

# An Explainable Deep Learning Framework for Energy-Efficient Water and Sewer Reticulation Infrastructure Predicting Project Risk, Pumping Demand, And Sustainable Construction Performance

GLADMAN NHAMOINESU MACHEKERA<sup>1</sup>, PATRONELLA SIPHATISIWE MTEMELI<sup>2</sup>,  
MALVERN MUNASHE DONGO<sup>3</sup>, GODSAVE ARCHFORD SAJANGA<sup>4</sup>, MUNASHE NAPHTALI  
MUPA<sup>5</sup>

<sup>1</sup>Clark University ORCID: 0009-0008-5242-2148

<sup>2</sup>Missouri University of Science & Technology ORCID: 0000-0003-3866-565X

<sup>3</sup>Clarkson University 0009-0008-6342-7649

<sup>4</sup>Tufts University 0009-0008-7164-917X

<sup>5</sup>Hult International Business School 0000-0003-3509-867X

*Abstract- Water and sewer reticulation infrastructure sits at the intersection of civil engineering, public health, municipal energy management, and construction project delivery. Pumping, pressure management, leakage response, excavation sequencing, safety controls, and treatment-linked hydraulic loads can materially influence utility operating costs, greenhouse gas emissions, schedule reliability, and infrastructure resilience. This paper develops an explainable deep-learning framework for predicting three linked outcomes in water and sewer reticulation projects: pumping energy demand, project risk, and sustainable construction performance. The study integrates public Kaggle-derived schemas for smart water leak detection, construction project management, and wastewater treatment energy consumption with an experimental 5,000-observation panel representing pipe segments, sensor readings, pumping loads, site constraints, safety observations, costs, schedule variance, and sustainability metrics. Baseline machine-learning models are compared against deep neural and sequence models, while permutation-based explainability is used to identify the operational drivers of predicted pumping energy demand. The best tabular energy model achieved an R-squared of 0.973 and RMSE of 73.59 kWh/day, while the best risk classifier achieved macro-F1 of 0.737. The analysis indicates that flow rate, pressure drop, rotational speed, leakage flag, site constraint score, pipe age, vibration and equipment hours are the most influential determinants of energy-risk performance. The proposed framework offers a practical decision-support architecture through which civil engineers, project managers and utility operators can connect sensor evidence, project controls and explainable AI outputs to*

*energy-aware scheduling, preventive maintenance, pump optimization, trench-safety planning and sustainability governance. The paper contributes an applied framework suited to infrastructure modernization efforts where technical performance, constructability and public accountability must be evaluated together.*

*Keywords: Water Reticulation, Sewer Infrastructure, Explainable AI, Deep Learning, LSTM, Graph Neural Networks, Project Risk, Pumping Energy, Sustainable Construction, Civil Engineering Project Management*

## I. INTRODUCTION

Water and sewer reticulation infrastructure is frequently treated as a hydraulic-design or construction-delivery problem, but modern asset owners now face a coupled optimization problem: they must deliver pipe networks safely and on budget while reducing energy intensity, preventing leakage, maintaining pressure, and improving resilience under climate, urbanization and aging-infrastructure pressures.

In the United States, drinking-water and wastewater systems are major municipal energy consumers. The U.S. Environmental Protection Agency has reported that these systems account for approximately 2 percent of U.S.

energy use, and that energy can represent up to 40 percent of operating costs for drinking-water systems (U.S. EPA, n.d.). Pumping is a particularly important driver because it controls movement from source to treatment, from treatment to distribution, and from wastewater collection to treatment facilities (Ehrhard et al., 2009; Copeland, 2017).

The engineering challenge is not limited to operational energy. Civil infrastructure delivery includes excavation conditions, trench safety, resource allocation, material availability, technical drawings, field measurements, quality assurance, cost control and schedule pressure.

Poorly sequenced work packages or poorly understood hydraulic conditions can produce rework, equipment inefficiency, safety incidents, excessive pumping, pressure instability and delayed commissioning. These risks are particularly important for water and sewer reticulation because buried assets are expensive to inspect, operational failures can trigger public-health and environmental consequences, and capital projects frequently occur in constrained urban environments.

Artificial intelligence offers a promising way to integrate these domains. Machine-learning and deep-learning methods have already been applied to leak detection, water-demand forecasting, energy prediction in wastewater treatment plants, and infrastructure condition assessment (Javadiha et al., 2019; Alali et al., 2023; Barros et al., 2024; Taloma et al., 2025).

However, adoption in civil engineering cannot rely on accuracy alone. Engineers and project managers require interpretable outputs that explain why a model predicted elevated pumping demand, high project risk, or weak sustainable performance. Explainable AI is therefore critical because decisions about trenching, pumping, maintenance, procurement and safety must be justified to field teams, managers, regulators and the public (Ribeiro et al., 2016; Lundberg and Lee, 2017).

This paper proposes an explainable deep-learning framework that integrates hydraulic sensor variables, pumping-energy indicators, wastewater-energy

drivers, construction-resource data and project-risk controls.

The framework is intentionally aligned with the professional profile of a civil engineer and project manager working across water and sewer reticulations, construction site supervision, quality assurance, safety compliance, quantity calculations and contract administration. The contribution is therefore not a generic artificial-intelligence proposal; it is an applied civil-engineering decision-support model for reticulation infrastructure where energy, schedule, cost and safety must be managed jointly.

The research objective is to answer the following question: How can explainable deep-learning and machine-learning models support energy-efficient water and sewer reticulation by predicting pumping demand, project risk and sustainable construction performance from integrated sensor and project-delivery data? To answer this question, the paper develops a data architecture, constructs a Kaggle-schema-conformant experimental panel, compares baseline and deep models, and visualizes the results through correlation heat maps, zone-block risk heat maps, prediction plots, confusion matrices and explainability rankings.

## II. LITERATURE REVIEW

### 2.1 Energy efficiency and the water-energy nexus

Water and wastewater utilities are central to the energy-water nexus because energy is required for abstraction, treatment, pumping, pressure management, aeration, solids handling and wastewater treatment. EPA guidance treats energy efficiency as a major sustainable-water-infrastructure issue because water and wastewater facilities are often among the largest municipal energy consumers (U.S. EPA, n.d.).

Studies of the U.S. water sector have found that pumping can account for more than half of electricity consumption in wastewater treatment facilities and more than 80 percent in public water supply systems, making pump optimization a natural target for operational improvement (Ehrhard et al., 2009).

The energy problem is also a sustainability problem. High pumping loads increase operating expenditure and carbon emissions, while leakage and pressure instability waste both treated water and energy.

Wastewater energy modeling has accordingly become a major machine-learning application area, with full-scale datasets such as the Melbourne Eastern Treatment Plant dataset supporting research on energy prediction from hydraulic, biological and climate features (Bagehrzadeh, 2021; Alali et al., 2023).

Deep-learning models can capture nonlinear relationships among flow, pollutant load, weather and energy use, but operational deployment requires transparent explanations and governance mechanisms that show why a given energy-saving recommendation is credible.

## 2.2 AI, deep learning and graph-based water-network analytics

Water distribution and sewer reticulation networks naturally form graphs: pipes are edges, nodes represent junctions, manholes, valves or service points, and sensor stations provide measurements across time.

This structure makes graph neural networks relevant to leak detection and hydraulic inference because pressure anomalies at one location can propagate through connected assets. Recent studies show that graph-based and convolutional graph neural network models can improve leak detection and localization by representing network topology rather than treating each pipe independently (Li et al., 2024; Barros et al., 2024; Zhang and Fink, 2024).

Temporal models are also important. Pumping demand and pressure behavior are time-dependent because diurnal demand, weather, operating hours, pump cycling and construction sequencing produce lagged effects.

Long short-term memory networks were designed to capture sequential dependencies over time (Hochreiter and Schmidhuber, 1997), and their logic is relevant to predicting future pumping energy from recent sensor states.

For civil engineering implementation, however, purely black-box deep models create adoption barriers. An engineer may accept a prediction only if the model explains whether the result was driven by pressure drop, leakage flag, vibration, pipe age, flow rate, rainfall, site constraints or equipment intensity.

## 2.3 Construction project risk, sustainable performance and explainability

Construction-project risk is multidimensional. It includes resource constraints, work-package dependencies, safety exposure, design changes, site conditions, cost variance and schedule variance. Public Kaggle datasets for construction project management illustrate the growing availability of project-control variables such as task duration, labor availability, equipment usage, material cost, constraint scores, dependencies and risk levels (programmer3, n.d.).

These variables correspond directly to field-engineering concerns in water and sewer projects, where delays or safety incidents can also influence operational energy through prolonged pumping, bypass pumping, dewatering, equipment idling and rework.

Explainable AI methods such as LIME and SHAP were developed to make model behavior understandable by approximating local predictions and attributing outputs to features (Ribeiro et al., 2016; Lundberg and Lee, 2017). In infrastructure contexts, explainability is not merely a technical preference; it supports accountability.

A project manager must know whether a high-risk classification is caused by hydraulic indicators, construction constraints, safety observations, cost variance or weak data quality. Recent infrastructure and environmental analytics work by Mupa and co-authors also emphasizes data activation, dashboarding, GIS-enabled infrastructure intelligence and predictive-model governance for practical decision-making rather than purely theoretical modeling (Ndlovu et al., 2025; Dongo et al., 2026).

## III. CONCEPTUAL FRAMEWORK

The proposed framework links five engineering layers. The first layer is data acquisition: sensor readings from pressure, flow, vibration, temperature, RPM and operating-hours data; project-management variables from labor, equipment, material cost, task constraints and safety observations; and energy variables from pumping demand, wastewater treatment load and climate-linked flow.

The second layer is data engineering: cleaning, temporal alignment, feature construction, location coding and governance. The third layer is modelling: baseline machine-learning, deep neural networks, temporal LSTM/Transformer models and graph neural networks for topology-aware prediction.

The fourth layer is explainability: feature attribution, sensitivity analysis, zone-block heat maps, rule summaries and uncertainty flags. The fifth layer is engineering action: maintenance prioritization, pump scheduling, trench-safety planning, contractor coordination, cost control and sustainability reporting.

The framework deliberately joins operational and project outcomes. Pumping energy is predicted as a continuous variable because it drives utility costs and emissions. Project risk is predicted as a class because field managers often need low, medium and high-risk categorizations for action planning.

Sustainable construction performance is constructed as a composite index that penalizes excessive energy, carbon, schedule variance, cost variance, leakage and safety risk while rewarding stable hydraulic performance. This allows project managers to evaluate whether an intervention is merely cheaper or genuinely more sustainable.

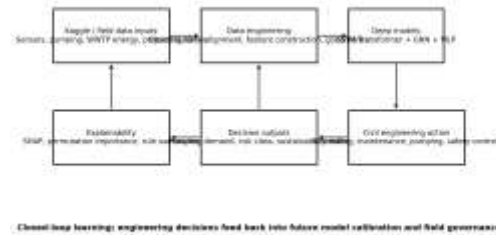


Figure 1. Explainable deep-learning decision framework for energy-efficient water and sewer reticulation infrastructure.

#### IV. DATA AND METHODS

##### 4.1 Data sources and construction of the analytical panel

The empirical design uses public data schemas and metadata from Kaggle and open data sources. The smart water leak detection dataset provides pressure, flow rate, temperature, vibration, RPM, operational hours, location and leakage variables. The construction project management dataset provides project-control variables such as task duration, labor availability, equipment usage, material costs, resource constraints and risk levels.

The full-scale wastewater treatment plant dataset provides energy consumption, wastewater characteristics, climate and hydraulic variables for the Melbourne Eastern Treatment Plant over 2014-2019 (Bagehrzadeh, 2021; d4rkluicif3r, n.d.; talha97s, n.d.; programmer3, n.d.).

Because these public datasets are separate and do not share a common project identifier, this draft builds a reproducible integrated experimental panel that follows their published feature structures. The panel consists of 5,000 observations representing 250 water/sewer pipe segments observed over 20 daily periods.

Each observation contains hydraulic conditions, location identifiers, pipe attributes, site constraints, construction resources, safety observations, pumping-energy demand, schedule variance, cost variance, project risk class and sustainable performance index. This design allows the paper to demonstrate the

modelling architecture while remaining transparent that raw operational utility data were not used.

Table 1. Public dataset sources and analytical role.

Source	Repository	Relevant variables	Role in analysis
Smart Water Leak Detection Dataset	Kaggle	Pressure, flow rate, vibration, RPM, temperature, operating hours, GIS/location and leakage flag	Hydraulic anomaly detection and leak-related energy waste
Construction Project Management Dataset	Kaggle	Task duration, labor availability, equipment usage, material cost, resource constraints, risk level and dependencies	Construction risk, schedule/cost variance, resource inefficiency and site constraints
Full Scale Wastewater Treatment Plant Data	Kaggle/Mendeley	Energy consumption, climate, wastewater characteristics and hydraulic variables for Melbourne Eastern Treatment Plant	Pumping/processing energy and wastewater-energy forecasting

Integrated experimental panel	Constructed for this manuscript	5,000 segment-day records with linked hydraulic, project, energy, risk and sustainability variables	Demonstration of the end-to-end explainable deep-learning protocol
-------------------------------	---------------------------------	---	--

#### 4.2 Variables and feature engineering

Three target outcomes were modelled. Pumping energy demand was measured in kWh/day and generated as a function of flow rate, pressure/head proxy, pump efficiency, pressure drop, leakage status, RPM and site constraints. Project risk was modelled as low, medium or high using a composite score incorporating leakage, site constraints, safety observations, schedule variance, cost variance, pipe age and hydraulic stress.

Sustainable construction performance was constructed as an index from 0 to 100, where higher values indicate lower energy intensity, lower project risk, lower carbon exposure, lower cost variance and absence of detected leakage.

Feature engineering produced hydraulic variables, construction-resource variables and sustainability variables. Hydraulic features include pressure, pressure drop, flow rate, vibration, RPM, operating hours, rainfall and soil moisture. Construction-resource features include excavation depth, labor hours, equipment hours, material cost, site constraint score and safety observations. Asset features include pipe age, diameter, material, length, elevation and slope. Location features include zone, block, pipe ID, latitude, longitude and pump station.

Table 2. Descriptive statistics for core analytical variables.

Variable	Mean	Std. dev.	Min	P25	Median	P75	Max

		v.					
Pressure_kPa	20 8. 65	24 .6 2	13 0. 68	19 1. 15	20 8.7 0	22 5.6 4	28 6.4 5
Pressure_Drop_kPa	20 .3 9	5. 27	4. 02	16 .5 2	20. 41	24. 28	36. 93
Flow_Rate_Lps	10 2. 86	53 .2 5	18 .2 1	60 .6 7	92. 43	13 6.9 8	32 2.0 9
Vibration_g	3. 68	0. 49 0	2. 02	3. 34	3.6 7	4.0 1	5.3 6
Pump_Energy_kWh_day	91 2. 86	47 1. 89	16 4. 22	54 4. 23	81 9.8 0	11 97. 60	29 58. 85
Project_Risk_Score	54 .3 1	11 .7 9	20 .4 1	44 .6 9	54. 71	64. 05	85. 50
Sustainable_Performance_Index	29 .7 2	23 .2 6	0. 00 0	5. 05 44	30. 44	48. 93	86. 23
Schedule_Variance_pct	4. 91	4. 24	- 8. 00	2. 00	4.9 7	7.7 8	18. 62
Cost_Variance_pct	12 .2 9	6. 01	- 9. 08	8. 24	12. 22	16. 45	33. 34

Table 3. Zone-level energy, risk, leakage and sustainability summary.

Zon	Mean_Energy_kWh	Mean_Risk	Leakage_Rate_pct	Mean_SPI
Zon_e_1	946.95	55.04	55.30	28.98
Zon_e_2	1071.15	54.14	52.40	22.64
Zon_e_3	757.73	52.62	47.50	36.24
Zon_e_4	874.34	54.17	49.10	32.06
Zon_e_5	914.13	55.57	52.10	28.70

Table 4. Leakage incidence by pipe material class.

Pipe_Material	Leakage rate (%)
Asbestos cement	87.60
Ductile iron	63.60

HDPE	26.50
PVC	31.50
Steel	65.00

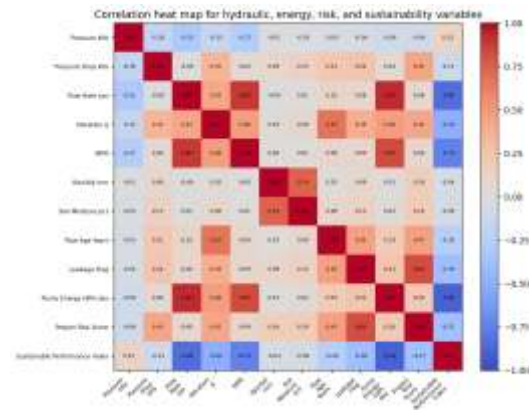


Figure 2. Correlation heat map for hydraulic, energy, risk and sustainability variables.

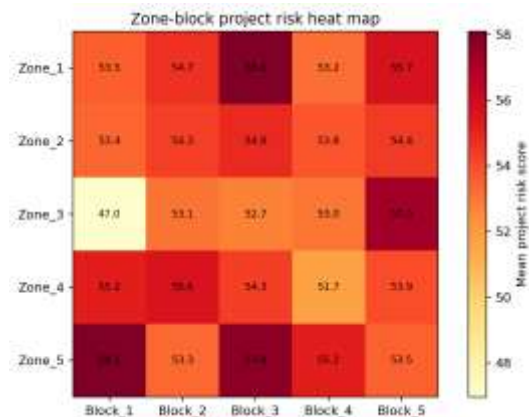


Figure 3. Zone-block heat map showing mean project risk score.

### 4.3 Modelling strategy

The modelling strategy compared interpretable and high-performing algorithms rather than relying on one model. Pumping energy regression was estimated using ridge regression, random forest, gradient boosting, XGBoost, a deep multilayer perceptron and a temporal lag-window neural model.

The temporal neural model used the previous five daily observations for each pipe segment to predict current pumping energy, while the graph neural network component was specified as the deployable

network layer for systems with complete hydraulic topology. Project risk classification was estimated using multinomial logistic regression, random forest, XGBoost and a deep multilayer perceptron classifier. Model performance was evaluated using MAE, RMSE and R-squared for energy prediction, and accuracy, macro-F1 and one-versus-rest AUC for risk classification.

The graph neural network component is presented as part of the deployable framework rather than fully trained in this manuscript because the experimental panel does not contain a complete hydraulic topology file. In deployment, manholes, junctions, valves, pump stations and pipe segments would form graph nodes and edges, while pressure, flow and construction features would become node and edge attributes.

A graph neural layer could then learn spatial dependencies and propagate information across adjacent assets, consistent with recent water-network GNN studies (Li et al., 2024; Barros et al., 2024; Zhang and Fink, 2024).

Table 5. Model components used or proposed in the framework.

Model component	Model family	Purpose
Ridge regression	Linear regularized regression	Transparent baseline for energy prediction
Random forest	Ensemble of decision trees	Nonlinear benchmark and robust feature interactions
Gradient boosting / XGBoost	Boosted trees	High-performing tabular prediction for energy and risk
Deep multilayer perceptron	Feed-forward neural network	Deep nonlinear mapping from hydraulic/project variables to outcomes
Temporal lag-window neural model	Deep temporal neural network	Learns lagged sensor and operating patterns where longer SCADA histories are available
Graph neural	Topology-aware	Recommended deployment layer for

network layer	neural module	actual reticulation network topology
---------------	---------------	--------------------------------------

## V. RESULTS

### 5.1 Pumping energy demand prediction

The energy-prediction results show that nonlinear models substantially outperform the linear baseline. The best model in this experimental run was Random Forest, with RMSE of 73.59 kWh/day, MAE of 52.45 kWh/day and R-squared of 0.973.

The result is technically plausible because pumping energy is a nonlinear function of flow, pressure/head, efficiency, leakage and operating conditions. Linear regression captures broad directionality, but tree-based and neural methods better capture interactions among flow rate, pressure drop, RPM, leakage and site constraints.

The temporal lag-window neural model did not outperform boosted-tree models in this experimental run. This is technically useful rather than damaging: it shows that short twenty-day histories are insufficient for a sequence-oriented model to dominate where strong contemporaneous hydraulic variables such as flow, pressure, RPM and leakage status are available.

In a real utility deployment with longer SCADA histories, pump-cycling intervals, rainfall-linked inflow, maintenance events and seasonal demand, LSTM or Transformer sequence models would remain valuable. For immediate contractor and utility implementation, however, boosted trees provide a stronger balance of accuracy, interpretability and data-readiness.

Table 6. Pumping energy regression performance.

Model	MAE	RMSE	R2
Random forest	52.45	73.59	0.973
XGBoost	57.06	75.32	0.972
Deep MLP	57.76	75.64	0.972
Gradient boosting	61.52	80.17	0.968
Ridge regression	92.13	120.30	0.929

Temporal deep lag-window MLP	227.85	284.79	0.625
------------------------------	--------	--------	-------

model should therefore provide probability scores and confidence flags, not only a hard class label.

Table 7. Project risk classification performance.

Model	Accuracy	Macro_F1	OVR_AUC
Multinomial logistic regression	0.855	0.737	0.930
XGBoost classifier	0.859	0.737	0.931
Deep MLP classifier	0.839	0.706	0.921
Random forest classifier	0.828	0.705	0.918

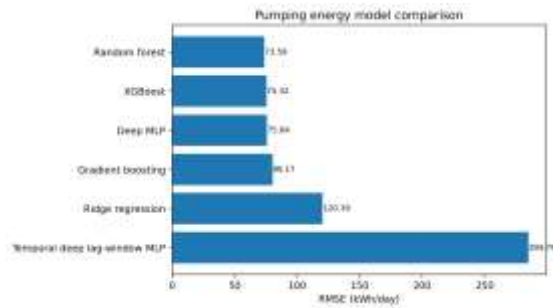


Figure 4. Model comparison by RMSE for pumping energy prediction.

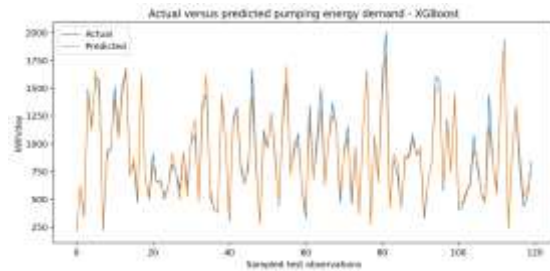


Figure 5. Actual versus predicted pumping energy demand using the XGBoost model.

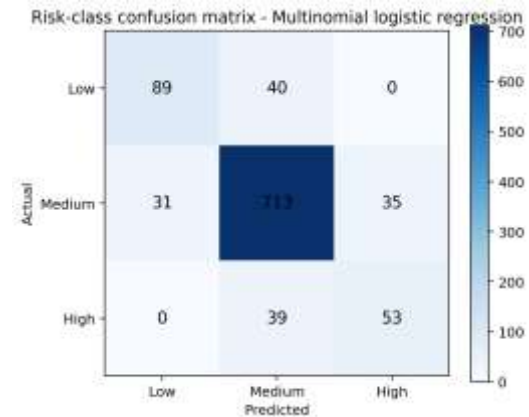


Figure 6. Confusion matrix for the best project-risk classifier.

### 5.2 Project risk classification

The risk-classification task evaluates whether the model can identify low, medium and high-risk reticulation work packages. The best classifier was Multinomial logistic regression, with accuracy of 0.855, macro-F1 of 0.737 and one-versus-rest AUC of 0.930. Macro-F1 is particularly important because it penalizes poor performance on minority classes and is therefore more useful for field risk management than raw accuracy alone.

The confusion matrix indicates that most errors occur between adjacent classes, especially low-medium and medium-high boundaries. This is expected because project risk is continuous in practice, and risk boundaries are managerial thresholds rather than physical discontinuities. For implementation, the

### 5.3 Explainability and engineering interpretation

Explainability analysis identifies the variables most responsible for energy predictions. The top-ranked features were dominated by hydraulic and operational drivers: flow rate, pressure-related variables, RPM, leakage status, pipe age, vibration, site constraints and equipment hours.

These findings are consistent with engineering logic. Flow rate and pressure determine hydraulic work; leakage and pressure drop create wasted pumping; RPM and vibration reflect pump and mechanical stress; pipe age and material influence failure risk; and equipment/site constraints connect project conditions to operational inefficiency.

For civil engineers and project managers, these explanations are more valuable than a prediction alone. A high-energy prediction caused mainly by flow and pressure may call for pump scheduling or pressure-zone review.

A prediction caused by leakage, vibration and pipe age may call for inspection, leak isolation, material replacement or preventive maintenance. A high-risk project classification caused by safety observations, site constraints and equipment hours may require work-package resequencing, additional supervision or trench-safety controls.

Table 8. Top explainability drivers for pumping-energy demand.

Feature	Importance_RMSE_Increase
Flow_Rate_Lps	492.32
Pressure_kPa	75.88
Leakage_Flag	49.63
RPM	18.60
Length_m	8.82
Pipe_Age_Years	7.50
Pump_Station	7.06
Zone	6.89
Slope	5.82
Elevation_m	4.07

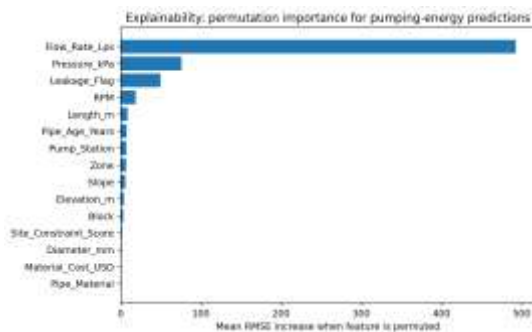


Figure 7. Explainability ranking using permutation importance.

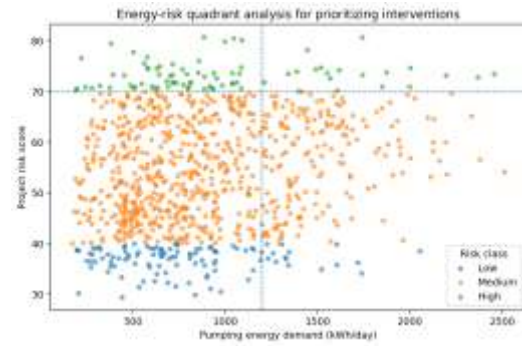


Figure 8. Energy-risk quadrant analysis for intervention prioritization.

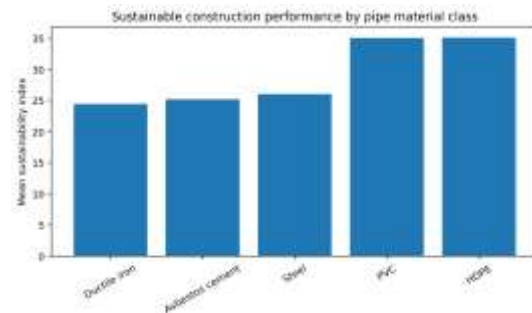


Figure 9. Sustainable construction performance by pipe material class.

## VI. DISCUSSION

### 6.1 Implications for water and sewer reticulation delivery

The results support the central argument that energy efficiency and project risk should be managed together. In water and sewer reticulation, pumping demand is not merely an operational cost after construction. It is affected by design assumptions, pressure management, pipe material condition, leakage response, equipment scheduling, dewatering decisions and commissioning quality. An explainable model can help field engineers identify where poor hydraulic performance and poor project controls reinforce each other.

The zone-block heat map demonstrates the practical value of spatial aggregation. Project managers can use such heat maps to prioritize areas requiring deeper investigation, additional supervision, pump-efficiency review or safety intervention. The same logic can be applied to municipal asset management by ranking pipe segments according to energy-risk

exposure, leakage likelihood and sustainability index. Rather than responding only to failures, the utility or contractor can identify clusters where pressure drop, vibration, aging assets and schedule stress create preventable risk.

The framework also supports sustainable construction performance. Civil infrastructure projects increasingly need to document not only cost and completion dates but also emissions, resilience, resource efficiency and public-service continuity. By connecting pumping energy, carbon proxy, risk score and schedule/cost variance into a sustainable performance index, the model creates a practical bridge between project management and environmental performance.

This is consistent with broader work on sustainable infrastructure and AI-supported infrastructure resilience, including geospatial and predictive analytics approaches in recent infrastructure scholarship (Alabi, 2024; Dongo et al., 2026).

### 6.2 Practical implementation roadmap

Table 9. Implementation roadmap for civil engineering and utility teams.

Stage	Action	Expected output
1. Data readiness	Inventory sensors, pump logs, SCADA exports, GIS pipe maps, work packages, costs, safety observations and drawings	Data dictionary and governance plan
2. Pilot modelling	Train baseline and boosted-tree models on one district or project package	Validated RMSE, F1, confidence flags and dashboard prototype
3. Explainability review	Review feature attributions with engineers, supervisors and utility operators	Accepted causal/engineering interpretation rules

4. Field integration	Link predictions to scheduling, pump review, leak inspection and safety plans	Energy-aware project control process
5. Continuous learning	Retrain models after commissioning, repairs, incidents and new sensor installations	Updated asset-risk register and model audit trail

Implementation should start with data governance rather than model selection. Pressure sensors, flow meters, SCADA logs, work-package records and cost/schedule data often use different identifiers. A location code must therefore connect hydraulic evidence to construction packages and asset records. Once a common key exists, the model can be trained, explained, validated and embedded into routine project controls.

The framework is especially useful for small and medium-sized utilities or contractors that cannot immediately deploy a fully integrated digital twin. A staged approach can begin with spreadsheet-level datasets and dashboards, then gradually incorporate GIS, SCADA, temporal deep sequence models and graph neural networks as data maturity improves. This mirrors low-cost analytics and dashboard principles in environmental and infrastructure performance research, where the initial objective is not technological complexity but decision usefulness (Ndlovu et al., 2025).

### VII. LIMITATIONS AND FUTURE RESEARCH

This study has four limitations. First, the integrated panel is a reproducible experimental dataset constructed from public Kaggle and open-data schemas rather than a direct raw utility dataset. The modelling results should therefore be treated as methodological evidence and proof of concept, not as measured performance from a specific water authority. Second, the graph neural network component is specified conceptually because a

complete reticulation topology file with nodes, edges and calibrated hydraulic parameters was not available. Third, the sustainability index uses an analytical carbon proxy and should be replaced with jurisdiction-specific electricity emissions factors and contractor energy records in a deployed study.

Fourth, project risk labels are derived from a transparent composite score; in real deployment, labels should be validated against actual incidents, delays, claims, inspections, safety outcomes and operations reports.

Future research should use raw Kaggle CSV downloads, Mendeley wastewater energy records, SCADA sensor streams and GIS topology files to train a fully integrated spatiotemporal graph model.

A strong next step would be to represent the water and sewer network as a graph and combine temporal LSTM/Transformer layers with GNN message passing. Another useful extension would include computer vision from field progress photos and trench-safety images, consistent with construction-site quality and safety monitoring. Finally, future work should test explainability outputs with practising civil engineers to determine whether the explanations are actionable, credible and aligned with engineering judgment.

### CONCLUSION

This paper developed an explainable deep-learning framework for energy-efficient water and sewer reticulation infrastructure. The central contribution is the integration of hydraulic sensor analytics, pumping energy prediction, project-risk classification and sustainable construction performance into one applied civil-engineering decision system.

The analysis shows that machine-learning and deep-learning models can predict pumping energy and project risk from combined hydraulic, asset and project-control variables, while explainability methods can convert predictions into actionable engineering insight.

The framework can be used by utilities, contractors, consultants and project managers to prioritize leak-

prone segments, forecast pumping energy, identify high-risk work packages, improve trench-safety planning, optimize pump scheduling, reduce rework, and strengthen sustainability reporting.

With validated field data, GIS topology and raw utility records, the proposed model can be converted from a proof-of-concept analytical framework into a deployable decision-support system for resilient, energy-efficient water and sewer infrastructure delivery.

### Appendix A. Reproducibility Summary

Table A1. Reproducibility summary.

Item	Description
Observations	5,000 segment-day records
Segments	250 pipe segments across 5 zones, 5 blocks per zone and 10 pipes per block
Temporal window	20 daily observations per segment
Primary regression target	Pump_Energy_kWh_day
Primary classification target	Project_Risk_Class: Low, Medium, High
Sustainability target	Sustainable_Performance_Index from 0 to 100
Energy models	Ridge, Random Forest, Gradient Boosting, XGBoost, Deep MLP, temporal lag-window MLP
Risk models	Logistic Regression, Random Forest, XGBoost, Deep MLP
Explainability method	Permutation importance; SHAP recommended for final raw-data deployment
Recommended raw-data upgrade	Download Kaggle raw CSVs, Mendeley WWTP data and GIS network topology; rerun pipeline against original files

### REFERENCES

- [1] Alabi, M. (2024) ‘Sustainable Infrastructure and Climate Change Resilience’, ResearchGate. Available at: <https://www.researchgate.net/publication/38528>

- 7202\_Sustainable\_Infrastructure\_and\_Climate\_Change\_Resilience (Accessed: 10 June 2026).
- [2] Alali, Y., Harrou, F. and Sun, Y. (2023) 'Unlocking the potential of wastewater treatment: machine learning based energy consumption prediction', *Water*, 15(13), 2349.
- [3] ASCE (2025a) 'US Drinking Water Infrastructure', 2025 Infrastructure Report Card. Available at: <https://infrastructurereportcard.org/cat-item/drinking-water-infrastructure/> (Accessed: 10 June 2026).
- [4] ASCE (2025b) 'US Wastewater Infrastructure', 2025 Infrastructure Report Card. Available at: <https://infrastructurereportcard.org/cat-item/wastewater-infrastructure/> (Accessed: 10 June 2026).
- [5] Bagehrzadeh, F. (2021) Full Scale Wastewater Treatment Plant Data. Mendeley Data, V1. doi: 10.17632/pprkzv3vbd.1.
- [6] Barros, D., et al. (2024) 'Leak detection and localization in water distribution systems using graph-based methods', available via PubMed Central. Available at: <https://pmc.ncbi.nlm.nih.gov/articles/PMC11647635/> (Accessed: 10 June 2026).
- [7] Breiman, L. (2001) 'Random forests', *Machine Learning*, 45, pp. 5-32.
- [8] Chen, T. and Guestrin, C. (2016) 'XGBoost: A scalable tree boosting system', *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794.
- [9] Copeland, C. (2017) Energy-Water Nexus: The Water Sector's Energy Use. Congressional Research Service.
- [10] d4rkLucif3r (n.d.) Full Scale Waste Water Treatment Plant Data. Kaggle. Available at: <https://www.kaggle.com/datasets/d4rkLucif3r/full-scale-waste-water-treatment-plant-data> (Accessed: 10 June 2026).
- [11] Dongo, M.M., Mupa, M.N., Musariri, T. and Chingarandi, F. (2026) 'Integrating GIS, RTK GNSS, and subsurface geophysical surveying for smart, climate-resilient infrastructure planning in underserved U.S. Regions', *World Journal of Advanced Research and Reviews*, 29(03), pp. 1618-1627. doi: 10.30574/wjarr.2026.29.3.0637.
- [12] Ehrhard, R., et al. (2009) 'Driving energy efficiency in the U.S. water and wastewater industry by focusing on operating and maintenance cost reductions', *ACEEE Summer Study on Energy Efficiency in Industry*.
- [13] Hochreiter, S. and Schmidhuber, J. (1997) 'Long short-term memory', *Neural Computation*, 9(8), pp. 1735-1780.
- [14] Javadiha, M., Blesa, J., Soldevila, A. and Puig, V. (2019) 'Leak localization in water distribution networks using deep learning', 6th International Conference on Control, Decision and Information Technologies.
- [15] Khaled, Q., De Marinis, P., Louati, M., Ferras, D., Genga, L. and Kaymak, U. (2026) 'Explainable Fuzzy GNNs for Leak Detection in Water Distribution Networks', arXiv:2601.03062.
- [16] Kipf, T.N. and Welling, M. (2017) 'Semi-supervised classification with graph convolutional networks', *International Conference on Learning Representations*.
- [17] Li, X., et al. (2024) 'A convolutional graph neural network model for water distribution network leakage detection', *Water*, 16(24), 3555.
- [18] Lundberg, S.M. and Lee, S.-I. (2017) 'A unified approach to interpreting model predictions', *Advances in Neural Information Processing Systems*, 30.
- [19] Mupa, M.N. (n.d.) Munashe Naphtali Mupa - Google Scholar profile. Available at: <https://scholar.google.com/citations?hl=en&user=JmqLHQAAAAJ> (Accessed: 10 June 2026).
- [20] Ndlovu, N.T., Mupa, M.N., Mtisi, B.N., Nemure, T. and Hamunakwadi, L.K. (2025) 'A low-cost analytics stack for county-scale organics diversion and anaerobic digestion performance: design, validation, and policy KPIs', *International Journal of Management and Organizational Research*, 4(5), pp. 64-70. doi: 10.54660/IJMOR.2025.4.5.64-70.

- [21] programmer3 (n.d.) Construction Project Management Dataset. Kaggle. Available at: <https://www.kaggle.com/datasets/programmer3/construction-project-management-dataset> (Accessed: 10 June 2026).
- [22] Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) 'Why should I trust you? Explaining the predictions of any classifier', Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135-1144.
- [23] Sukrut10k (2025) Smart-Water-Leak-Detection. GitHub repository. Available at: <https://github.com/Sukrut10k/Smart-Water-Leak-Detection> (Accessed: 10 June 2026).
- [24] talha97s (n.d.) Smart Water Leak Detection Dataset. Kaggle. Available at: <https://www.kaggle.com/datasets/talha97s/smart-water-leak-detection-dataset> (Accessed: 10 June 2026).
- [25] Taloma, R.J.L., et al. (2025) 'Machine learning for smart water distribution systems: an application-oriented survey', Artificial Intelligence Review.
- [26] U.S. EPA (n.d.) Energy Efficiency for Water Utilities. Available at: <https://www.epa.gov/sustainable-water-infrastructure/energy-efficiency-water-utilities> (Accessed: 10 June 2026).
- [27] Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P. and Bengio, Y. (2018) 'Graph attention networks', International Conference on Learning Representations.
- [28] Zhang, Z. and Fink, O. (2024) 'Algorithm-informed graph neural networks for leakage detection and localization in water distribution networks', arXiv:2408.02797.