, M. N., Shava, F. B., Gamundani, A. M., & Hartmann, A. (2024). Machine learning applications for anomaly detection in smart water metering networks: A systematic review. Physics and Chemistry of the Earth, Parts A/B/C, 134, 103558. https://doi.org/10.1016/j.pce.2024.103558

11. Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008) (pp. 413–422). IEEE. https://doi.org/10.1109/ICDM.2008.17

12. Lowe, M., Qin, R., & Mao, X. (2022). A review on machine learning, artificial intelligence, and smart technology in water treatment and monitoring. Water, 14(9), 1384. https://doi.org/10.3390/w14091384

13. Maurya, P., Hazra, A., Kumari, P., Sørensen, T. B., & Das, S. K. (2024). A comprehensive survey of data-driven solutions for LoRaWAN: Challenges and future directions. ACM Transactions on Internet of Things, 6(1), 1–42. https://doi.org/10.1145/3711953

14. Ministry of Natural Resources and Environment of Vietnam. (2008). QCVN 14:2008/BTNMT: National technical regulation on domestic wastewater. MONRE.

15. Ministry of Natural Resources and Environment of Vietnam. (2011). QCVN 40:2011/BTNMT: National technical regulation on industrial wastewater. MONRE.

 DOI: ***10.5281/zenodo.22270023***

16. Nguyen, B. Q., Kantoush, S., Binh, D. V., Saber, M., Vo, D. N., & Sumi, T. (2023). Understanding the anthropogenic development impacts on long-term flow regimes in a tropical river basin, Central Vietnam. Hydrological Sciences Journal, 68(3), 341–354. https://doi.org/10.1080/02626667.2022.2159935

17. Piro, G., Grieco, L. A., Boggia, G., & Camarda, P. (2021). A multi-layer LoRaWAN infrastructure for smart waste management. Sensors, 21(8), 2600. https://doi.org/10.3390/s21082600

18. Ramos Villalta, M. K., Huanca Humpiri, J. A., & Cutipa Coila, R. M. (2024). Development of an IoT system based on LoRaWAN to monitor water quality parameters. International Journal of Electrical and Electronics Engineering, 11(10), 223–230. https://doi.org/10.14445/23488379/IJEEE-V11I10P123

19. Saravanan, K., Anusuya, E., Kumar, R., & Son, L. H. (2024). Artificial intelligence technologies revolutionising wastewater treatment: Current trends and future prospective. Water, 16(2), 314. https://doi.org/10.3390/w16020314

20. Taloma, R. J. L., Cuomo, F., Comminiello, D., & Pisani, P. (2025). Machine learning for smart water distribution systems: Exploring applications, challenges and future perspectives. Artificial Intelligence Review, 58, Article 120. https://doi.org/10.1007/s10462-024-11093-7

21. Uddin, M. R., & Chakrovorty, R. S. (2024). IoT-based real-time water quality monitoring system in water treatment plants. Heliyon, 10(23), e40746. https://doi.org/10.1016/j.heliyon.2024.e40746

22. Vietnam Government. (2022). Decree No. 08/2022/ND-CP detailing a number of articles of the Law on Environmental Protection. Government of Vietnam.

23. Liu, X., Gong, Z., & Zhang, X. (2025). Research on Anomaly Detection in Wastewater Treatment Systems Based on a VAE-LSTM Fusion Model. Water, 17(19), 2842. https://doi.org/10.3390/w17192842

 DOI: ***10.5281/zenodo.22270023***