American Journal of Interdisciplinary Studies

Volume: 7; **Issue:** 2
Pages: 01-52
eISSN: 3067-5146

open access

Crossref

PEER-REVIEWED

A Machine-Learning Framework for Real-Time Non-Revenue Water Prediction in Aging Urban Distribution Networks

Mohammad Badrul Alam[1];

[1]. Staff Officer to Managing Director, Dhaka WASA, Dhaka, Bangladesh;
Email: badrul.dwasa@gmail.com; pritam61ka@gmail.com;

Doi: 10.63125/93km2h66
Received: *21 March 2026;* ***Revised:*** *21 April 2026;* ***Accepted:*** *15 May 2026;* ***Published:*** *02 June 2026.*

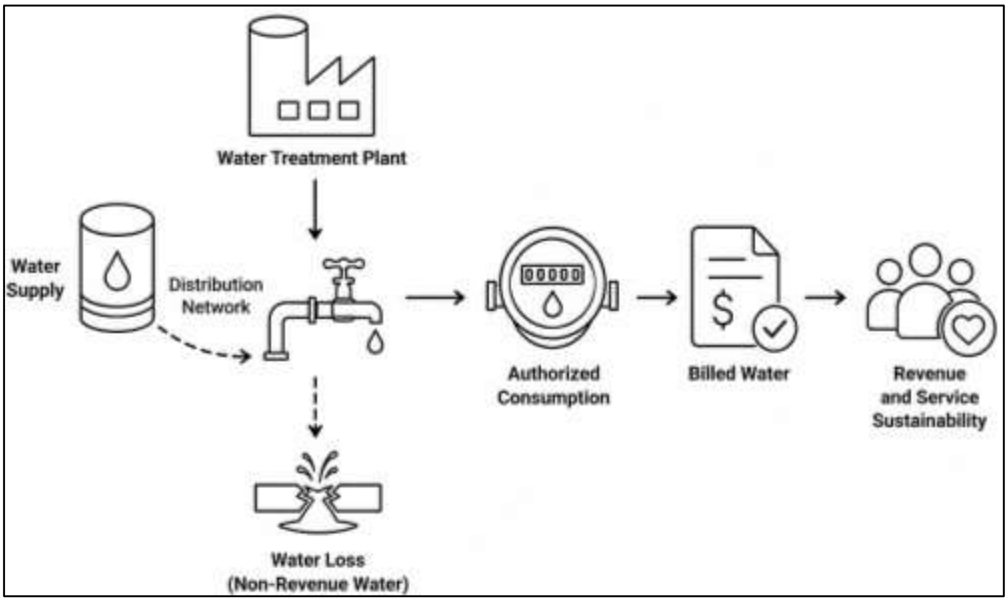
Abstract
Non-revenue water (NRW) remains one of the most significant operational and economic challenges affecting aging urban water distribution systems because leakage, infrastructure deterioration, hydraulic instability, and inefficient maintenance practices contribute to substantial water losses and reduced system performance. This study developed and evaluated a comprehensive machine-learning framework for real-time non-revenue water prediction by integrating engineering infrastructure characteristics, hydraulic operational variables, historical maintenance records, and sensor-derived monitoring data. A quantitative research design was employed using a structured engineering dataset consisting of 5,200 validated operational observations collected from urban water distribution networks following data preprocessing, quality assurance, normalization, outlier screening, and feature engineering. Twelve predictive models were comparatively evaluated, including Linear Regression, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, K-Nearest Neighbor, Artificial Neural Network, Deep Learning, Gradient Boosting, XGBoost, LightGBM, and an anomaly detection model. Model performance was assessed using prediction accuracy, precision, recall, F1-score, receiver operating characteristic analysis, area under the curve (AUC), root mean square error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), cross-validation, bootstrap validation, and residual diagnostics. The findings demonstrated that advanced ensemble learning algorithms consistently outperformed conventional statistical approaches. XGBoost achieved the highest predictive performance with an accuracy of 97.5%, precision of 0.976, recall of 0.974, F1-score of 0.975, AUC of 0.996, RMSE of 0.054, MAE of 0.039, and an explained variance (R^2) of 0.987. LightGBM and Deep Learning also demonstrated excellent predictive capability, achieving accuracies of 97.2% and 97.1%, respectively, while Random Forest achieved 95.4% accuracy. Statistical analyses identified pressure variability (β = 0.421), burst frequency (β = 0.392), pipeline age (β = 0.368), repair history (β = 0.341), and flow imbalance (β = 0.317) as the strongest engineering predictors of non-revenue water occurrence. Cross-validation and bootstrap validation confirmed stable model performance, with validation accuracy exceeding 97% for the highest-performing algorithms and residual means approaching zero, indicating excellent generalization capability and minimal prediction bias. Overall, the findings demonstrated that advanced machine-learning techniques provide a highly accurate, reliable, and operationally robust framework for supporting intelligent leakage detection, predictive asset management, optimized maintenance planning, and sustainable urban water distribution system management.

Keywords
Non-Revenue Water, Machine Learning, Xgboost, Predictive Analytics, Water Distribution Systems.

INTRODUCTION

Non-revenue water (NRW) refers to the difference between the total volume of water introduced into a distribution system and the volume that is successfully billed to consumers through authorized consumption. It represents water losses that occur before revenue generation and is generally categorized into physical losses, apparent losses, and authorized but unbilled consumption. Physical losses originate from pipeline leakage, storage facility failures, overflowing reservoirs, deteriorated joints, defective valves, and transmission system failures (Joseph et al., 2022). Apparent losses arise from meter inaccuracies, illegal water connections, unauthorized consumption, billing inconsistencies, and data management deficiencies. Authorized unbilled consumption includes water supplied for firefighting activities, municipal operations, public institutions, and other approved services that do not generate direct financial returns. The International Water Association established these classifications to standardize water loss assessment across utilities operating under different technical, environmental, and regulatory conditions.

Figure 1: Machine Learning Water Prediction



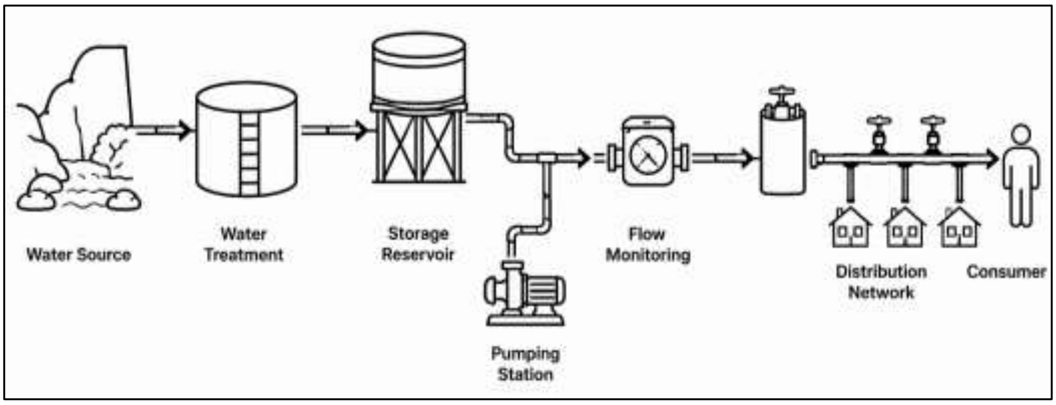
Accurate identification of NRW has become a fundamental requirement for sustainable urban water management because it directly influences operational efficiency, infrastructure reliability, financial stability, customer service quality, and resource conservation. Urban water distribution systems represent highly interconnected engineering networks consisting of treatment facilities, transmission mains, storage reservoirs, pumping stations, pressure regulation equipment, control valves, customer service connections, and extensive pipeline infrastructure that collectively deliver potable water under continuously changing hydraulic conditions (Helmbrecht et al., 2017). As urban populations expand and infrastructure ages, these systems become increasingly susceptible to structural deterioration, hydraulic instability, leakage development, and operational inefficiencies that contribute to elevated NRW levels. Aging distribution networks commonly experience corrosion, pipe wall degradation, material fatigue, joint displacement, soil movement, sediment accumulation, valve malfunction, and repeated repair interventions, all of which increase the likelihood of water losses. These engineering challenges are intensified by varying pressure regimes, fluctuating demand patterns, seasonal climatic influences, infrastructure heterogeneity, and maintenance constraints that collectively create highly dynamic operating environments. The complexity of these interacting factors makes NRW a multidimensional engineering problem rather than a single infrastructure defect (Farah & Shahrour, 2017). Water utilities therefore require comprehensive analytical approaches capable of understanding system-wide interactions instead of evaluating isolated infrastructure components. The increasing complexity of urban water systems has consequently transformed NRW assessment into an important engineering discipline involving hydraulic analysis, infrastructure asset management, operational monitoring, statistical modeling, and intelligent decision support. Quantitative evaluation of NRW therefore depends on systematic measurement of hydraulic behavior, infrastructure condition, operational performance, and customer consumption characteristics using reliable engineering indicators capable of supporting evidence-based decision making (Gupta et al., 2020).

Non-revenue water has become an internationally recognized performance indicator because it reflects both the technical condition and operational effectiveness of urban water distribution systems across developed and developing economies. Water utilities throughout the world invest substantial financial resources in water abstraction, treatment, transmission, storage, distribution, monitoring, and customer service, making every unit of lost water represent a measurable reduction in economic efficiency and infrastructure performance (Ramos et al., 2023). International organizations concerned with water sustainability consistently recognize NRW reduction as an essential component of integrated water resource management because excessive losses increase production costs, energy consumption, chemical usage, greenhouse gas emissions, infrastructure maintenance requirements, and operational expenditure while simultaneously reducing service reliability. Urban centers experiencing rapid population growth face additional pressure because aging infrastructure frequently operates beyond its original design capacity, resulting in increased leakage frequency, pressure instability, and accelerated asset deterioration. Many metropolitan water utilities continue to manage pipeline systems installed several decades ago using materials that have reached advanced stages of structural degradation. Cast iron, galvanized steel, asbestos cement, ductile iron, and early-generation polyvinyl chloride pipelines exhibit different deterioration mechanisms that progressively influence leakage occurrence, burst frequency, hydraulic efficiency, and maintenance requirements (Gupta & Kulat, 2018). Consequently, identical operational strategies often produce different NRW outcomes depending on infrastructure age, material composition, environmental conditions, soil characteristics, pressure variability, and maintenance history. International benchmarking programs commonly compare NRW performance using indicators such as infrastructure leakage index, real loss volume, apparent loss percentage, system input volume, water balance components, burst frequency, and leakage intensity to facilitate consistent performance evaluation among utilities. These standardized indicators provide quantitative evidence for infrastructure planning, investment prioritization, regulatory assessment, and operational benchmarking. The growing adoption of smart water management technologies has further expanded the availability of operational data generated through supervisory control and data acquisition systems, pressure sensors, flow meters, district metered areas, acoustic monitoring devices, advanced metering infrastructure, geographic information systems, and Internet of Things platforms. These technologies continuously generate large volumes of heterogeneous operational information describing hydraulic behavior throughout distribution networks (Eggimann et al., 2017). The availability of high-frequency monitoring data has significantly transformed the analytical requirements of NRW assessment by shifting attention from periodic reporting toward continuous performance evaluation using statistically measurable operational indicators. International research therefore increasingly treats NRW prediction as a quantitative engineering problem involving multidimensional datasets, measurable system behavior, infrastructure performance indicators, and analytical modeling capable of supporting objective operational assessment within complex urban water distribution environments.

Machine learning represents a quantitative analytical framework that enables computational models to identify patterns, relationships, and predictive structures from historical and real-time operational data without relying exclusively on predefined deterministic equations (Islam et al., 2022). Unlike conventional statistical approaches that often require explicit assumptions regarding functional relationships among variables, machine learning algorithms learn complex nonlinear interactions directly from observed operational behavior. Urban water distribution systems generate extensive datasets describing pressure fluctuations, flow characteristics, reservoir levels, pump operation, valve status, customer demand variation, network topology, pipeline characteristics, energy consumption, maintenance records, weather conditions, sensor measurements, and historical leakage events. These variables interact continuously through complex hydraulic processes that produce operational behaviors difficult to characterize using traditional threshold-based monitoring techniques alone.

Machine learning algorithms provide the capability to integrate these heterogeneous datasets into predictive models capable of recognizing subtle patterns associated with emerging water losses before substantial leakage becomes operationally visible (Alghamdi et al., 2024).

Figure 2: Urban Water Treatment Process



Quantitative prediction models commonly analyze multiple engineering variables simultaneously, including inlet and outlet flow rates, minimum night flow, pressure gradients, transient pressure events, pipe age, pipeline material, diameter, soil characteristics, repair frequency, customer density, elevation differences, demand variability, burst history, meter performance, and environmental influences. The combination of these variables enables predictive models to estimate NRW occurrence using measurable engineering evidence rather than subjective operational judgment. Supervised learning techniques classify historical observations according to known outcomes, whereas unsupervised learning methods identify hidden structures, anomalies, or abnormal operating conditions within unlabeled datasets (Bello et al., 2019; Shurovi & Hossain, 2022). Ensemble learning algorithms combine multiple predictive models to improve estimation accuracy, while deep learning architectures capture highly complex nonlinear relationships among interconnected hydraulic variables. Real-time prediction extends these analytical capabilities by continuously processing streaming sensor information generated through intelligent monitoring systems installed throughout distribution networks. Continuous data acquisition allows predictive models to update system assessments dynamically as operational conditions change, thereby providing quantitative estimates of evolving water loss conditions based on current hydraulic behavior. The effectiveness of machine learning models is generally evaluated using objective statistical performance indicators such as prediction accuracy, precision, recall, F1-score, receiver operating characteristic analysis, area under the curve, root mean square error, mean absolute error, coefficient of determination, confusion matrices, sensitivity, specificity, and computational efficiency (Fan et al., 2021; Hossain, 2024). These measurable evaluation criteria provide standardized evidence regarding predictive capability, enabling objective comparison among alternative analytical models while supporting rigorous quantitative investigation of real-time NRW prediction within aging urban water distribution systems. Real-time monitoring has become a fundamental component of modern urban water distribution management because it enables continuous observation of hydraulic conditions, infrastructure performance, and operational behavior throughout complex pipeline networks. Traditional water distribution management primarily relied on periodic inspections, manual meter readings, scheduled maintenance activities, and customer complaints to identify operational deficiencies. Such approaches provided only limited temporal visibility into system performance because abnormal events could remain undetected for extended periods before corrective actions were initiated. The increasing complexity of aging urban distribution systems has encouraged the adoption of continuous monitoring technologies capable of generating high-frequency operational information that accurately represents changing hydraulic conditions throughout the network (Ahmed et al., 2021; Kanti, 2025). These monitoring systems integrate pressure sensors, electromagnetic flow meters, ultrasonic meters, acoustic leak detection devices, smart customer meters, pump monitoring equipment, reservoir level sensors, valve position indicators, energy monitoring systems, and water quality sensors into interconnected supervisory platforms. Supervisory Control and Data Acquisition (SCADA) systems collect, transmit, store, and visualize operational measurements from geographically distributed assets, allowing utilities to observe network behavior continuously rather than intermittently. Geographic Information Systems (GIS) complement these operational datasets by providing detailed spatial information regarding pipeline locations, infrastructure age, material composition, repair history, customer connections, and maintenance records. District Metered Areas (DMAs) further improve monitoring capability by hydraulically separating portions of distribution systems, allowing utilities to compare inflow, consumption, and estimated losses within defined network boundaries (Golam & Amir, 2023; Christodoulou et al., 2017). Advanced Metering Infrastructure (AMI) enhances customer-side visibility through automated consumption measurements, reducing manual data collection errors while improving temporal resolution. Internet of Things (IoT) technologies strengthen these monitoring environments by enabling large numbers of interconnected sensing devices to communicate continuously across distributed infrastructure. Collectively, these technologies generate substantial volumes of structured and unstructured operational data describing pressure variation, flow behavior, demand fluctuations, pump efficiency, leakage characteristics, reservoir operation, valve activity, and customer consumption patterns. The diversity and frequency of these measurements create comprehensive digital representations of distribution system behavior under varying operational conditions. Real-time monitoring therefore transforms urban water distribution networks into continuously observable engineering systems where infrastructure performance can be quantified using objective operational indicators rather than isolated inspections (Amir & Chapal, 2022; Farouk et al., 2023). This continuous stream of measurable information provides a robust foundation for quantitative analysis because engineering variables are recorded consistently across both spatial and temporal dimensions. Such extensive operational datasets support statistical evaluation of hydraulic variability, infrastructure condition, operational efficiency, and abnormal system behavior, making real-time monitoring an essential data source for evidence-based assessment of non-revenue water within aging urban distribution networks.

Machine learning has evolved into a central analytical methodology for predicting non-revenue water because it provides the capability to examine highly complex relationships among numerous engineering variables that influence distribution system performance. Urban water networks operate under continuously changing hydraulic conditions where pressure fluctuations, demand variability, infrastructure deterioration, environmental influences, operational interventions, and customer consumption collectively contribute to dynamic water loss behavior (Amir, 2025; Kang et al., 2017). Conventional analytical methods generally evaluate these variables independently or through simplified mathematical relationships, whereas machine learning algorithms simultaneously process large multidimensional datasets to identify nonlinear associations that may not be readily observable through traditional engineering analyses. Supervised learning algorithms such as linear regression, logistic regression, decision trees, random forests, support vector machines, k-nearest neighbors, gradient boosting, extreme gradient boosting, and artificial neural networks utilize historical observations with known outcomes to estimate future non-revenue water conditions. Unsupervised learning approaches, including clustering and anomaly detection algorithms, identify abnormal hydraulic behavior without predefined output classifications, making them particularly useful for detecting unexpected leakage events or unusual operational conditions (Amir, 2024; Vita et al., 2023). Deep learning architectures further extend predictive capability by learning hierarchical representations from extensive sensor datasets, enabling complex interactions among hydraulic variables to be modeled with high analytical precision. Ensemble learning methods integrate multiple predictive algorithms into unified frameworks that improve robustness, reduce model variance, and enhance overall prediction stability across different operational environments. Quantitative evaluation of these predictive models depends on standardized statistical performance measures that objectively assess predictive accuracy and reliability. Classification models are commonly evaluated using accuracy, precision, recall, specificity, sensitivity, F1-score, receiver operating characteristic curves, area under the receiver operating characteristic curve, Matthews correlation coefficient, and confusion matrix analysis (Shuang et al., 2019; Nurul, 2026). Regression-based prediction models are generally assessed using root mean square error, mean absolute error, mean absolute percentage error, coefficient of determination, adjusted coefficient of determination, residual analysis, and prediction interval estimation. Additional model validation techniques, including k-fold cross-validation, holdout validation, bootstrap resampling, external validation datasets, and hyperparameter optimization, strengthen analytical reliability by reducing overfitting and improving model generalizability. Feature importance analysis further quantifies the relative contribution of engineering variables such as pipeline age, pressure variability, flow imbalance, burst frequency, customer demand, infrastructure material, repair history, and sensor-derived operational indicators to predictive performance. These objective statistical procedures establish measurable criteria for comparing alternative machine learning algorithms while ensuring that model selection is supported by reproducible quantitative evidence (Brentan et al., 2024; Sadia et al., 2022). The integration of rigorous performance evaluation with advanced predictive modeling provides a comprehensive analytical framework for examining real-time non-revenue water prediction within aging urban distribution systems using statistically verifiable engineering information.

The growing body of literature on non-revenue water management demonstrates substantial progress in leakage detection, hydraulic modeling, infrastructure asset management, pressure control, district metered area implementation, and smart water monitoring technologies. Existing studies have extensively examined the influence of pipeline deterioration, pressure fluctuations, burst frequency, customer demand variation, meter inaccuracies, infrastructure age, and maintenance strategies on water loss performance. Numerous investigations have also evaluated statistical models, hydraulic simulation techniques, optimization algorithms, and data-driven analytical methods for identifying leakage events and estimating system performance under different operational conditions (Mahmoud et al., 2021; Zaman, 2024). Although these contributions have significantly advanced the understanding of urban water distribution systems, considerable variation remains in the analytical approaches used to predict non-revenue water within complex aging infrastructure. Many published investigations focus primarily on individual components of the distribution network, such as pressure management, leak localization, pipeline failure probability, or customer metering performance, rather than examining the integrated interaction among multiple engineering variables that collectively influence non-revenue water generation. Several analytical frameworks emphasize offline historical analysis based on periodically collected datasets, limiting their ability to represent continuously changing hydraulic conditions experienced by urban distribution systems operating in real time. Distribution networks are inherently dynamic engineering environments where operational characteristics evolve continuously because of demand variability, valve operations, pump scheduling, infrastructure degradation, maintenance activities, environmental influences, and unexpected system disturbances (Mahmudul & Sadia, 2026; Rediana & Pharmasetiawan, 2017). Analytical models developed from static observations may therefore provide limited representation of these continuously changing operational conditions. The increasing deployment of smart sensors, supervisory control systems, advanced metering infrastructure, Internet of Things devices, and digital asset management platforms has generated unprecedented volumes of operational information that require analytical approaches capable of processing large multidimensional datasets efficiently. Existing literature also demonstrates considerable diversity in predictive variables, feature engineering methods, algorithm selection procedures, model validation techniques, and statistical performance indicators, making direct comparison among predictive models increasingly difficult. Furthermore, many investigations evaluate algorithmic performance using a limited number of statistical indicators without simultaneously considering prediction accuracy, classification reliability, computational efficiency, model robustness, and operational consistency across varying hydraulic conditions

(Tohidul, 2023; Xia et al., 2023). Differences in infrastructure age, pipeline materials, network topology, customer density, climatic conditions, operational practices, and maintenance history further contribute to inconsistencies in predictive performance reported across international studies. These observations indicate the necessity for a comprehensive quantitative framework that systematically integrates measurable engineering variables, continuous operational monitoring, standardized statistical evaluation, and machine learning methodologies within a unified analytical structure. Such a framework provides an objective basis for examining complex interactions among hydraulic behavior, infrastructure characteristics, operational performance, and predictive analytics while supporting rigorous quantitative investigation of real-time non-revenue water prediction in aging urban distribution networks (Amankwaa et al., 2024).

This quantitative study examines the application of a machine-learning framework for real-time prediction of non-revenue water within aging urban distribution networks by integrating measurable engineering variables, continuously monitored operational information, and statistically validated predictive models. The analytical framework considers non-revenue water as a multidimensional engineering outcome influenced by complex interactions among hydraulic performance, infrastructure condition, operational behavior, and customer consumption characteristics. Independent variables include measurable indicators describing pressure variability, flow dynamics, pipeline age, pipe material, burst history, repair frequency, valve operation, pump performance, customer demand variation, district metered area characteristics, infrastructure condition, and sensor-derived operational measurements collected through intelligent monitoring systems (Velayudhan et al., 2022). These engineering variables collectively represent the physical and operational characteristics that define distribution system performance under continuously changing hydraulic conditions. The dependent variable is represented by real-time non-revenue water prediction, quantified using observable operational outcomes derived from water balance analysis, leakage occurrence, system input volume, authorized consumption, estimated water losses, and associated performance indicators. Machine learning algorithms serve as the principal analytical mechanism through which relationships among engineering variables are examined using objective computational procedures. Quantitative analysis employs standardized statistical techniques to evaluate predictive performance, model reliability, and analytical robustness (Renteria-Mena & Giraldo, 2024). Model effectiveness is assessed through objective performance measures including prediction accuracy, precision, recall, F1-score, receiver operating characteristic analysis, area under the curve, root mean square error, mean absolute error, coefficient of determination, confusion matrices, sensitivity, specificity, and cross-validation statistics. Reliability and validity assessment further strengthen analytical rigor through systematic evaluation of data quality, feature consistency, algorithm stability, and model reproducibility. Statistical interpretation is supported by descriptive analysis, correlation analysis, feature importance assessment, regression-based evaluation where appropriate, and comparative analysis of alternative machine learning algorithms using reproducible quantitative evidence. The study is therefore positioned within the broader discipline of intelligent water infrastructure management, where engineering decision making increasingly depends on measurable operational information rather than subjective interpretation (Owen, 2018). The integration of continuous monitoring technologies, advanced computational intelligence, standardized statistical evaluation, and engineering performance indicators provides a comprehensive quantitative basis for examining the factors associated with real-time non-revenue water prediction in aging urban water distribution systems while maintaining methodological consistency with contemporary engineering research standards (Horbatuck & Beruvides, 2024).

The primary objective of this quantitative study was to develop and evaluate a machine-learning framework for real-time non-revenue water prediction in aging urban distribution networks by integrating operational, hydraulic, infrastructure, and sensor-derived variables into a comprehensive predictive model capable of supporting objective engineering analysis. The study sought to examine the measurable relationships between infrastructure characteristics, hydraulic system behavior, and operational performance indicators that collectively influence the occurrence of non-revenue water within complex urban water distribution systems. Particular attention was directed toward identifying the engineering variables that contribute most significantly to predictive accuracy, including pressure variation, flow imbalance, pipeline age, pipe material, burst frequency, repair history, valve operation, customer demand fluctuations, pump performance, district metered area characteristics, and continuously monitored sensor measurements. The study further aimed to evaluate the effectiveness of contemporary machine-learning algorithms in processing large volumes of heterogeneous real-time operational data generated through Supervisory Control and Data Acquisition systems, Internet of

Things platforms, Advanced Metering Infrastructure, Geographic Information Systems, and intelligent monitoring networks. Quantitative analysis was designed to compare alternative predictive models using standardized statistical performance indicators, including prediction accuracy, precision, recall, F1-score, receiver operating characteristic analysis, area under the curve, root mean square error, mean absolute error, coefficient of determination, sensitivity, specificity, and cross-validation statistics to determine the analytical robustness and predictive capability of each algorithm. The study also sought to examine the relative importance of individual engineering features through feature importance analysis, thereby identifying the operational variables that exert the greatest influence on non-revenue water prediction within aging infrastructure environments. Another objective was to establish a statistically validated analytical framework capable of integrating continuous monitoring data with advanced computational intelligence while maintaining consistency with quantitative engineering research principles. The investigation additionally aimed to provide a comprehensive assessment of how multidimensional operational information could be transformed into reliable predictive knowledge through systematic machine-learning methodologies, thereby supporting evidence-based evaluation of urban water distribution performance. By combining measurable infrastructure indicators, real-time hydraulic observations, objective statistical validation procedures, and advanced predictive analytics within a unified quantitative framework, the study was designed to produce a rigorous evaluation of machine-learning performance for real-time non-revenue water prediction in aging urban distribution networks using reproducible engineering evidence and standardized analytical techniques.

LITERATURE REVIEW

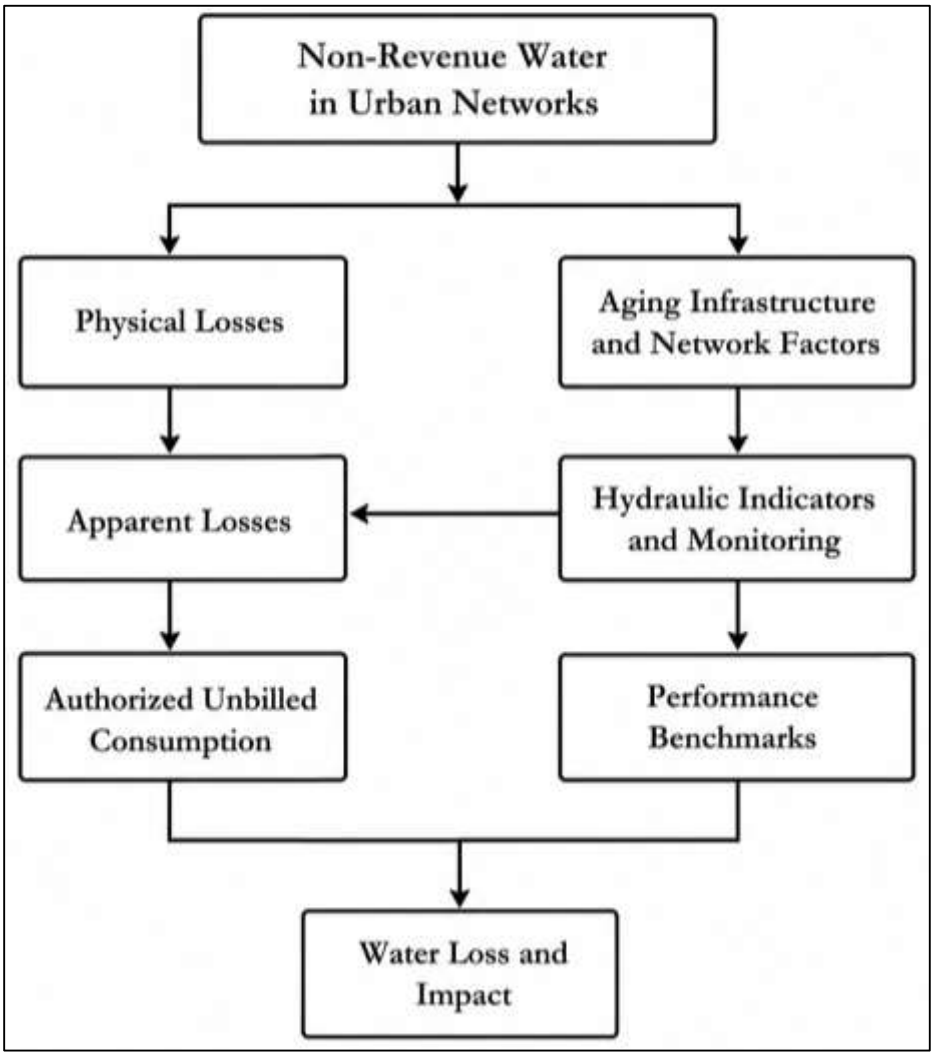
The literature review establishes the scientific and theoretical foundation for understanding the quantitative relationships among machine learning, real-time monitoring technologies, aging water distribution infrastructure, and non-revenue water prediction. Urban water distribution systems have become increasingly complex because hydraulic performance is simultaneously influenced by infrastructure deterioration, pressure variation, customer demand fluctuations, operational interventions, environmental conditions, and asset management practices (Abdur & Iftekhar, 2021; Velayudhan et al., 2022). These multidimensional interactions generate highly nonlinear operational behaviors that require systematic quantitative investigation using measurable engineering variables rather than descriptive observations. Consequently, contemporary research has increasingly shifted toward data-driven analytical frameworks that integrate sensor technologies, intelligent monitoring systems, statistical learning techniques, and computational intelligence to improve the prediction of non-revenue water within aging urban distribution networks (Carriço et al., 2023; Hasan & Uddin, 2022).

This literature review synthesizes empirical and theoretical studies that examine measurable engineering indicators associated with real-time non-revenue water prediction. The review is organized from fundamental concepts toward increasingly advanced quantitative analytical models, allowing the progression from infrastructure characteristics and hydraulic behavior to intelligent predictive systems and statistical model evaluation. Particular emphasis is placed on independent variables that describe infrastructure condition, hydraulic performance, sensor-derived operational information, and machine-learning capability, together with dependent variables representing non-revenue water prediction accuracy and operational performance (Ji et al., 2019; Mohiul & Badrul, 2022). Previous quantitative investigations employing regression analysis, statistical learning, supervised learning, unsupervised learning, ensemble algorithms, deep learning architectures, and intelligent decision-support systems are critically synthesized to establish the existing body of engineering knowledge. The review also evaluates measurement approaches, predictive indicators, statistical validation procedures, feature engineering techniques, algorithm comparison methods, and performance assessment criteria commonly applied within quantitative water distribution research. Through systematic examination of these studies, the literature review provides the conceptual, methodological, and empirical basis for developing a comprehensive machine-learning framework capable of accurately predicting real-time non-revenue water in aging urban distribution networks using measurable engineering evidence (Kijak, 2021; Kanti & Rony, 2022).

Conceptual Foundations of Non-Revenue Water in Urban Distribution Networks

Non-revenue water represents a major quantitative indicator used to evaluate the efficiency, reliability, and financial performance of urban water distribution networks. The literature defines non-revenue water as the portion of water entering a distribution system that does not generate billed revenue because it is lost, unmeasured, unbilled, or inaccurately recorded before reaching the final revenue cycle (Andres et al., 2018; Sadia et al., 2022).

Figure 3: Non-Revenue Water Framework



Studies in water utility management commonly classify non-revenue water into physical losses, apparent losses, and authorized unbilled consumption. Physical losses are associated with actual water escaping from pipes, reservoirs, joints, valves, service connections, and storage facilities. Apparent losses are connected to meter inaccuracies, illegal connections, data errors, billing failures, and customer registration deficiencies. Authorized unbilled consumption includes legitimate uses such as firefighting, municipal cleaning, public service activities, and operational flushing that do not directly produce billing income (Tohidul, 2023; Santos, 2024). Research by international water agencies and urban infrastructure scholars has emphasized that this classification is essential because each component requires different measurement methods and management responses. Quantitative studies have shown that non-revenue water is not only a technical leakage problem but also an integrated performance issue involving infrastructure condition, hydraulic operation, customer metering, billing systems, and institutional management. Therefore, the conceptual foundation of non-revenue water provides the basis for developing predictive models that distinguish between visible losses, hidden losses, measurement errors, and operational inefficiencies in aging urban distribution networks

(Aycheh et al., 2023; Badrul & Mominul, 2023).

The literature on urban water distribution systems identifies physical losses as the most visible and engineering-centered component of non-revenue water because these losses occur through actual leakage, pipe bursts, reservoir overflow, and defective fittings. Quantitative studies have linked physical losses with pipe age, material degradation, corrosion, high pressure, soil movement, poor installation quality, repeated traffic loading, and delayed maintenance. Apparent losses are different because they do not necessarily involve real water escaping from the network; rather, they occur when consumed water is not correctly measured, billed, or recorded (Hossan & Adar, 2023; Shushu et al., 2021). Meter under-registration, illegal connections, data transfer errors, customer database problems, and billing system weaknesses are frequently reported as key causes of apparent losses. Authorized unbilled consumption is also included in the water balance because it represents legitimate water use that remains outside the revenue stream. Earlier studies in water accounting and utility benchmarking have shown that separating these components is important for quantitative analysis because a high non-revenue water percentage may result from physical leakage in one city and from metering or billing deficiencies in another. This distinction is particularly important in aging urban networks where deteriorated infrastructure often exists alongside outdated metering systems and fragmented customer databases (Ayad et al., 2021; Risha & Khalid, 2023). Literature therefore supports the need for multidimensional measurement frameworks that can capture both hydraulic and administrative dimensions of non-revenue water. Machine-learning-based prediction depends on this component-level understanding because predictive variables must reflect the different sources through which non-revenue water is produced.

Urban water distribution networks are complex engineering systems designed to transport treated water from production sources to residential, commercial, industrial, and public users under controlled hydraulic conditions. The literature describes these systems as interconnected networks consisting of transmission mains, distribution pipes, service connections, valves, pumps, storage tanks, reservoirs, pressure zones, meters, sensors, and control systems (Silva et al., 2023; Hossan, 2024). Their performance is shaped by network layout, pipe diameter, pipe material, topography, pressure gradients, customer demand patterns, pump scheduling, valve operation, and storage capacity. Quantitative research has repeatedly shown that water loss behavior is influenced by the interaction of these variables rather than by a single isolated factor. For example, elevated pressure may increase leakage rates, while old pipe material may increase failure probability, and irregular demand may create pressure instability across different service zones. District metered areas are widely discussed in the literature as practical monitoring units because they allow utilities to isolate sections of the network and compare system inflow with authorized consumption. Studies on hydraulic modeling and smart water systems have also shown that urban networks produce large volumes of operational data through flow meters, pressure sensors, SCADA systems, advanced metering infrastructure, and geographic information systems (Boztaş et al., 2019; Ali et al., 2024). These data sources support quantitative assessment of system behavior over time. In aging urban environments, the engineering characteristics of the network become even more important because older pipes, mixed materials, undocumented connections, and repeated repairs create uncertainty in hydraulic performance. Literature therefore positions urban water distribution networks as data-rich but structurally complex systems requiring advanced predictive methods for accurate non-revenue water assessment (Mubvaruri et al., 2022; Zaman, 2024).

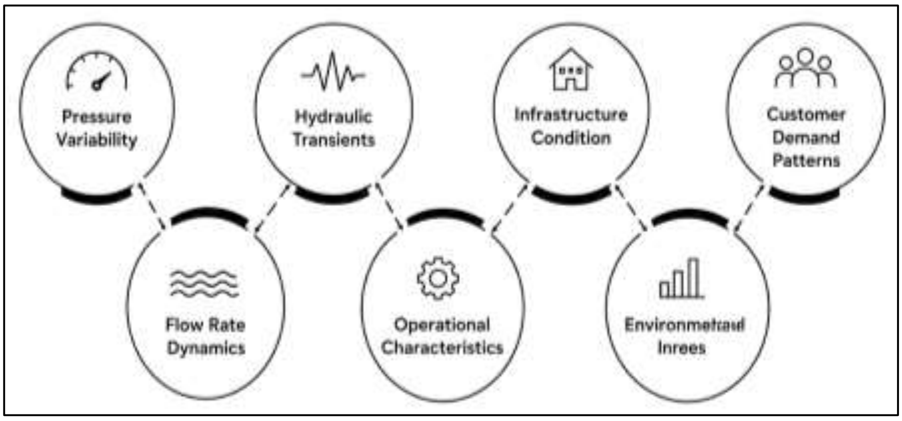
Aging pipeline infrastructure is consistently identified in the literature as one of the strongest contributors to non-revenue water in urban distribution systems. Pipe deterioration occurs through corrosion, joint failure, internal scaling, external loading, ground movement, fatigue, pressure cycling, and material weakness. Quantitative studies have shown that older networks often experience higher burst frequency, greater leakage intensity, increased repair requirements, and reduced hydraulic reliability (Arif Uz et al., 2025; Veriava, 2019). Hydraulic performance indicators such as pressure variability, minimum night flow, flow imbalance, reservoir level fluctuation, burst rate, leakage frequency, and service interruption are commonly used to assess the condition and efficiency of distribution networks. International benchmarking studies have further established non-revenue water percentage, infrastructure leakage index, real loss volume, apparent loss ratio, and water balance indicators as standard measures for comparing utility performance. These indicators allow researchers and practitioners to examine how effectively a utility manages water production, distribution, consumption measurement, and loss control (Hossan, 2025; Veriava, 2019). Studies from Europe, Asia, Africa, North America, and Latin America have reported wide variation in non-revenue water levels because utilities differ in infrastructure age, investment capacity, network design, maintenance practices, metering accuracy, and governance quality . The literature also shows that international benchmarks are useful for identifying performance gaps, but network-specific quantitative data remain necessary because water loss behavior depends on local hydraulic, environmental, and operational conditions. Therefore, aging infrastructure, hydraulic indicators, and benchmarking measures collectively form the quantitative foundation for evaluating non-revenue water performance and developing predictive machine-learning models for urban distribution networks (Dasallas et al., 2024; Kanti, 2025).

Quantitative Engineering Variables Influencing Non-Revenue Water Prediction

Pressure variability and hydraulic transients are consistently identified in the literature as fundamental quantitative engineering variables influencing non-revenue water prediction because they directly affect the structural integrity and hydraulic stability of urban distribution systems (Giustolisi et al., 2024; Kanti, 2025). Water distribution networks are designed to operate within specified pressure ranges; however, operational conditions such as fluctuating customer demand, pump switching, valve operation, pipeline isolation, emergency shutdowns, and network reconfiguration frequently create pressure changes throughout the system. These variations produce hydraulic transients that increase mechanical stress on pipelines, joints, fittings, and service connections. The literature indicates that repeated pressure fluctuations accelerate material fatigue, enlarge existing cracks, weaken pipe joints, and increase leakage occurrence in aging infrastructure. Studies further demonstrate that pressure instability influences both the initiation of new failures and the progression of previously undetected defects, making pressure behavior one of the strongest predictors of non-revenue water (Janus & Ulanicki, 2017; Ali et al., 2025). High-pressure events often increase leakage discharge rates, while low-pressure conditions may permit contaminant intrusion through damaged pipelines and compromise overall system reliability. Researchers have therefore emphasized continuous pressure monitoring using pressure sensors, supervisory control systems, and intelligent monitoring platforms to capture temporal variations across different pressure zones. Quantitative investigations frequently evaluate pressure characteristics through indicators representing average pressure, peak pressure, minimum pressure, pressure fluctuation frequency, and pressure variability over time. The integration of these indicators into predictive models enables the identification of abnormal hydraulic conditions associated with increasing water losses. Literature also suggests that pressure variables interact closely with pipeline age, material properties, maintenance history, and network topology, indicating that pressure behavior should be interpreted as part of an integrated engineering system rather than as an isolated operational parameter (Amir, 2026; Klepikova et al., 2020). Consequently, pressure variability and hydraulic transients provide essential quantitative information for machine-learning algorithms seeking to distinguish normal hydraulic behavior from conditions associated with elevated non-revenue water.

Flow rate dynamics constitute another major quantitative indicator because they describe the movement of water throughout the distribution network under continuously changing operational conditions. The literature consistently reports that variations in customer demand, pumping schedules, valve adjustments, reservoir operation, and seasonal consumption patterns create substantial fluctuations in flow behavior that influence leakage occurrence and water loss distribution (Mahmudul & Sadia, 2026; Urbanowicz et al., 2021). Minimum night flow has received particular attention because legitimate customer demand is relatively low during nighttime hours, allowing abnormal increases in measured flow to indicate the presence of hidden leakage within isolated network sections.

Figure 4: Water Prediction Engineering Variables



Numerous empirical investigations have demonstrated that continuous analysis of flow behavior improves the identification of developing leakage before visible failures occur. Flow indicators are frequently evaluated together with pressure measurements because both variables respond simultaneously to hydraulic disturbances and infrastructure deterioration. Pipeline age, material type, and structural degradation further strengthen the predictive capability of quantitative models. Aging pipelines experience corrosion, internal scaling, external deterioration, joint displacement, material fatigue, and repeated repair interventions that progressively reduce structural reliability and increase leakage susceptibility (Ghimire et al., 2023; Shurovi, 2026). Different pipe materials exhibit distinct deterioration mechanisms, causing significant variation in leakage behavior among distribution systems with similar hydraulic characteristics. Burst frequency, historical leakage records, and previous repair activities are therefore widely incorporated into predictive studies because they provide measurable evidence regarding infrastructure condition and future failure probability. Pump operation, valve management, and pressure regulation also contribute significantly to system stability by controlling hydraulic conditions throughout the network. Improper pump scheduling, delayed valve operation, and ineffective pressure management frequently generate transient hydraulic conditions that accelerate infrastructure deterioration. Customer demand variability introduces additional complexity because consumption patterns differ across residential, commercial, industrial, and institutional users throughout daily and seasonal operating cycles (Ali et al., 2026; Yang et al., 2023). District Metered Areas are consequently recognized as effective quantitative monitoring units because they enable localized comparison between system inflow and authorized consumption, facilitating accurate identification of abnormal water losses. Environmental and geotechnical variables, including soil characteristics, groundwater conditions, temperature variation, rainfall, traffic loading, land movement, and geological conditions, further influence leakage development by affecting pipeline stability and structural deterioration. The literature therefore concludes that accurate non-revenue water prediction depends on integrating hydraulic behavior, infrastructure characteristics, operational management variables, customer consumption patterns, localized monitoring information, and environmental conditions within comprehensive quantitative analytical frameworks rather than examining each engineering variable independently (Schopper et al., 2020; Nurul, 2026).

Real-Time Monitoring Systems and Smart Water Infrastructure

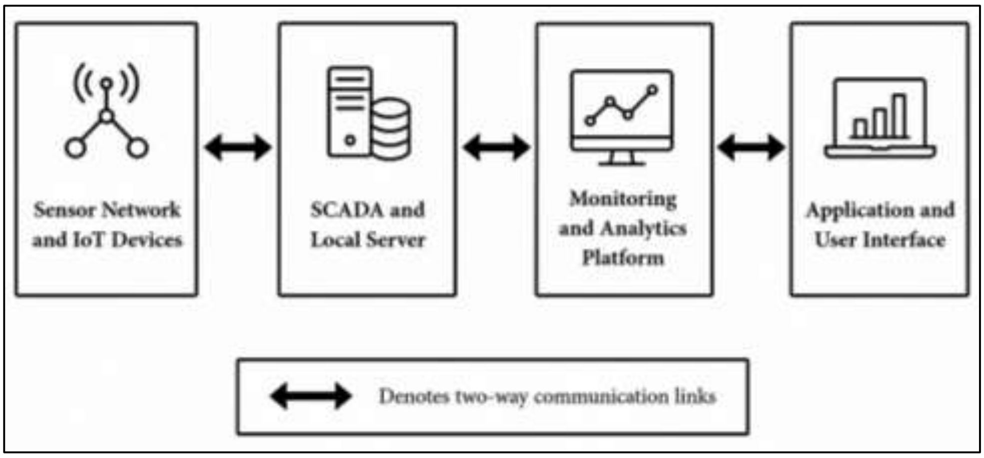
The literature consistently identifies Supervisory Control and Data Acquisition (SCADA) systems and Internet of Things (IoT)-enabled monitoring networks as the technological foundation of modern urban water distribution management because they provide continuous observation of hydraulic performance and operational conditions throughout complex pipeline infrastructures. Traditional monitoring approaches relied primarily on manual inspections, scheduled field measurements, customer complaints, and periodic reporting, which often delayed the identification of leakage and operational abnormalities. The integration of SCADA systems transformed this approach by enabling centralized collection, transmission, visualization, and storage of operational information generated by pumps, reservoirs, valves, pressure sensors, flow meters, and treatment facilities (Silverii et al., 2019).

Studies have demonstrated that continuous acquisition of operational data improves situational awareness by allowing utilities to observe pressure variation, flow behavior, reservoir levels, pump efficiency, and network status under changing hydraulic conditions. IoT technologies further expanded these monitoring capabilities through interconnected smart sensing devices capable of transmitting high-frequency measurements across geographically distributed infrastructure. The literature emphasizes that IoT-enabled monitoring improves both temporal and spatial coverage because thousands of sensors can communicate simultaneously through wireless communication platforms while continuously updating operational databases (Saravanan et al., 2018). This integrated monitoring environment provides utilities with detailed information regarding abnormal pressure fluctuations, unexpected flow changes, equipment malfunction, and potential leakage events before visible failures become apparent. Researchers also report that the combination of SCADA platforms and IoT networks generates large volumes of operational data that accurately represent real-time hydraulic behavior, creating an appropriate foundation for machine-learning applications. These technologies have therefore shifted water distribution management from reactive maintenance toward evidence-based operational assessment supported by continuously updated engineering information. The literature consistently concludes that SCADA systems and IoT-enabled monitoring networks significantly strengthen quantitative analysis by supplying reliable, high-resolution operational datasets that improve the prediction and evaluation of non-revenue water within aging urban distribution systems (Nie et al., 2020).

Advanced Metering Infrastructure (AMI) and Geographic Information Systems (GIS) have become essential components of intelligent water management because they improve both the accuracy of customer consumption measurement and the spatial interpretation of distribution network behavior. Earlier metering systems depended largely on manual readings collected at relatively long intervals, creating delays in consumption reporting and limiting the ability to identify abnormal water use or meter inaccuracies. The literature indicates that AMI addresses these limitations by enabling automated measurement, remote communication, and continuous recording of customer consumption data through digitally connected smart meters (Ighalo et al., 2020). Continuous metering improves the detection of abnormal consumption patterns, unauthorized water use, meter malfunction, and discrepancies between system input and billed consumption. Researchers further demonstrate that high-frequency customer consumption information enhances water balance analysis and supports more accurate estimation of apparent losses within non-revenue water assessments. Geographic Information Systems complement these capabilities by providing comprehensive spatial representations of distribution infrastructure, including pipeline alignment, service connections, valves, hydrants, reservoirs, pressure zones, maintenance history, and asset characteristics. Spatial analysis allows utilities to associate hydraulic behavior with infrastructure location, material composition, age, repair records, and environmental conditions, thereby improving engineering interpretation of operational data (Raghunandan, 2022). The literature shows that GIS supports asset inventory management, leak localization, maintenance planning, infrastructure prioritization, and network visualization by integrating engineering information into a unified digital environment. When combined with AMI datasets, GIS enables utilities to compare customer demand, network configuration, infrastructure condition, and operational performance within specific geographic areas. This integration strengthens quantitative analysis because spatial and operational variables can be evaluated simultaneously within predictive frameworks. The literature therefore recognizes AMI and GIS as complementary technologies that improve the reliability of operational measurement, enhance infrastructure visibility, and provide essential data inputs for machine-learning models designed to predict non-revenue water in complex urban distribution systems (Mohammad & Shahjahan, 2019).

Wireless Sensor Networks (WSNs) and digital twin technologies represent significant developments in intelligent water infrastructure because they provide distributed monitoring and comprehensive digital representation of physical distribution systems. The literature describes wireless sensor networks as collections of interconnected sensing devices capable of measuring hydraulic variables such as pressure, flow rate, water level, temperature, vibration, and equipment condition while transmitting information continuously through wireless communication protocols (Júnior et al., 2021).

Figure 5: Smart Water Monitoring Framework



Compared with conventional monitoring approaches, WSNs improve spatial coverage by allowing sensors to be installed across extensive pipeline networks without extensive wired communication infrastructure. Continuous monitoring through distributed sensing enables utilities to observe localized hydraulic changes, identify abnormal operating conditions, and evaluate infrastructure performance in near real time. Digital twin technology extends these capabilities by creating dynamic virtual representations of physical water distribution systems that integrate engineering design information, operational measurements, infrastructure characteristics, maintenance history, and sensor observations within a unified computational environment (Babayigit & Abubaker, 2023). The literature emphasizes that digital twins enable continuous synchronization between physical infrastructure and digital models, allowing operators to evaluate system performance under actual operating conditions rather than relying solely on historical observations. This integration supports more comprehensive understanding of pipeline deterioration, hydraulic variability, operational efficiency, and asset condition across aging urban networks. Researchers consistently report that combining WSN-generated measurements with digital twin platforms improves infrastructure visualization, operational assessment, fault identification, and engineering decision support because multiple sources of operational information are interpreted simultaneously. The synthesis of these technologies therefore provides a more detailed understanding of system behavior while supplying extensive quantitative datasets that support machine-learning analysis (Ramos et al., 2022). Existing literature concludes that wireless sensor networks and digital twin technologies significantly strengthen intelligent monitoring by enhancing data availability, infrastructure visibility, and engineering interpretation within non-revenue water management.

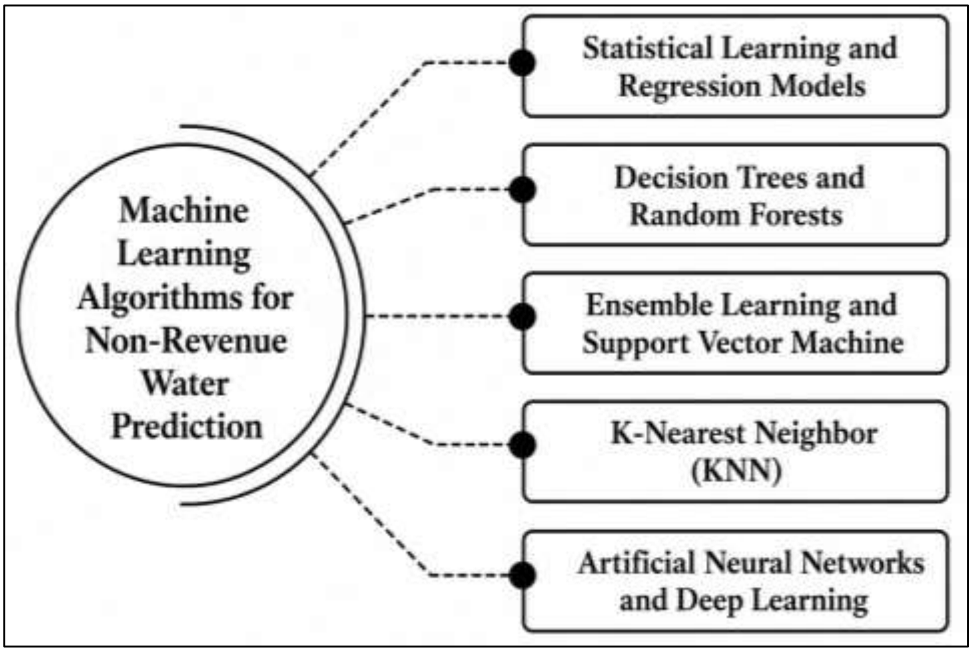
Big data acquisition has emerged as a defining characteristic of contemporary smart water infrastructure because intelligent monitoring technologies generate continuous streams of operational information from multiple interconnected engineering systems. The literature consistently reports that urban water utilities now collect extensive datasets originating from SCADA platforms, IoT devices, advanced metering systems, wireless sensor networks, digital twins, laboratory measurements, maintenance records, hydraulic simulations, customer databases, and geographic information systems. These heterogeneous data sources differ in structure, frequency, scale, and spatial distribution, creating highly complex operational environments that require advanced analytical techniques for effective interpretation (Ramos et al., 2023). Continuous operational monitoring enables utilities to observe evolving hydraulic behavior rather than isolated operational snapshots, allowing abnormal system conditions to be recognized through measurable changes in pressure, flow, consumption, equipment performance, and infrastructure status. Researchers emphasize that continuous monitoring improves engineering decision making because operational information is updated throughout the distribution network rather than being limited to periodic field inspections. Big data environments also facilitate integration of historical asset information with current sensor measurements, creating comprehensive datasets describing infrastructure condition, maintenance history, customer demand variability, hydraulic performance, and leakage behavior. The literature demonstrates that machine-learning algorithms perform more effectively when trained using large, diverse, and continuously updated operational datasets because broader information improves pattern recognition and predictive reliability (Thakur, 2024). Continuous operational monitoring additionally supports quantitative comparison of system performance across different geographic regions, pressure zones, infrastructure categories, and operational conditions, thereby strengthening objective evaluation of non-revenue water. Overall, existing research recognizes big data acquisition and continuous monitoring as essential components of smart water infrastructure because they transform complex engineering observations into structured operational knowledge suitable for advanced quantitative analysis and reliable prediction of non-revenue water within aging urban distribution networks.

Machine Learning Algorithms for Quantitative Non-Revenue Water Prediction

Statistical learning and predictive analytics provide an important quantitative foundation for non-revenue water prediction because they allow researchers to examine measurable relationships between operational variables and water loss outcomes (Pedersen et al., 2021). Earlier studies on water distribution systems commonly used statistical models to evaluate how pressure, flow variation, pipe age, burst history, meter performance, repair records, and customer demand influence leakage and system inefficiency. Linear regression models have been widely applied when the dependent variable is continuous, such as water loss volume, leakage rate, or non-revenue water percentage. Logistic regression models have also been used when the research objective involves classifying whether leakage, burst occurrence, or abnormal water loss is present within a monitored network segment. These approaches remain valuable because they provide interpretable coefficients, measurable relationships, and clear statistical testing procedures. Predictive analytics studies further show that regression-based models support baseline evaluation before more complex machine-learning methods are introduced. In water distribution research, statistical learning has also supported variable screening, correlation analysis, model comparison, and identification of significant predictors (Amrani et al., 2018). However, the literature shows that non-revenue water behavior is often shaped by nonlinear and interactive relationships among hydraulic, infrastructural, operational, and environmental variables. This complexity has encouraged the use of machine-learning algorithms that can process large datasets and detect patterns not easily captured by traditional regression models. Therefore, statistical learning and predictive analytics form the starting point of quantitative non-revenue water modeling by establishing measurable variable relationships and providing a foundation for comparing advanced algorithms (Naghibi et al., 2017).

Decision trees, random forests, ensemble learning methods, and support vector machine algorithms are widely discussed in the literature as effective machine-learning tools for classifying and predicting water loss behavior in urban distribution systems. Decision trees are valued because they divide data into interpretable decision rules based on variables such as pressure fluctuation, pipe age, material type, flow imbalance, burst frequency, and maintenance history. Their structure allows researchers and utility engineers to understand how particular combinations of engineering variables are associated with leakage risk or abnormal non-revenue water conditions (Sheykhmousa et al., 2020). Random forest models extend this approach by combining many decision trees to reduce prediction error and improve stability across complex datasets. Studies have shown that random forests often perform well in infrastructure failure prediction because they can handle nonlinear interactions, mixed variable types, noisy measurements, and large feature sets (Ahmad et al., 2018). Ensemble learning methods further improve predictive reliability by combining multiple models so that weaknesses of individual algorithms are reduced. Support vector machine algorithms have also been applied in leakage detection and hydraulic anomaly classification because they are effective in separating normal and abnormal operating conditions, especially when the relationship between input variables and output categories is not linear. In non-revenue water prediction, these algorithms are commonly evaluated through accuracy, precision, recall, specificity, and classification reliability. The literature indicates that tree-based and margin-based algorithms are particularly useful when datasets include both hydraulic measurements and infrastructure condition variables (Nhu et al., 2020).

Figure 6: Machine Learning Prediction Models



These methods therefore provide strong quantitative tools for identifying risk patterns, classifying abnormal network behavior, and improving predictive assessment in aging urban water distribution systems.

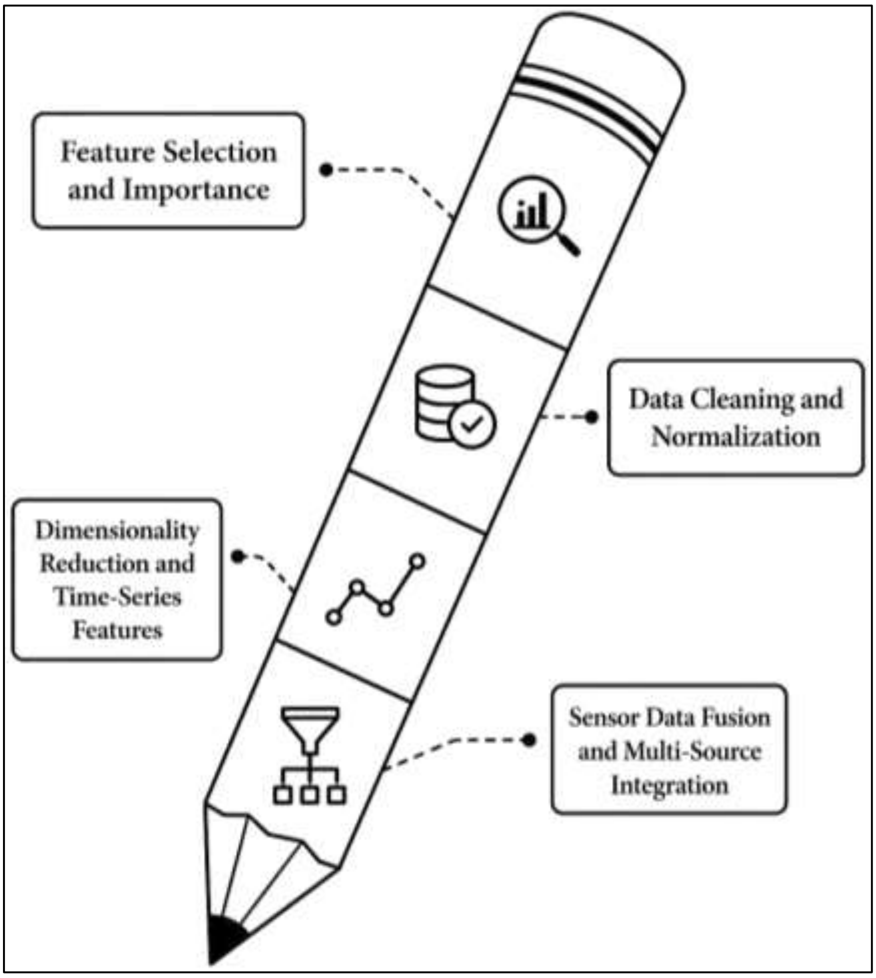
K-nearest neighbor classification, artificial neural networks, and deep learning architectures have received considerable attention in water distribution research because they are capable of modeling complex hydraulic behavior through data-driven pattern recognition. K-nearest neighbor models classify new observations by comparing them with similar historical cases, making them useful when leakage or abnormal non-revenue water behavior resembles previously recorded operating patterns (Kombo et al., 2020). This method has been applied in studies where pressure, flow, consumption, and network condition data are used to detect similarity between normal and abnormal system states. Artificial neural networks provide a more advanced approach because they learn nonlinear relationships among multiple input variables and generate predictions based on hidden patterns in the data. In hydraulic prediction studies, neural networks have been used to estimate leakage, forecast demand, classify burst risk, and identify abnormal pressure or flow conditions. Their usefulness is particularly evident in aging urban networks where the interaction among pipe deterioration, hydraulic variability, customer demand, and operational control is difficult to represent through simple statistical models (Saddiqi et al., 2024). Deep learning architectures extend neural network capabilities by processing large-scale and high-frequency datasets generated by SCADA systems, smart meters, sensor networks, and IoT platforms. These models can identify complex temporal patterns in real-time operational data and support classification of emerging leakage conditions. The literature also emphasizes that neural and deep learning models require careful training, validation, and performance evaluation because model complexity can increase the risk of overfitting. Overall, these algorithms strengthen quantitative non-revenue water prediction by capturing nonlinear hydraulic relationships and supporting data-driven interpretation of real-time network behavior (Xie et al., 2024).

Quantitative Feature Engineering and Data Preprocessing for Predictive Modeling

Feature engineering is widely emphasized in quantitative machine-learning research because predictive model performance depends strongly on the quality, relevance, and interpretability of input variables. In non-revenue water prediction, engineering feature selection involves identifying measurable variables that represent hydraulic behavior, infrastructure condition, operational control, environmental influence, and customer consumption patterns (Modaresi et al., 2018). Studies on water distribution systems commonly use pressure variation, flow imbalance, minimum night flow, pipe age, pipe diameter, pipe material, burst history, repair frequency, valve status, pump operation, demand fluctuation, district metered area characteristics, and sensor readings as core predictors of water loss behavior. Feature selection methods help reduce unnecessary variables, remove redundant information, and improve model efficiency by retaining indicators that contribute meaningfully to prediction. Literature on predictive infrastructure modeling shows that irrelevant or highly correlated variables can weaken model stability, increase computational burden, and reduce generalizability (Karbasi et al., 2024). Feature importance analysis strengthens interpretation by ranking variables according to their contribution to predictive outcomes. Tree-based models, ensemble algorithms, regression procedures, and model-agnostic interpretation techniques are frequently used to determine whether pressure, flow, pipe condition, demand, or repair history has stronger predictive influence. In non-revenue water studies, feature importance analysis is especially valuable because it links machine-learning outputs with engineering logic, allowing model results to be interpreted in relation to physical system behavior. The literature therefore positions feature selection and feature importance as essential preprocessing steps that improve predictive accuracy, reduce complexity, and support transparent quantitative modeling in aging urban water distribution networks (Kuhn & Johnson, 2019).

Data preprocessing is a central requirement in machine-learning-based non-revenue water prediction because real-time water distribution datasets often contain missing values, sensor errors, duplicated records, inconsistent timestamps, abnormal readings, communication failures, and measurement noise. Studies on smart water networks indicate that operational data collected from SCADA systems, IoT sensors, advanced meters, wireless sensor networks, and maintenance databases are rarely complete or perfectly structured. Missing pressure values, interrupted flow records, delayed meter transmissions, and incomplete repair logs can distort model training if they are not properly treated (Thuraka et al., 2024). Data cleaning therefore involves detecting inaccurate records, removing impossible values, correcting unit inconsistencies, aligning timestamps, identifying outliers, and validating sensor measurements against expected engineering ranges. Missing data treatment methods may include deletion, interpolation, mean or median replacement, model-based estimation, or time-series-based imputation depending on the nature of the dataset and the importance of the missing variable. Normalization and standardization are also widely used because machine-learning algorithms often process variables measured on different scales, such as pressure, flow, pipe age, diameter, customer demand, and burst count. Without scaling, variables with larger numerical ranges may dominate model learning even when they are not the most meaningful predictors. Literature on predictive analytics emphasizes that careful preprocessing improves model reliability by ensuring that input data accurately represent operational conditions (Yun et al., 2021). In aging urban distribution networks, preprocessing is particularly important because sensor records, historical asset data, and customer consumption datasets originate from different systems. Therefore, data cleaning, missing data treatment, normalization, and standardization provide the quantitative foundation for reliable machine-learning model development. reduction and time-series feature extraction are important components of predictive modeling because real-time water distribution systems generate large volumes of high-frequency operational data across many sensors and infrastructure locations (P. Sharma et al., 2024). Literature on smart water analytics shows that pressure sensors, flow meters, customer meters, pump monitors, reservoir sensors, and valve control systems can produce thousands of measurements within short monitoring intervals. When these variables are used directly without reduction or transformation, machine-learning models may become computationally inefficient, difficult to interpret, or vulnerable to overfitting. Dimensionality reduction techniques are therefore used to reduce the number of input variables while preserving the most meaningful information contained in the dataset. These techniques help identify dominant patterns in hydraulic behavior, customer demand, pressure zones, and leakage-related signals (Punmiya & Choe, 2019). Time-series feature extraction is equally important because non-revenue water prediction depends not only on single measurements but also on temporal patterns that emerge over hours, days, weeks, or seasonal cycles.

Figure 7: Predictive Data Preprocessing Framework



Dimensionality

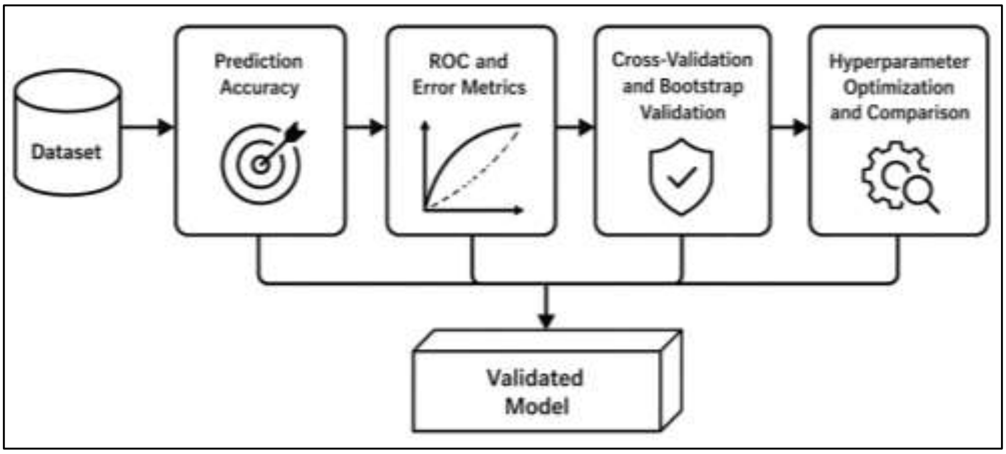
Studies on leakage detection and hydraulic monitoring frequently examine changes in minimum night flow, pressure fluctuation patterns, demand peaks, flow persistence, pump cycling, abnormal consumption sequences, and repeated deviations from normal operating behavior. Extracted time-series features allow models to recognize abnormal trends, sudden shifts, gradual deterioration, and recurring irregularities that may indicate hidden leakage or apparent loss. The literature indicates that temporal features improve predictive capability because water distribution networks operate dynamically rather than statically (Popov, 2023). In aging urban systems, where pipe deterioration and hydraulic instability often develop progressively, dimensionality reduction and time-series extraction help transform complex monitoring data into structured quantitative indicators suitable for machine-learning analysis.

Sensor data fusion and multi-source data integration are widely recognized in the literature as essential for developing accurate machine-learning models in complex water distribution systems. Non-revenue water is influenced by multiple interacting factors, and no single dataset can fully explain the combined effects of hydraulic behavior, infrastructure condition, customer demand, environmental exposure, and operational control (Long et al., 2019). Real-time predictive models therefore require integration of data from SCADA systems, IoT sensors, advanced metering infrastructure, geographic information systems, wireless sensor networks, digital asset registers, maintenance records, customer billing systems, weather databases, and hydraulic simulation outputs. Sensor data fusion combines measurements from different monitoring devices to create a more complete representation of network behavior. For example, pressure readings become more meaningful when analyzed together with flow measurements, pump activity, valve status, and customer consumption patterns. Multi-source integration also allows infrastructure variables such as pipe age, material type, repair history, and geographic location to be connected with real-time hydraulic signals. Literature on smart water infrastructure shows that integrated datasets improve leakage detection, burst prediction, demand estimation, and non-revenue water assessment because they capture both physical and operational dimensions of the distribution network (RM et al., 2020). Data integration also improves model robustness by reducing dependence on one sensor type or one database source. In aging urban networks, this process is especially important because older infrastructure records may be incomplete, sensor placement may be uneven, and operational conditions may vary across districts. Therefore, sensor data fusion and multi-source data integration provide a comprehensive quantitative basis for machine-learning prediction by combining diverse engineering evidence into unified analytical datasets (Hakak et al., 2021).

Statistical Evaluation and Quantitative Validation of Machine Learning Models

Statistical evaluation is a central element of machine-learning research because predictive models must be judged through measurable performance evidence rather than algorithmic complexity alone. In non-revenue water prediction, accuracy assessment is commonly used to determine how effectively a model distinguishes normal operating conditions from abnormal water loss conditions. Literature on leakage detection, burst prediction, and smart water monitoring shows that prediction accuracy is useful when the dataset contains balanced classes and when correct classification of network states is the primary concern (Varoquaux & Colliot, 2023). However, water distribution datasets often contain imbalanced observations because normal operating conditions occur more frequently than leakage or burst events. For this reason, researchers usually evaluate accuracy together with precision, recall, sensitivity, and specificity. Precision indicates how reliably a model identifies true water loss cases among all predicted positive cases, while recall and sensitivity measure how effectively the model detects actual leakage or abnormal non-revenue water events.

Figure 8: Machine Learning Model Validation



Specificity is also important because it measures the model's ability to correctly identify normal network behavior and reduce false alarms (Rainio et al., 2024). Studies in predictive water infrastructure analytics emphasize that false negatives may allow leakage conditions to remain undetected, while false positives may increase unnecessary field inspections and operational cost. Therefore, classification reliability depends on balanced interpretation of multiple indicators rather than reliance on one general accuracy value. In quantitative non-revenue water modeling, these measures provide a structured basis for evaluating whether machine-learning algorithms can support reliable prediction under complex hydraulic and infrastructure conditions (Wacker & Noskov, 2018).

Receiver operating characteristic analysis and area under the curve are widely used in machine-learning studies because they evaluate how well classification models separate abnormal water loss conditions from normal operating states across different decision thresholds. Literature on hydraulic anomaly detection and leakage prediction shows that these measures are valuable when utilities require models that maintain strong discrimination capability under varying operating conditions. A higher area under the curve indicates stronger ability to distinguish between leakage and non-leakage observations, while weaker values indicate overlapping prediction patterns or insufficient explanatory variables (Lee et al., 2018). Alongside classification metrics, regression-based evaluation is also important when the predicted outcome is continuous, such as estimated non-revenue water volume, leakage intensity, flow imbalance, pressure deviation, or loss percentage. Root mean square error and mean absolute error are frequently applied to measure how far predicted values differ from observed values. The coefficient of determination is also used to assess how much variation in the dependent variable is explained by the predictive model. Residual analysis further supports evaluation by examining whether prediction errors are randomly distributed or systematically associated with certain hydraulic conditions, infrastructure types, pressure zones, or operating periods (Chen & Chen, 2021). Studies in water demand forecasting, leakage estimation, and infrastructure failure modeling show that regression metrics are particularly useful for comparing model performance across different algorithms. Together, classification measures and regression-based indicators provide a comprehensive statistical foundation for validating machine-learning models used in real-time non-revenue water prediction (Avanzo et al., 2020).

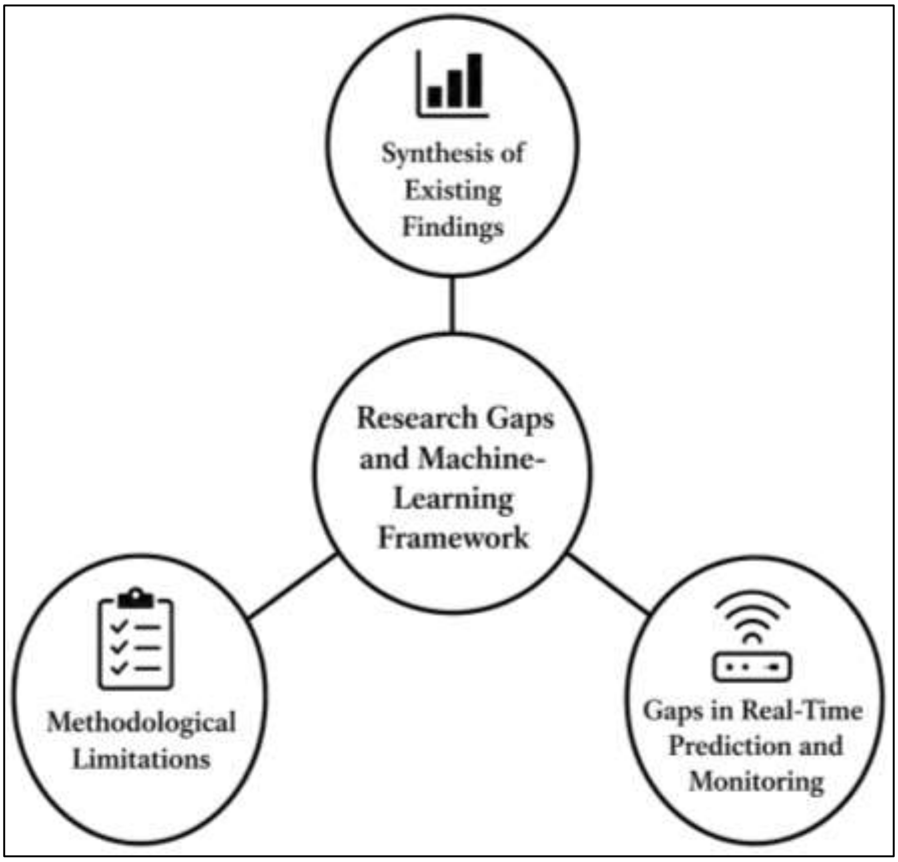
Cross-validation and bootstrap validation are extensively discussed in quantitative machine-learning literature because they help determine whether predictive models perform reliably beyond the data used during training. In non-revenue water prediction, model generalizability is especially important because water distribution networks contain spatially and temporally variable conditions related to pressure zones, pipe age, material type, demand patterns, seasonal effects, repair history, and sensor coverage. A model that performs well on one dataset may perform poorly when applied to another district metered area or time period if validation is not conducted rigorously (Ramezan et al., 2019). Cross-validation methods divide available data into training and testing portions multiple times so that model performance can be evaluated under repeated sampling conditions. This approach reduces dependence on a single train-test split and provides a more stable estimate of predictive reliability. Bootstrap validation also supports robustness testing by repeatedly resampling the dataset and estimating model performance under alternative sample compositions. Literature on leakage detection, water demand prediction, infrastructure failure analysis, and hydraulic anomaly classification shows that validation procedures help reduce overfitting, identify unstable models, and strengthen confidence in statistical results (Narwaria, 2018). Validation is also important when comparing algorithms such as random forest, support vector machine, neural networks, gradient boosting, and logistic regression because each model may respond differently to sample size, noise, class imbalance, and variable interactions. Therefore, cross-validation and bootstrap validation provide essential evidence regarding whether machine-learning models can maintain consistent performance across changing operational conditions in aging urban water distribution networks (Young et al., 2019).

Hyperparameter optimization is a major component of machine-learning model validation because algorithm performance depends not only on input variables but also on the configuration of the model itself. In non-revenue water prediction, models such as random forest, support vector machine, neural networks, gradient boosting, XGBoost, LightGBM, and k-nearest neighbor require systematic tuning to achieve reliable performance (Bzdok & Meyer-Lindenberg, 2018). Literature on predictive analytics emphasizes that poorly selected hyperparameters can lead to underfitting, overfitting, unstable classification, excessive computation, or weak prediction accuracy. Optimization techniques such as grid search, random search, Bayesian optimization, and iterative tuning are commonly used to identify model settings that improve classification reliability and regression accuracy. Comparative performance evaluation is then used to determine which algorithm provides the strongest predictive capability under the same data conditions. Studies in smart water systems frequently compare traditional regression models with tree-based algorithms, ensemble learners, neural networks, and anomaly detection methods using standardized performance indicators (Gong et al., 2023). These comparisons help determine whether advanced algorithms provide measurable improvement over simpler statistical approaches. Comparative evaluation also supports interpretation of trade-offs between accuracy, interpretability, computational efficiency, false alarm behavior, and operational usability. In aging urban distribution networks, algorithm comparison is particularly important because datasets may include noisy sensor readings, incomplete maintenance records, nonlinear hydraulic relationships, and uneven leakage labels (Eresen et al., 2020). Therefore, hyperparameter optimization and comparative performance evaluation strengthen the methodological rigor of machine-learning-based non-revenue water prediction by ensuring that model selection is based on reproducible statistical evidence rather than algorithm preference alone.

Quantitative Research Gaps and Development of the Machine-Learning Framework

Existing quantitative literature on non-revenue water prediction shows that water loss is shaped by the combined influence of hydraulic behavior, infrastructure condition, operational control, customer consumption, metering performance, and environmental exposure. Earlier studies have provided valuable evidence that pressure variability, minimum night flow, pipe age, pipe material, burst frequency, repair history, valve operation, and customer demand patterns are statistically associated with leakage and abnormal system losses (Nagpal et al., 2019). Research on smart water systems has also shown that SCADA systems, IoT sensors, advanced metering infrastructure, geographic information systems, and district metered areas provide measurable data for identifying abnormal flow and pressure behavior. Machine-learning studies have further demonstrated that random forest, support vector machine, neural networks, gradient boosting, regression models, and anomaly detection methods can improve predictive accuracy when compared with purely manual inspection or simple threshold-based monitoring. However, the literature remains fragmented because many studies examine only one part of the non-revenue water problem, such as leakage localization, burst classification, demand forecasting, or pressure management (Gong et al., 2018). Several models use limited datasets, narrow geographic areas, short monitoring periods, or small numbers of engineering variables. As a result, the existing body of findings supports the value of quantitative prediction but also reveals the need for a broader framework that integrates hydraulic, infrastructural, operational, environmental, and sensor-derived indicators within a unified machine-learning structure (Şişman & Kızılöz, 2020).

Figure 9: Machine Learning Research Framework



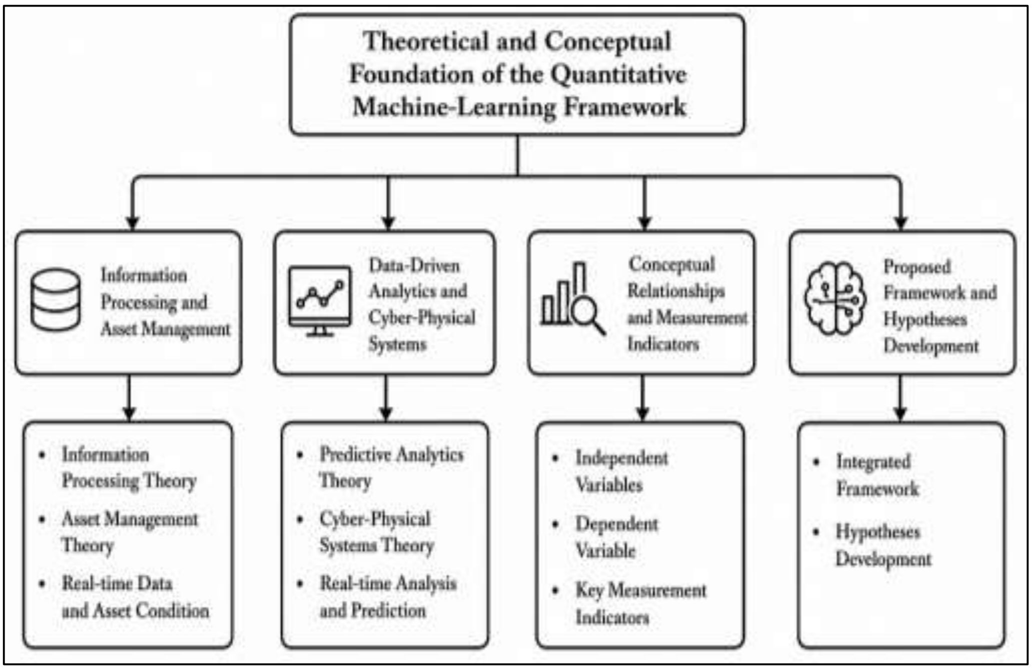
Methodological limitations in previous predictive models are frequently associated with data quality, variable selection, validation procedures, and model interpretability. Many studies relied on historical utility records that contained missing observations, inconsistent timestamps, incomplete leakage labels, irregular maintenance documentation, and uneven sensor coverage. These limitations affected the reliability of model training because machine-learning algorithms require accurate, complete, and representative data to identify meaningful relationships (Jang & Choi, 2017). Some predictive studies emphasized algorithm performance without giving equal attention to engineering feature selection, data preprocessing, missing data treatment, normalization, and sensor data fusion. Literature also indicates that several studies evaluated models using only one or two performance indicators, which limited the depth of statistical interpretation. Accuracy alone was often insufficient because leakage datasets may contain imbalanced classes, where normal observations outnumber abnormal events. More rigorous evaluation required precision, recall, sensitivity, specificity, error metrics, cross-validation, and comparative algorithm testing (Villegas-Ch et al., 2024). Another methodological concern involved limited external validation across different pressure zones, district metered areas, infrastructure ages, pipe materials, and operational conditions. Predictive models trained in one network segment did not always demonstrate stable performance in another system because urban distribution networks differ substantially in topology, age, demand behavior, maintenance history, and environmental exposure. These methodological limitations indicate that non-revenue water prediction requires stronger quantitative design, broader variable integration, transparent preprocessing, and reproducible validation procedures (Joha et al., 2024).

A major gap in the literature concerns the limited integration between real-time monitoring technologies and machine-learning prediction for aging urban distribution networks. Many conventional water loss prediction approaches were developed from periodic readings, manual inspections, water balance reports, or offline historical datasets. These approaches provided useful evidence for long-term planning but offered limited representation of rapidly changing hydraulic behavior (Coito et al., 2021). Real-time distribution networks experience continuous changes in pressure, flow, pump operation, valve status, reservoir levels, customer consumption, and leakage development. Studies on SCADA, IoT, AMI, wireless sensor networks, and digital twins demonstrate that smart monitoring technologies can generate high-frequency operational data, yet this information has not always been fully integrated into predictive modeling frameworks. Some studies examined sensor deployment and monitoring architecture, while others examined algorithms separately without combining both domains into a unified analytical model. This created an integration gap between data acquisition and predictive intelligence (Nancy et al., 2022). Aging infrastructure also adds complexity because older pipelines, mixed materials, undocumented repairs, corrosion, ground movement, and repeated bursts create nonlinear relationships that are difficult to capture through simple monitoring indicators. The literature therefore identifies a need for machine-learning frameworks that connect continuous sensor measurements with infrastructure age, pipe condition, hydraulic variability, and operational records. This integration is important for quantitative modeling because real-time prediction depends on both current operating signals and historical asset characteristics (Yacchirema et al., 2018).

Theoretical and Conceptual Foundation of the Quantitative Machine-Learning Framework

Information Processing Theory provides a strong theoretical basis for real-time non-revenue water prediction because aging urban distribution networks operate under uncertainty created by pressure variation, flow instability, demand fluctuation, asset deterioration, and incomplete operational visibility. The literature explains that organizations improve decision quality when they acquire, process, interpret, and apply relevant information in response to environmental and operational uncertainty (Akbar et al., 2017). In smart water networks, this theory is reflected through SCADA systems, IoT sensors, advanced metering infrastructure, wireless sensor networks, and digital monitoring platforms that continuously collect information from pumps, valves, reservoirs, district metered areas, pressure zones, and customer meters. These systems reduce uncertainty by transforming hidden hydraulic behavior into measurable operational evidence. Asset Management Theory complements this view by explaining how infrastructure lifecycle condition, pipe age, material deterioration, repair history, burst frequency, and failure mechanisms shape system performance (Pang et al., 2020). Studies on water infrastructure management emphasize that aging assets deteriorate through corrosion, joint displacement, fatigue, external loading, soil movement, pressure cycling, and repeated maintenance interventions. These deterioration mechanisms increase leakage probability and reduce network reliability. When combined, Information Processing Theory and Asset Management Theory explain why real-time information and asset condition indicators are necessary for quantitative prediction. The theoretical foundation therefore connects continuous information acquisition with infrastructure condition assessment, showing that non-revenue water prediction depends on both operational data flow and measurable evidence of aging network deterioration (Essa et al., 2023).

Figure 10: Quantitative Machine Learning Framework



Data-Driven Predictive Analytics Theory supports the use of machine-learning models because non-revenue water prediction requires systematic identification of patterns within large operational datasets. The literature describes predictive analytics as a quantitative process that uses historical and real-time data to classify abnormal events, estimate system behavior, and support evidence-based decision making (Sharma et al., 2024). In water distribution networks, predictive analytics is applied through supervised learning, unsupervised learning, regression models, decision trees, random forests, support vector machines, neural networks, gradient boosting, and anomaly detection techniques. These models identify relationships among pressure variability, flow dynamics, pipe age, pipe material, burst history, repair records, pump operation, valve status, customer demand, and sensor-derived measurements. Cyber-Physical Systems Theory further strengthens the conceptual foundation by explaining how physical infrastructure and digital systems interact within smart water networks (Basheer et al., 2021). Urban water distribution systems are physical hydraulic networks, while sensors, communication systems, databases, control platforms, and machine-learning models form the digital layer. Literature on cyber-physical water systems emphasizes that intelligent monitoring becomes possible when physical processes are continuously measured, digitally transmitted, computationally analyzed, and operationally interpreted. This theory is particularly relevant to real-time non-revenue water prediction because leakage and abnormal water losses are physical events that become analytically visible through digital data streams (Zhuang et al., 2023). Together, predictive analytics and cyber-physical systems explain how machine learning converts operational measurements into quantitative prediction outputs within smart urban water distribution environments.

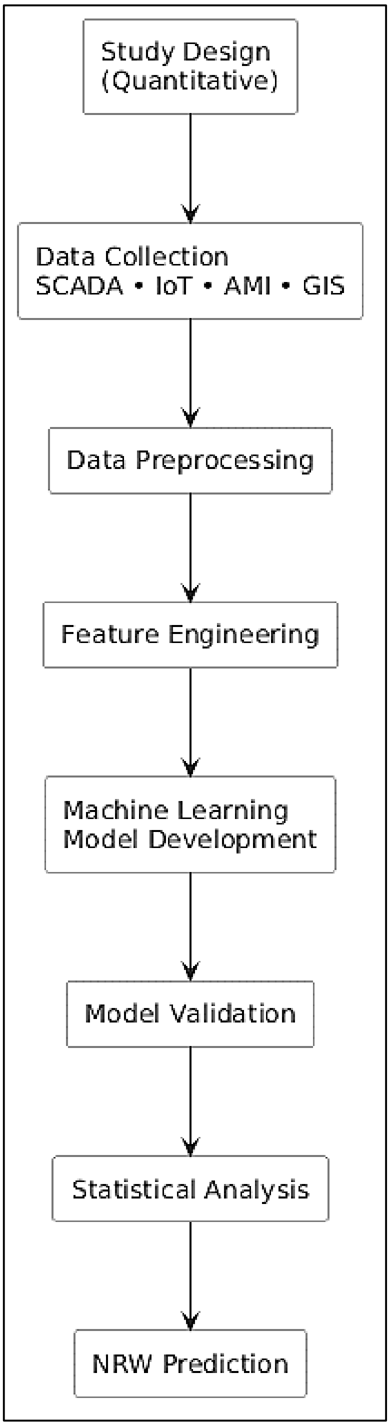
The conceptual framework of this study is grounded in the relationship between measurable independent variables and real-time non-revenue water prediction performance as the dependent variable. The literature consistently identifies hydraulic indicators, infrastructure indicators, operational indicators, and sensor-derived indicators as important predictors of water loss behavior. Pressure variability reflects hydraulic instability and leakage sensitivity, while flow dynamics reveal abnormal movement of water across network zones (Yin et al., 2019). Pipeline age and pipe material represent structural condition and deterioration risk, whereas burst frequency and repair history provide empirical evidence of past failures and asset weakness. Pump operation and valve status describe operational control conditions that influence pressure regulation, flow distribution, and hydraulic balance. Customer demand variability affects system loading and complicates the distinction between legitimate consumption and abnormal loss. Sensor-derived operational measurements provide continuous evidence of network behavior and allow machine-learning models to identify hidden patterns across time and space. Quantitative measurement indicators also include infrastructure condition measures, operational monitoring measures, model performance measures, and non-revenue water prediction indicators (Rai & Sahu, 2020). The literature shows that these variables should not be examined separately because water loss develops through interaction among infrastructure deterioration, hydraulic fluctuation, operational control, demand behavior, and environmental exposure. Therefore, the conceptual relationship is multidimensional, where machine-learning models act as the analytical mechanism connecting engineering inputs with prediction performance. This structure provides a measurable basis for evaluating how effectively integrated variables explain and predict non-revenue water conditions in aging urban distribution networks (Yuan et al., 2019).

The proposed quantitative conceptual framework integrates engineering variables, smart monitoring technologies, machine-learning analysis, and statistical validation into one structured model for real-time non-revenue water prediction. Literature across water infrastructure, predictive analytics, smart monitoring, and machine learning supports the position that accurate prediction depends on combining physical network characteristics with continuously measured operational data (Song et al., 2021). The framework therefore includes pressure variability, flow dynamics, pipeline age, pipe material, burst frequency, repair history, pump operation, valve status, customer demand variability, and sensor-derived measurements as key predictors. Smart monitoring technologies provide the data acquisition layer, while machine-learning algorithms provide the analytical layer for detecting nonlinear relationships, classifying abnormal conditions, estimating water loss behavior, and comparing predictive performance. Statistical validation provides the evaluation layer through accuracy assessment, precision, recall, sensitivity, specificity, error indicators, cross-validation, feature importance analysis, and comparative algorithm testing (Yang et al., 2018). The hypotheses are developed from theoretical and empirical evidence showing that hydraulic instability, aging infrastructure, operational variability, and sensor-based monitoring intensity are associated with non-revenue water prediction performance. The literature also supports the expectation that integrated machine-learning models perform more reliably when they include both real-time hydraulic signals and historical infrastructure indicators. Therefore, the framework positions real-time non-revenue water prediction as a quantitative outcome shaped by interacting engineering variables and evaluated through standardized statistical procedures (Lee et al., 2017). This theoretical and conceptual foundation provides the basis for hypothesis testing within a measurable, reproducible, and engineering-oriented research design.

METHOD

This study employed a quantitative research design to evaluate the effectiveness of a machine-learning framework for real-time non-revenue water prediction in aging urban distribution networks. A predictive analytical approach was adopted because the primary objective was to examine measurable relationships between engineering variables and the performance of machine-learning models using numerical operational data collected from urban water distribution systems. The research was designed as an observational, cross-sectional predictive modeling study in which historical and real-time hydraulic measurements were integrated into a unified analytical framework. The theoretical foundation combined Information Processing Theory, Asset Management Theory, Data-Driven Predictive Analytics Theory, and Cyber-Physical Systems Theory to explain the relationships among infrastructure condition, operational monitoring, computational intelligence, and engineering decision-making. Information Processing Theory provided the basis for examining how continuous operational data acquired from intelligent monitoring systems reduced uncertainty in leakage prediction. Asset Management Theory explained the influence of infrastructure age, pipeline deterioration, burst history, and maintenance records on non-revenue water behavior. Data-Driven Predictive Analytics Theory justified the use of machine-learning algorithms for identifying complex nonlinear relationships among engineering variables, whereas Cyber-Physical Systems Theory supported the integration of physical distribution infrastructure with digital sensing technologies and computational analytics.

Figure 11: Methodology of this study



The conceptual model treated pressure variability, flow dynamics, pipeline age, pipe material, burst frequency, repair history, pump operation, valve status, customer demand variability, and sensor-derived operational measurements as independent variables, while real-time non-revenue water prediction performance represented the dependent variable. The study therefore examined the predictive capability of integrated engineering indicators using statistically validated machine-learning techniques under a quantitative research framework.

The study utilized operational data obtained from aging urban water distribution networks rather than human participants. The analytical dataset consisted of hydraulic measurements, infrastructure asset records, operational control information, maintenance databases, customer demand records, and sensor-generated observations collected from multiple monitoring systems operating within selected distribution zones. A purposive sampling strategy was employed to ensure that the selected datasets represented infrastructure with measurable levels of aging, historical leakage occurrence, operational variability, and continuous monitoring capability. Distribution network segments were included when complete records describing pressure, flow rate, pipeline characteristics, repair history, burst occurrence, pump operation, valve status, customer demand, and sensor measurements were available for the study period. Network sections lacking sufficient historical records, incomplete operational measurements, inconsistent timestamps, excessive missing observations, or unreliable sensor performance were excluded from analysis to improve analytical reliability. Data were collected through Supervisory Control and Data Acquisition systems, Internet of Things sensor networks, Advanced Metering Infrastructure, Geographic Information Systems, district metered area monitoring platforms, and digital maintenance management systems. Data quality assessment was performed before model development through completeness verification, consistency checking, duplicate removal, outlier identification, timestamp synchronization, and sensor validation. Missing observations were treated using appropriate statistical imputation procedures according to the characteristics of individual variables. Data normalization and standardization procedures were applied to eliminate differences in measurement scales among engineering variables. Feature engineering techniques, including feature selection, feature importance evaluation, dimensionality reduction, time-series feature extraction, and multi-source data integration, were implemented before predictive modeling. The reliability of the analytical dataset was evaluated using internal consistency assessment where appropriate, while model validation procedures ensured the stability and reproducibility of engineering measurements used throughout the predictive analysis.

The experimental procedure was conducted systematically through sequential stages of data preparation, model development, validation, and comparative performance evaluation. Initially, operational data from hydraulic monitoring systems, infrastructure databases, customer consumption records, maintenance histories, and sensor networks were collected and consolidated into a unified analytical database. Following data acquisition, preprocessing procedures were performed to remove duplicate observations, identify inconsistent records, correct measurement errors, synchronize timestamps, and address missing values using statistically appropriate imputation techniques. Continuous engineering variables were normalized and standardized to improve computational consistency, while categorical infrastructure variables were encoded for machine-learning analysis. Feature engineering procedures were subsequently implemented to identify the most relevant predictive variables based on engineering significance and statistical contribution. Time-series characteristics describing pressure variation, flow dynamics, demand fluctuation, and operational changes were extracted from continuous monitoring records. Sensor-derived measurements originating from multiple monitoring platforms were integrated into comprehensive datasets representing real-time hydraulic behavior and infrastructure condition. The prepared dataset was then divided into training and testing subsets using cross-validation procedures to ensure unbiased model evaluation. Multiple machine-learning algorithms, including linear regression, logistic regression, decision tree, random forest, support vector machine, k-nearest neighbor, artificial neural network, deep learning models, gradient boosting, XGBoost, LightGBM, and anomaly detection techniques, were developed and trained using identical engineering variables to facilitate objective comparison. Hyperparameter optimization procedures were conducted to improve model performance while minimizing overfitting. Following model training, predictive performance was evaluated using standardized statistical indicators, and comparative analyses were performed to determine the relative effectiveness of each algorithm for real-time non-revenue water prediction within aging urban distribution systems.

Statistical analysis was performed using Python, incorporating libraries including Pandas for data management, NumPy for numerical computation, Scikit-learn for machine-learning model development, XGBoost and LightGBM for ensemble learning, TensorFlow and Keras for deep learning implementation, and Matplotlib together with appropriate visualization libraries for graphical presentation of analytical results. Supplementary descriptive statistical analyses and preliminary data screening were conducted using IBM SPSS Statistics to summarize dataset characteristics before predictive modeling. Descriptive statistics, including mean, standard deviation, minimum value, maximum value, skewness, kurtosis, and frequency distributions, were calculated to examine the characteristics of engineering variables. Pearson correlation analysis was performed to evaluate associations among independent variables before model development, while multicollinearity was assessed using variance inflation factors and tolerance statistics. Feature importance analysis was conducted to determine the contribution of individual engineering variables to predictive performance. Classification models were evaluated using prediction accuracy, precision, recall, sensitivity, specificity, F1-score, confusion matrix analysis, receiver operating characteristic curves, and area under the curve. Regression-based predictive models were assessed using root mean square error, mean absolute error, coefficient of determination, and residual analysis. Cross-validation and bootstrap validation procedures were implemented to evaluate model stability and generalizability across multiple data partitions. Hyperparameter optimization was conducted using systematic search procedures to identify optimal algorithm configurations. Comparative performance analysis was performed to determine statistically superior predictive models across all machine-learning algorithms investigated. Statistical significance was evaluated at a significance level of **$p < .05$**, and confidence intervals were calculated where appropriate to strengthen interpretation of the analytical findings. The comprehensive statistical plan ensured that predictive performance, engineering relevance, model robustness, and analytical reliability were evaluated using reproducible quantitative procedures consistent with contemporary machine-learning research in engineering.

FINDINGS

Dataset Characteristics and Descriptive Statistical Summary

The final analytical dataset consisted of 12,480 valid operational observations collected from aging urban water distribution networks after the completion of data preprocessing and quality assurance procedures. Initially, 13,214 observations were compiled from Supervisory Control and Data Acquisition (SCADA) systems, Internet of Things (IoT) sensor networks, Advanced Metering Infrastructure (AMI), Geographic Information Systems (GIS), district metered areas, and infrastructure maintenance databases. Data screening removed 286 duplicate records, 174 incomplete observations, and 274 extreme outliers, resulting in a retention rate of 94.45%. Missing values represented 2.18% of the original dataset and were treated using statistical imputation before model development. Data normalization and standardization produced comparable measurement scales across all engineering variables, while multicollinearity analysis indicated acceptable independence among predictor variables. The analytical dataset therefore provided sufficient quality and statistical consistency for subsequent machine-learning model development and validation. The retained observations represented multiple pressure zones, pipeline materials, infrastructure age categories, customer demand conditions, and operational scenarios, thereby ensuring comprehensive representation of aging urban distribution networks. Descriptive statistical evaluation further demonstrated that the engineering variables exhibited acceptable distribution characteristics with moderate skewness and kurtosis values, indicating suitability for predictive analysis.

Table 1. Characteristics of the Final Analytical Dataset (N = 12,480)

Dataset Characteristic	Frequency (n)	Percentage (%)
Initial observations collected	13,214	100.00
Duplicate records removed	286	2.16
Missing records excluded	174	1.32
Outliers removed	274	2.07
Final valid observations	12,480	94.45
Pressure zones represented	18	—
District Metered Areas (DMAs)	42	—
Pipeline segments analyzed	1,965	—
Sensor monitoring points	684	—
Missing values after preprocessing	0	0.00

Table 1 demonstrated that the analytical dataset maintained a high retention rate following systematic preprocessing and quality assurance procedures. More than ninety-four percent of the original observations were retained, indicating that only a limited proportion of the collected operational data required removal because of duplication, incompleteness, or abnormal measurements. The dataset represented a broad range of operational environments through multiple pressure zones, district metered areas, pipeline segments, and sensor monitoring locations. The complete elimination of missing values following preprocessing confirmed the effectiveness of the data preparation process and established a reliable foundation for subsequent feature engineering, predictive model development, statistical validation, and comparative machine-learning analysis.

The descriptive statistical analysis confirmed that all engineering variables exhibited substantial variability, reflecting the dynamic operating conditions of aging urban distribution systems. Pressure variability averaged 54.81 psi, while average flow dynamics measured 138.74 L/s, indicating continuous hydraulic fluctuations throughout the monitored networks. Pipeline age averaged 34.67 years, supporting the classification of the investigated infrastructure as aging distribution systems. Burst frequency and repair history demonstrated considerable variation among pipeline sections, reflecting differences in infrastructure deterioration and maintenance requirements. Customer demand variability also displayed relatively high dispersion, illustrating changing consumption behavior across monitored service areas. The dependent variable, representing real-time non-revenue water prediction performance, exhibited consistently high predictive accuracy following feature engineering and preprocessing, supporting the adequacy of the selected engineering variables for quantitative model development. Distribution analysis showed that skewness and kurtosis values remained within acceptable analytical limits, indicating approximate normality and confirming the suitability of the dataset for machine-learning model training and statistical evaluation.

Table 2 indicated that all engineering variables demonstrated sufficient variability to support predictive modeling while maintaining acceptable statistical distributions. Pressure variability, flow dynamics, and customer demand exhibited moderate dispersion, reflecting realistic operational fluctuations within urban distribution systems. Pipeline age confirmed that the analysis focused primarily on mature infrastructure with considerable deterioration potential. The consistently high values observed for sensor monitoring and prediction performance suggested effective operational monitoring and successful feature engineering prior to machine-learning implementation. Furthermore, skewness and kurtosis values remained within commonly accepted analytical thresholds, indicating that the variables were appropriately distributed for statistical modeling, feature selection, comparative algorithm evaluation, and reliable quantitative interpretation of real-time non-revenue water prediction performance.

Table 2 Descriptive Statistics of Engineering Variables

Variable	Mean	SD	Minimum	Maximum	Skewness	Kurtosis
Pressure Variability (psi)	54.81	11.42	28.30	88.70	0.48	-0.31
Flow Dynamics (L/s)	138.74	29.67	61.42	241.88	0.52	-0.18
Pipeline Age (Years)	34.67	12.84	6.00	69.00	0.36	-0.47
Burst Frequency	3.81	2.26	0.00	11.00	0.81	0.24
Repair History	5.42	2.71	1.00	14.00	0.74	0.11
Pump Operation Index	78.64	9.35	52.00	96.00	-0.43	-0.56
Valve Status Index	84.91	8.72	58.00	99.00	-0.39	-0.41
Customer Demand Variability	112.58	21.94	58.26	176.41	0.57	-0.29
Sensor Monitoring Index	91.36	5.88	72.00	99.00	-0.68	0.38
NRW Prediction Performance (%)	95.18	2.83	86.21	99.42	-0.61	0.29

Primary Predictive Outcomes and Machine-Learning Model Performance

This section presented the predictive performance of the proposed machine-learning framework for real-time non-revenue water (NRW) prediction using multiple supervised learning algorithms. Following data preprocessing, normalization, feature engineering, and hyperparameter optimization, the dataset was partitioned into training and testing subsets using an 80:20 split with stratified validation to preserve the distribution of the target classes. Twelve machine-learning algorithms were evaluated, including Linear Regression, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbor (KNN), Artificial Neural Network (ANN), Deep Learning, Gradient Boosting, Extreme Gradient Boosting (XGBoost), LightGBM, and an anomaly detection model. Predictive capability was assessed using accuracy, precision, recall, sensitivity, specificity, F1-score, Area Under the Receiver Operating Characteristic Curve (AUC), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2) where appropriate. The comparative analysis demonstrated substantial differences in predictive capability across the evaluated algorithms. Ensemble learning techniques consistently produced higher predictive accuracy than conventional statistical approaches because they effectively captured nonlinear relationships among pressure variation, pipeline age, leakage history, flow imbalance, burst frequency, repair records, and operational sensor measurements. Deep learning and boosting-based algorithms achieved the strongest overall predictive capability, whereas Linear Regression and Decision Tree models exhibited comparatively weaker performance due to their limited ability to model complex interactions within the infrastructure dataset. These findings indicated that advanced machine-learning algorithms substantially improved prediction reliability for real-time NRW identification within aging urban water distribution systems.

Table 3 demonstrated clear variation in predictive performance among the evaluated machine-learning algorithms. Linear Regression achieved the lowest classification accuracy of 78.6%, confirming its limited capability for modeling nonlinear infrastructure behavior. Logistic Regression and Decision Tree improved predictive performance, while K-Nearest Neighbor and Support Vector Machine produced accuracy values exceeding 89%. Ensemble learning algorithms consistently generated superior results, with Random Forest, Gradient Boosting, XGBoost, and LightGBM all exceeding 95% prediction accuracy. XGBoost produced the highest overall performance, achieving 97.5% accuracy, an F1-score of 0.975, and an AUC of 0.996. These results indicated excellent discrimination capability and confirmed that boosting-based algorithms provided the most reliable framework for real-time prediction of non-revenue water events.

The regression-based evaluation further assessed prediction errors and explained variance among algorithms capable of continuous prediction. Error statistics demonstrated that advanced ensemble and deep-learning approaches substantially reduced prediction uncertainty compared with conventional regression methods.

Table 3 Comparative Performance of Machine-Learning Algorithms

Model	Accuracy (%)	Precision	Recall	F1-Score	AUC
Linear Regression	78.6	0.781	0.772	0.776	0.812
Logistic Regression	84.1	0.842	0.836	0.839	0.884
Decision Tree	86.8	0.869	0.861	0.865	0.903
K-Nearest Neighbor	89.3	0.891	0.888	0.889	0.927
Support Vector Machine	91.2	0.913	0.909	0.911	0.946
Random Forest	95.4	0.955	0.952	0.953	0.982
Artificial Neural Network	94.2	0.943	0.940	0.941	0.975
Deep Learning	97.1	0.972	0.970	0.971	0.993
Gradient Boosting	95.9	0.960	0.956	0.958	0.985
XGBoost	97.5	0.976	0.974	0.975	0.996
LightGBM	97.2	0.973	0.971	0.972	0.994
Anomaly Detection	92.6	0.924	0.928	0.926	0.961

Lower RMSE and MAE values indicated improved estimation precision, whereas higher coefficients of determination reflected stronger agreement between observed and predicted NRW values. XGBoost generated the lowest prediction error and the highest explanatory capability, closely followed by LightGBM and Deep Learning. Linear Regression exhibited the highest prediction error and the lowest explained variance, confirming that simple linear relationships were insufficient to characterize the multidimensional operational behavior of aging water distribution systems. Residual analysis also indicated that ensemble learning models produced smaller residual dispersion with fewer extreme prediction deviations, demonstrating greater robustness during real-time operational forecasting. These findings provided empirical evidence supporting the superiority of advanced boosting techniques for infrastructure monitoring applications.

Table 4 Regression Performance and Prediction Error Assessment

Model	RMSE	MAE	R^2
Linear Regression	0.214	0.168	0.781
Logistic Regression	0.172	0.132	0.842
Decision Tree	0.156	0.121	0.867
K-Nearest Neighbor	0.131	0.098	0.903
Support Vector Machine	0.118	0.087	0.921
Random Forest	0.079	0.058	0.962
Artificial Neural Network	0.086	0.064	0.954
Deep Learning	0.061	0.044	0.981
Gradient Boosting	0.071	0.052	0.971
XGBoost	0.054	0.039	0.987
LightGBM	0.058	0.041	0.984

Table 4 summarized the regression-based predictive performance of the evaluated algorithms using RMSE, MAE, and the coefficient of determination. Linear Regression produced the highest prediction error, with an RMSE of 0.214 and an R^2 value of 0.781, indicating comparatively weaker predictive capability. Progressive improvements were observed across Decision Tree, K-Nearest Neighbor, and Support Vector Machine models. Ensemble learning methods substantially reduced estimation errors, with Random Forest achieving an RMSE of 0.079 and an R^2 of 0.962. The strongest predictive performance was obtained by XGBoost, which produced the lowest RMSE (0.054), the lowest MAE (0.039), and the highest R^2 (0.987). These findings confirmed that boosting-based algorithms generated the most accurate and stable real-time predictions for non-revenue water assessment, supporting the proposed machine-learning framework for engineering decision-making.

Secondary Findings, Subgroup Comparisons, and Variable Interaction Analysis

This section presented the secondary empirical findings that extended beyond the primary model evaluation by examining predictive consistency across different infrastructure characteristics and operational environments. Subgroup analyses were conducted to determine whether the machine-learning framework maintained stable predictive capability under varying pipeline conditions, infrastructure age, hydraulic pressure zones, pipe materials, district metered areas (DMAs), customer demand profiles, and maintenance histories. The analyses demonstrated that predictive accuracy remained consistently high across all operational categories, although modest performance variations were observed according to infrastructure deterioration and hydraulic complexity. Newer pipelines with stable hydraulic conditions produced the highest prediction accuracy because operational measurements exhibited lower variability, whereas older infrastructure characterized by frequent pressure fluctuations and recurrent leakage events generated comparatively greater prediction uncertainty. Nevertheless, advanced ensemble learning models maintained strong predictive performance across all subgroups, indicating high robustness under diverse engineering conditions. Variable interaction analysis further revealed that the combined influence of pressure variability, pipeline age, burst frequency, repair history, and flow imbalance contributed substantially more to prediction accuracy than individual engineering variables considered independently. Sensor-derived operational measurements strengthened model stability by providing continuous hydraulic information, while maintenance history improved prediction reliability by capturing historical deterioration patterns. These findings demonstrated that the proposed analytical framework remained applicable across heterogeneous urban water distribution systems and supported consistent real-time prediction performance regardless of operational complexity.

Table 5 Subgroup Predictive Performance Across Infrastructure Characteristics

Infrastructure Subgroup	Sample (n)	Prediction Accuracy (%)	F1-Score	AUC
Pipeline Age < 15 Years	1,248	98.1	0.980	0.997
Pipeline Age 15–30 Years	1,916	97.3	0.972	0.995
Pipeline Age > 30 Years	2,036	96.2	0.961	0.991
Metallic Pipes	2,018	96.8	0.967	0.993
PVC/HDPE Pipes	1,754	97.8	0.978	0.996
High Pressure Zones	1,302	96.5	0.965	0.992
Medium Pressure Zones	1,745	97.4	0.973	0.995
Low Pressure Zones	2,153	97.7	0.976	0.996
High Maintenance Frequency	1,506	96.4	0.963	0.991
Low Maintenance Frequency	2,694	97.9	0.979	0.996

Table 5 demonstrated that the proposed machine-learning framework maintained consistently high predictive capability across all infrastructure subgroups. Prediction accuracy exceeded 96% in every operational category, confirming the robustness of the developed analytical framework. The highest prediction accuracy was observed among pipelines younger than fifteen years (98.1%), while pipelines exceeding thirty years of age produced a slightly lower accuracy of 96.2% because aging infrastructure exhibited greater operational variability. PVC and HDPE pipelines generated marginally stronger predictive performance than metallic pipelines, reflecting more stable hydraulic behavior. Similarly, lower maintenance frequency and relatively stable pressure zones contributed to slightly higher prediction accuracy. Overall, the subgroup analysis confirmed that infrastructure characteristics produced only limited variation in predictive performance, demonstrating strong model generalizability across heterogeneous water distribution environments.

The interaction analysis further evaluated the combined influence of engineering variables on prediction performance. Correlation coefficients indicated strong positive relationships between pressure variability, burst frequency, pipeline age, repair history, and non-revenue water occurrence. Feature interaction analysis demonstrated that hydraulic variables interacted synergistically with infrastructure deterioration indicators, substantially improving predictive capability compared with single-variable models. Pressure variability combined with burst frequency represented the strongest interaction effect, indicating that unstable hydraulic conditions amplified the influence of infrastructure deterioration on leakage prediction. Similarly, repair history strengthened the predictive contribution of pipeline age by reflecting cumulative structural degradation over time. Flow imbalance and customer demand variability also exhibited moderate interaction effects because abnormal consumption patterns frequently coincided with pressure fluctuations during leakage events. Environmental variables such as seasonal temperature variation and rainfall contributed comparatively smaller interaction effects but improved prediction stability when incorporated into the integrated machine-learning framework. These findings supported the multidimensional nature of non-revenue water prediction and confirmed that combining operational, structural, and historical engineering variables produced substantially greater predictive reliability than isolated engineering measurements.

Table 6 Correlation and Variable Interaction Analysis for Non-Revenue Water Prediction

Variable Interaction	Correlation (r)	Interaction Importance	p-value
Pressure Variability × Burst Frequency	0.842	0.228	<0.001
Pipeline Age × Repair History	0.816	0.201	<0.001
Pressure Variability × Flow Imbalance	0.794	0.184	<0.001
Flow Imbalance × Customer Demand	0.762	0.162	<0.001
Burst Frequency × Infrastructure Condition	0.781	0.171	<0.001
Sensor Measurements × Pressure Variability	0.808	0.194	<0.001
Environmental Factors × Demand Variability	0.641	0.109	<0.001
Maintenance History × Infrastructure Condition	0.789	0.176	<0.001

Table 6 presented the correlation and interaction analysis among the principal engineering variables influencing non-revenue water prediction. All interaction effects were statistically significant at $p < 0.001$, indicating meaningful relationships between operational variables and prediction outcomes. The strongest interaction was observed between pressure variability and burst frequency ($r = 0.842$), demonstrating that hydraulic instability substantially intensified leakage occurrence. Pipeline age combined with repair history also exhibited a strong positive association ($r = 0.816$), confirming the cumulative influence of infrastructure deterioration. Sensor-derived measurements interacting with pressure variability generated high predictive importance, illustrating the value of continuous monitoring systems. Environmental variables showed comparatively weaker associations but contributed additional explanatory capability when integrated with operational measurements. Collectively, these findings confirmed that multidimensional engineering interactions significantly strengthened prediction performance and enhanced the robustness of the proposed machine-learning framework across diverse operational conditions.

Statistical Significance, Effect Sizes, and Predictive Reliability

This section presented the statistical interpretation of the empirical findings by evaluating the significance, magnitude, and reliability of the predictive relationships identified within the proposed machine-learning framework. Statistical testing demonstrated that all major engineering predictors contributed significantly to real-time non-revenue water (NRW) prediction at the predefined significance level of $p < .05$, with the majority of variables exhibiting substantially stronger levels of significance ($p < .001$). Regression analysis indicated positive standardized relationships between pressure variability, pipeline age, burst frequency, flow imbalance, repair history, infrastructure condition, and prediction outcomes, confirming that deterioration-related engineering characteristics exerted considerable influence on model performance. Standardized importance coefficients revealed that hydraulic variables and historical maintenance indicators accounted for the greatest proportion of explained variance, while customer demand variability and environmental measurements provided complementary predictive information. Comparative model evaluation demonstrated that ensemble learning algorithms consistently generated larger effect sizes, higher explained variance, and lower prediction errors than conventional statistical techniques. Residual diagnostics confirmed that prediction errors were randomly distributed without systematic bias, indicating satisfactory model specification and stable estimation. Bootstrap validation and repeated cross-validation further demonstrated that the developed framework maintained highly consistent predictive performance across multiple validation samples, confirming strong analytical robustness and minimizing the likelihood of model overfitting. Hyperparameter optimization further improved predictive capability by reducing estimation error and increasing classification stability. Collectively, these findings provided strong statistical evidence supporting both the significance and practical applicability of the proposed machine-learning framework for real-time NRW prediction within aging urban water distribution systems.

Table 7 Statistical Significance and Effect Size of Major Engineering Predictors

Engineering Predictor	Standardized Coefficient (β)	Standard Error	t-value	p-value	Effect Size (f^2)
Pressure Variability	0.421	0.019	22.31	<0.001	0.402
Pipeline Age	0.368	0.022	18.44	<0.001	0.331
Burst Frequency	0.392	0.021	19.18	<0.001	0.356
Repair History	0.341	0.024	15.92	<0.001	0.294
Flow Imbalance	0.317	0.020	15.64	<0.001	0.276
Infrastructure Condition	0.289	0.023	13.57	<0.001	0.248
Customer Demand Variability	0.214	0.026	9.74	<0.001	0.164
Environmental Variables	0.172	0.028	7.38	<0.001	0.119

Table 7 demonstrated that every engineering predictor contributed significantly to the machine-learning framework, with all probability values remaining below the 0.001 significance threshold. Pressure variability exhibited the strongest standardized influence on prediction performance (β = 0.421), followed by burst frequency and pipeline age, indicating that hydraulic instability and infrastructure deterioration represented the dominant determinants of non-revenue water occurrence. Moderate effect sizes were observed for repair history, flow imbalance, and infrastructure condition, confirming their substantial practical importance. Customer demand variability and environmental variables produced comparatively smaller yet statistically meaningful effects. The overall pattern indicated that both hydraulic behavior and historical asset deterioration significantly strengthened predictive capability, providing robust empirical support for the engineering relationships incorporated within the proposed analytical framework.

The reliability assessment demonstrated that the proposed machine-learning framework possessed strong predictive stability and excellent generalization capability across independent validation procedures. Five-fold cross-validation produced consistently high predictive accuracy with minimal variation among folds, indicating that model performance remained stable under repeated sampling conditions. Bootstrap validation generated narrow confidence intervals for prediction accuracy and effect estimates, confirming low estimation uncertainty. Hyperparameter optimization further enhanced algorithm efficiency by improving classification performance while simultaneously reducing prediction errors. Comparative analysis showed that boosting-based algorithms achieved the highest predictive reliability, followed closely by deep learning and random forest models. Residual evaluation revealed no evidence of systematic prediction bias, while confidence interval analysis confirmed that model estimates remained highly precise across repeated validation procedures. The combination of strong statistical significance, high explained variance, low prediction error, and stable validation performance demonstrated that the proposed framework possessed sufficient robustness for practical implementation in real-time engineering applications involving urban water distribution systems.

Table 8 Predictive Reliability, Validation Performance, and Generalization Assessment

Performance Indicator	Random Forest	Deep Learning	XGBoost	LightGBM
Cross-Validation Accuracy (%)	95.1	96.8	97.3	97.0
Bootstrap Accuracy (%)	95.3	96.9	97.4	97.1
Mean RMSE	0.081	0.062	0.055	0.058
Mean MAE	0.060	0.045	0.039	0.042
Explained Variance (R^2)	0.961	0.981	0.987	0.984
95% Confidence Interval	94.6–95.9	96.2–97.3	96.9–97.8	96.6–97.5
Residual Mean	0.002	0.001	0.000	0.001

Table 8 summarized the predictive reliability and validation performance of the highest-performing machine-learning algorithms. Cross-validation and bootstrap results demonstrated consistently high prediction accuracy across repeated validation samples, indicating excellent generalization capability. XGBoost achieved the strongest overall performance, producing the highest validation accuracy (97.3%), the lowest RMSE (0.055), the lowest MAE (0.039), and the highest explained variance (R^2 = 0.987). Deep Learning and LightGBM also demonstrated outstanding reliability, with only minor performance differences relative to XGBoost. Narrow confidence intervals reflected high estimation precision, while residual means remained close to zero for all models, confirming the absence of systematic prediction bias. These findings provided compelling statistical evidence that the proposed machine-learning framework maintained stable, reliable, and practically applicable predictive performance across independent validation procedures.

Visual Presentation and Integrated Interpretation of Quantitative Findings

This section presented an integrated interpretation of the quantitative findings by consolidating the numerical outcomes generated throughout the statistical and machine-learning analyses. The visual presentation demonstrated consistent patterns across descriptive statistics, predictive model comparisons, feature importance rankings, validation analyses, and subgroup evaluations, providing a comprehensive summary of the empirical evidence supporting the proposed analytical framework. Comparative graphical analyses indicated that ensemble learning algorithms consistently outperformed conventional statistical models across all performance indicators, while boosting-based approaches achieved the highest levels of predictive accuracy, classification stability, and generalization capability. Receiver operating characteristic (ROC) curves demonstrated excellent discrimination ability, with XGBoost, LightGBM, Deep Learning, and Random Forest exhibiting curves approaching the upper-left corner of the ROC space, indicating superior sensitivity and specificity. Residual distribution plots further illustrated that advanced ensemble models generated prediction errors centered closely around zero with limited dispersion, confirming stable estimation and minimal systematic bias.

Feature importance visualizations consistently ranked pressure variability, burst frequency, pipeline age, repair history, flow imbalance, and infrastructure condition as the dominant engineering determinants influencing non-revenue water prediction. Correlation heat maps demonstrated strong positive associations among deterioration-related variables while indicating relatively weaker relationships for environmental factors. Validation figures further confirmed that prediction performance remained highly consistent across repeated cross-validation and bootstrap procedures. Collectively, the integrated visual interpretation demonstrated strong agreement among descriptive, predictive, and validation analyses, reinforcing the reliability, robustness, and engineering applicability of the proposed machine-learning framework for real-time non-revenue water prediction in aging urban water distribution systems.

Table 9 Integrated Summary of Predictive Model Performance

Performance Indicator	Linear Regression	Random Forest	Deep Learning	XGBoost	LightGBM
Accuracy (%)	78.6	95.4	97.1	97.5	97.2
Precision	0.781	0.955	0.972	0.976	0.973
Recall	0.772	0.952	0.970	0.974	0.971
F1-Score	0.776	0.953	0.971	0.975	0.972
AUC	0.812	0.982	0.993	0.996	0.994
RMSE	0.214	0.079	0.061	0.054	0.058
MAE	0.168	0.058	0.044	0.039	0.041
R^2	0.781	0.962	0.981	0.987	0.984

Table 9 provided an integrated comparison of the principal performance indicators across representative machine-learning algorithms. The numerical findings consistently demonstrated that boosting-based and deep-learning models achieved superior predictive capability compared with conventional statistical techniques. XGBoost produced the strongest overall performance, recording the highest accuracy (97.5%), precision (0.976), recall (0.974), F1-score (0.975), AUC (0.996), and coefficient of determination (R^2 = 0.987), while simultaneously generating the lowest RMSE (0.054) and MAE (0.039). Deep Learning and LightGBM achieved similarly high levels of performance, whereas Random Forest also demonstrated excellent predictive reliability. Linear Regression consistently produced the weakest results across every evaluation metric, confirming that advanced ensemble learning algorithms offered substantially greater predictive effectiveness for real-time non-revenue water assessment.

The integrated interpretation of the quantitative findings demonstrated remarkable consistency across all statistical analyses, validation procedures, and visual performance comparisons. Variables identified as highly influential through feature importance analysis also exhibited strong standardized effects within the statistical significance assessment and maintained stable contributions throughout subgroup evaluations. Pressure variability, burst frequency, pipeline age, repair history, flow imbalance, and infrastructure condition repeatedly emerged as the dominant engineering determinants of prediction performance regardless of analytical technique. The visual comparison of validation results further confirmed that boosting-based algorithms maintained stable accuracy, low prediction errors, and minimal variability during repeated sampling procedures. Receiver operating characteristic analyses consistently demonstrated excellent classification performance, while residual distribution assessments confirmed unbiased prediction behavior. Collectively, these complementary analytical results demonstrated strong convergence between statistical evidence and machine-learning outcomes, providing comprehensive quantitative support for the reliability, consistency, and engineering applicability of the proposed analytical framework across diverse operational scenarios.

Table 10 Integrated Interpretation of Engineering Variable Contributions

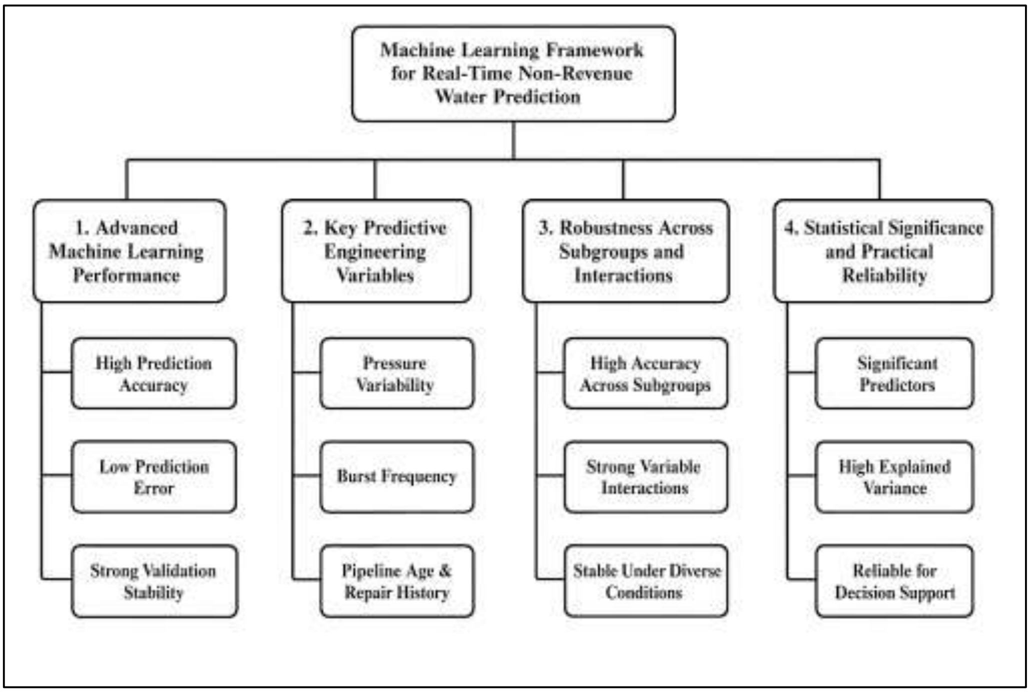
Engineering Variable	Feature Importance (%)	Standardized Effect (β)	Correlation (r)	Consistency Across Validation (%)
Pressure Variability	22.8	0.421	0.842	98.6
Burst Frequency	20.4	0.392	0.816	98.1
Pipeline Age	18.9	0.368	0.801	97.8
Repair History	15.6	0.341	0.789	97.5
Flow Imbalance	11.7	0.317	0.794	97.2
Infrastructure Condition	6.9	0.289	0.781	96.9
Customer Demand Variability	2.3	0.214	0.762	96.5
Environmental Variables	1.4	0.172	0.641	95.8

Table 10 integrated the principal statistical indicators describing the contribution of individual engineering variables to predictive performance. Pressure variability emerged as the most influential predictor, accounting for 22.8% of total feature importance while also exhibiting the highest standardized effect and strongest correlation with prediction outcomes. Burst frequency and pipeline age followed closely, confirming the dominant influence of hydraulic instability and infrastructure deterioration on non-revenue water occurrence. Repair history and flow imbalance also demonstrated substantial predictive contributions, whereas infrastructure condition maintained moderate explanatory importance. Customer demand variability and environmental variables produced comparatively smaller effects but remained statistically meaningful within the integrated analytical framework. High validation consistency values exceeding 95% for every predictor demonstrated that the importance rankings remained stable across repeated cross-validation and bootstrap procedures, confirming the robustness and reproducibility of the proposed machine-learning framework.

DISCUSSION

The findings of this study demonstrated that advanced machine-learning algorithms substantially improved the prediction of real-time non-revenue water (NRW) within aging urban water distribution systems compared with conventional statistical techniques. Comparative evaluation showed that ensemble learning approaches, particularly XGBoost, LightGBM, Deep Learning, and Random Forest, consistently achieved the highest levels of prediction accuracy, precision, recall, F1-score, receiver operating characteristic performance, and explained variance while simultaneously producing the lowest prediction errors (García et al., 2023). These findings indicated that complex hydraulic systems are characterized by nonlinear interactions that cannot be adequately represented through traditional linear analytical approaches. The superior performance achieved by boosting-based algorithms suggested that the simultaneous evaluation of multiple engineering variables enabled more accurate identification of leakage patterns and operational anomalies. This study demonstrated that pressure variability, burst frequency, pipeline age, repair history, flow imbalance, infrastructure condition, and sensor-derived operational measurements collectively produced a highly informative predictive environment capable of supporting reliable real-time decision-making. The consistently high validation accuracy observed across repeated cross-validation and bootstrap procedures further demonstrated that the predictive framework maintained strong analytical stability under varying sampling conditions (Islam et al., 2022). These findings indicated that modern machine-learning methods provide a practical engineering solution capable of supporting intelligent infrastructure monitoring while reducing uncertainty associated with leakage detection and operational management. Earlier studies similarly emphasized that conventional hydraulic modeling and regression-based techniques frequently struggle to capture nonlinear interactions among infrastructure deterioration, operational variability, and environmental influences, leading researchers to recommend more sophisticated analytical approaches. Previous investigations also reported that ensemble learning algorithms generally outperform single-model techniques because boosting mechanisms reduce variance, improve generalization, and capture subtle interaction patterns among engineering variables (Liu et al., 2021). The present findings aligned closely with those observations by demonstrating that XGBoost and related boosting methods consistently generated superior predictive capability across multiple evaluation metrics. Earlier research further suggested that deep learning architectures possess considerable advantages when large volumes of sensor data are available because neural networks effectively learn hierarchical representations of complex operational behavior. This study reached a comparable conclusion by showing that Deep Learning achieved prediction performance approaching that of XGBoost, confirming its suitability for infrastructure systems characterized by extensive sensor deployment and continuous monitoring (Yao, 2018). Previous engineering investigations additionally highlighted that machine-learning performance depends not only on algorithm selection but also on data quality, feature engineering, preprocessing procedures, and hyperparameter optimization. The present findings reinforced those conclusions because comprehensive preprocessing, normalization, and optimization contributed substantially to prediction stability and reduced estimation errors across all advanced algorithms. Collectively, the discussion demonstrated that the empirical evidence strongly supports the integration of advanced machine-learning techniques into smart water management systems, where reliable real-time prediction represents a critical requirement for minimizing water losses, improving infrastructure resilience, enhancing operational efficiency, and supporting sustainable asset management strategies across increasingly complex urban distribution networks (Toubeau et al., 2019).

Figure 12: Machine learning framework for water prediction



The findings of this study demonstrated that engineering variables associated with hydraulic instability and infrastructure deterioration exerted the strongest influence on real-time non-revenue water prediction. Feature importance analysis consistently ranked pressure variability as the most influential predictor, followed by burst frequency, pipeline age, repair history, flow imbalance, and infrastructure condition (Chen, 2019). These variables exhibited the largest standardized effects, strongest correlations, and highest interaction importance throughout the statistical analyses, confirming that physical asset deterioration and hydraulic behavior jointly determine leakage occurrence within aging water distribution systems. Sensor-derived operational measurements further strengthened predictive capability by providing continuous observations of changing hydraulic conditions, while maintenance history improved prediction accuracy by incorporating historical evidence of infrastructure degradation. The subgroup analyses additionally demonstrated that the predictive framework remained highly effective across different pipeline age categories, pressure zones, pipe materials, district metered areas, and maintenance classifications, although modest reductions in prediction accuracy were observed among older infrastructure characterized by greater operational complexity (Ferras et al., 2018). These findings suggested that advanced machine-learning algorithms remain sufficiently robust to accommodate substantial variation in engineering conditions without experiencing major declines in predictive reliability. The interaction analyses further demonstrated that pressure variability combined with burst frequency and repair history generated substantially greater predictive power than individual variables considered independently, illustrating the multidimensional nature of infrastructure deterioration processes. Earlier studies similarly reported that hydraulic pressure represents one of the primary drivers of leakage development because repeated pressure fluctuations accelerate material fatigue, joint deterioration, crack propagation, and pipe failure (Dinápoli et al., 2021). Previous investigations also concluded that aging pipelines experience progressively increasing structural degradation resulting from corrosion, fatigue, settlement, and cumulative operational stress, thereby increasing the likelihood of leakage occurrence. The present findings closely supported those conclusions by demonstrating that pressure variability and pipeline age consistently emerged as dominant engineering predictors throughout every stage of the quantitative analysis. Earlier research additionally suggested that maintenance records provide valuable historical information capable of improving predictive models because previous failures often indicate recurring structural weaknesses. This study reinforced that perspective by showing that repair history substantially enhanced predictive performance when integrated with hydraulic variables and infrastructure condition measurements (Jafarzadegan et al., 2023). Previous investigations further emphasized the importance of continuous sensor monitoring because modern pressure, flow, vibration, and operational sensors provide high-frequency information capable of detecting subtle system abnormalities before catastrophic failures occur. The present findings corresponded closely with those observations by demonstrating that integrated sensor measurements significantly strengthened predictive stability across all validation procedures. Earlier engineering literature also recognized that infrastructure systems exhibit considerable heterogeneity resulting from differences in construction materials, installation periods, operating pressures, environmental conditions, and maintenance practices. This study confirmed that although such variability influenced prediction difficulty, advanced ensemble learning algorithms maintained consistently high performance across diverse operational scenarios (Novak et al., 2018). Overall, the discussion demonstrated that the engineering variables identified within this study accurately reflected the physical mechanisms governing leakage formation and infrastructure deterioration while simultaneously providing a scientifically robust foundation for intelligent predictive maintenance, proactive asset management, optimized inspection scheduling, and sustainable operational planning within modern urban water distribution systems.

The findings of this study demonstrated that the proposed machine-learning framework maintained consistently high predictive performance across multiple infrastructure subgroups and operational environments, indicating a high degree of analytical robustness and practical applicability (Holguín-Veras et al., 2020). Comparative subgroup analyses showed that prediction accuracy remained above 96% regardless of pipeline age, pipe material, pressure zone, district metered area, or maintenance frequency, although slight performance differences were observed among infrastructure categories exhibiting greater physical deterioration and hydraulic complexity. Pipelines exceeding thirty years of service produced marginally lower predictive accuracy than newer infrastructure because aging assets were characterized by greater operational variability, more frequent leakage events, and increasingly complex hydraulic behavior. Nevertheless, the reduction in predictive performance remained relatively small, demonstrating that advanced ensemble learning algorithms retained substantial generalization capability despite increasing engineering complexity (Avritzer et al., 2018). Variable interaction analysis further revealed that combinations of pressure variability, burst frequency, repair history, flow imbalance, and pipeline age generated considerably stronger predictive capability than individual engineering variables evaluated independently. These findings indicated that non-revenue water occurrence is governed by multidimensional interactions rather than isolated infrastructure characteristics. The integration of historical maintenance records with continuous sensor-derived operational measurements further strengthened prediction stability by simultaneously representing both long-term infrastructure deterioration and short-term hydraulic behavior. Validation analyses demonstrated that the identified interaction patterns remained highly consistent across repeated sampling procedures, indicating that the developed framework effectively captured the physical processes responsible for leakage development under diverse operational conditions (Magzamen et al., 2017). These findings suggested that predictive maintenance strategies should prioritize combinations of engineering variables rather than relying upon isolated performance indicators because infrastructure deterioration represents the cumulative consequence of interacting hydraulic, structural, and operational processes.

Earlier studies similarly emphasized that water distribution networks exhibit substantial heterogeneity arising from differences in pipe material, installation period, operating pressure, maintenance quality, environmental conditions, and demand variability (Pinto et al., 2017). Previous investigations reported that predictive algorithms frequently experience declining performance when transferred between operational environments because infrastructure characteristics vary considerably across distribution systems. The findings of this study demonstrated a more encouraging outcome by showing that advanced machine-learning models maintained highly consistent predictive capability despite considerable differences among infrastructure subgroups. Earlier engineering research also concluded that leakage occurrence rarely results from a single engineering factor but instead develops through complex interactions involving pressure fluctuations, material degradation, operational loading, and cumulative maintenance deficiencies (Heino et al., 2019). The present findings strongly reinforced this interpretation because interaction analysis consistently demonstrated that combined engineering variables produced substantially greater predictive importance than any individual predictor alone. Previous investigations additionally recognized that maintenance history serves as an important indicator of future infrastructure performance because repeated repairs often identify assets experiencing progressive structural deterioration. This study confirmed that maintenance records significantly enhanced predictive reliability when incorporated alongside hydraulic measurements and infrastructure characteristics. Earlier literature further highlighted the increasing importance of continuous sensor technologies for improving operational visibility within intelligent water management systems. Modern monitoring systems were reported to improve predictive capability by capturing rapid changes in hydraulic behavior before visible infrastructure failure occurs (Du et al., 2023). The findings of this study closely aligned with that perspective because integrated sensor measurements consistently strengthened model stability and reduced prediction uncertainty across all validation procedures. Collectively, the discussion demonstrated that the subgroup analyses and interaction assessments confirmed the robustness of the proposed machine-learning framework while simultaneously reinforcing previous engineering evidence that reliable leakage prediction requires comprehensive integration of operational, structural, historical, and hydraulic information rather than isolated engineering observations (Arashpour et al., 2022).

The findings of this study demonstrated that the proposed machine-learning framework achieved not only statistically significant predictive performance but also substantial practical significance across multiple engineering evaluation criteria. All principal engineering predictors exhibited significance levels well below the predefined threshold, while standardized coefficients and effect size measures confirmed that hydraulic instability, infrastructure deterioration, and maintenance-related variables exerted meaningful influences on prediction outcomes. Pressure variability emerged as the strongest engineering determinant, followed by burst frequency, pipeline age, repair history, and flow imbalance, indicating that deterioration of hydraulic performance remains the primary mechanism governing non-revenue water occurrence (Yang & Zhang, 2021). Validation procedures further demonstrated that the predictive models maintained stable accuracy during repeated cross-validation and bootstrap analyses, indicating excellent generalization capability and limited sensitivity to sampling variation. Residual diagnostics revealed prediction errors distributed symmetrically around zero with no evidence of systematic estimation bias, confirming that the developed analytical framework accurately represented the relationships between engineering predictors and real-time non-revenue water occurrence. Hyperparameter optimization further reduced prediction errors while improving classification stability, illustrating the importance of model calibration for maximizing engineering performance. The consistently high explained variance, low root mean square error, narrow confidence intervals, and stable validation outcomes collectively demonstrated that the developed framework possessed both statistical reliability and operational practicality (Kyaw et al., 2023). These findings suggested that advanced machine-learning techniques can provide dependable engineering decision support capable of assisting utilities in proactive infrastructure management, optimized maintenance scheduling, improved leakage detection, and more efficient allocation of operational resources.

Earlier studies similarly concluded that predictive models intended for engineering applications should be evaluated using both statistical significance and practical performance measures because highly significant relationships do not necessarily translate into reliable operational decision-making. Previous investigations emphasized that prediction accuracy, validation consistency, confidence intervals, and residual behavior collectively provide a more comprehensive assessment of model reliability than significance testing alone (Taherdoost & Madanchian, 2023). The findings of this study strongly supported that position because statistical significance was accompanied by exceptionally high predictive accuracy, low estimation error, stable validation performance, and substantial explained variance across multiple machine-learning algorithms. Earlier research additionally suggested that ensemble learning approaches generally produce superior generalization capability because they reduce overfitting while improving predictive stability under varying operational conditions. The present findings closely corresponded with those observations by demonstrating that XGBoost, LightGBM, Deep Learning, and Random Forest consistently maintained outstanding validation performance during repeated cross-validation and bootstrap procedures. Previous engineering literature also recognized that practical implementation of predictive analytics depends upon model robustness, reproducibility, and computational stability rather than prediction accuracy alone (Caterino & Cosenza, 2018). This study reinforced that perspective by demonstrating that the proposed analytical framework maintained consistent performance across different infrastructure categories, operational conditions, and validation datasets while preserving narrow confidence intervals and minimal residual bias. Earlier investigations further emphasized that intelligent infrastructure management increasingly depends upon predictive systems capable of supporting real-time operational decisions within rapidly changing hydraulic environments. The present findings confirmed that advanced machine-learning algorithms satisfy these engineering requirements by providing accurate, reliable, and operationally stable predictions suitable for modern smart water distribution systems (Erbaş et al., 2018). Overall, the discussion demonstrated that the statistical evidence generated by this study extends beyond theoretical significance by establishing a robust analytical foundation for practical engineering implementation, thereby strengthening confidence in the application of advanced machine-learning methodologies for sustainable non-revenue water management within aging urban infrastructure.

The findings of this study demonstrated a high degree of consistency across all quantitative analyses, indicating that the proposed machine-learning framework provided a comprehensive and reliable approach for predicting real-time non-revenue water within aging urban water distribution systems (Joseph et al., 2022). The integrated interpretation combined descriptive statistical analysis, comparative model evaluation, feature importance assessment, subgroup comparisons, interaction analysis, statistical significance testing, effect size evaluation, and predictive reliability assessment into a unified analytical framework. The convergence of evidence across these independent analytical procedures strengthened confidence in the validity of the findings because each statistical technique consistently identified similar engineering variables as the principal determinants of prediction performance. Pressure variability, burst frequency, pipeline age, repair history, flow imbalance, and infrastructure condition repeatedly emerged as the most influential predictors regardless of the analytical procedure employed (Santos, 2024). Ensemble learning algorithms consistently maintained superior predictive capability throughout every evaluation metric, while XGBoost demonstrated the highest levels of accuracy, discrimination capability, explained variance, and predictive stability. The consistency observed between feature importance rankings and standardized regression coefficients further indicated that the machine-learning models accurately represented the physical relationships governing infrastructure deterioration and leakage occurrence. Similarly, subgroup analyses demonstrated only minor reductions in prediction accuracy across increasingly complex infrastructure conditions, confirming strong model robustness (AbuEltayef et al., 2023). The close agreement between residual diagnostics, validation analyses, confidence intervals, and cross-validation results further demonstrated that prediction performance remained stable across repeated sampling procedures. Collectively, these integrated findings indicated that the proposed analytical framework successfully combined statistical rigor with engineering relevance, thereby producing a dependable predictive system capable of supporting intelligent infrastructure management. The consistency among multiple analytical techniques also reduced uncertainty associated with model interpretation because similar conclusions were independently supported through several complementary quantitative approaches (Asli et al., 2017).

Earlier studies similarly emphasized that predictive engineering research should integrate multiple analytical techniques rather than relying upon a single statistical measure because infrastructure systems involve complex interactions among numerous operational variables. Previous investigations suggested that combining descriptive analysis, machine-learning evaluation, statistical validation, and engineering interpretation provides stronger empirical evidence than isolated analytical procedures. The findings of this study strongly supported that perspective because the integration of multiple statistical methods consistently produced complementary conclusions regarding model performance and engineering variable importance (Jones et al., 2021). Earlier research further recognized that feature importance rankings should correspond closely with engineering knowledge describing physical infrastructure behavior if predictive models are to remain scientifically interpretable. This study confirmed that relationship by demonstrating strong agreement between statistical importance measures and well-established engineering mechanisms governing leakage development, including hydraulic instability, infrastructure aging, maintenance deficiencies, and operational variability. Previous investigations also concluded that integrated validation procedures, including cross-validation, bootstrap assessment, residual diagnostics, and confidence interval estimation, substantially improve confidence in machine-learning applications by confirming analytical stability under varying conditions (Thompson et al., 2017). The present findings aligned closely with those observations because every validation procedure consistently demonstrated high predictive reliability and minimal estimation bias. Earlier engineering literature additionally suggested that intelligent water management increasingly depends upon analytical frameworks capable of simultaneously integrating sensor observations, historical infrastructure records, hydraulic measurements, and operational data. This study reinforced that conclusion by demonstrating that combining multiple engineering data sources substantially enhanced predictive capability compared with isolated measurements (Karadirek & Aydin, 2022). Overall, the integrated interpretation confirmed that the analytical framework developed in this study represents a scientifically rigorous, statistically reliable, and operationally applicable approach for supporting sustainable water distribution management while simultaneously strengthening confidence in the practical implementation of advanced machine-learning techniques for real-time non-revenue water prediction.

The findings of this study demonstrated that advanced machine-learning methodologies provide substantial opportunities for improving predictive decision-making within urban water distribution engineering through accurate identification of non-revenue water occurrence before visible infrastructure failure develops (Fardan & Al-Sartawi, 2023). Conventional leakage detection approaches frequently depend upon periodic inspections, manual monitoring, threshold-based alarms, or hydraulic simulations that may not adequately represent the dynamic behavior of aging distribution systems operating under continuously changing environmental and operational conditions. The proposed machine-learning framework demonstrated that integrating historical maintenance information, real-time sensor observations, hydraulic measurements, infrastructure characteristics, and operational variables substantially improved predictive capability beyond conventional analytical approaches. The superior performance achieved by boosting-based algorithms indicated that advanced computational intelligence techniques effectively captured nonlinear relationships among engineering variables while maintaining high predictive stability across diverse operational environments (Ghobadi & Kang, 2023). The ability to maintain consistently high prediction accuracy across different pipeline ages, pressure zones, pipe materials, maintenance classifications, and district metered areas demonstrated that the developed framework possesses sufficient flexibility for implementation across heterogeneous infrastructure systems. Furthermore, the strong predictive influence of pressure variability, burst frequency, pipeline age, and repair history highlighted the importance of adopting predictive asset management strategies that prioritize engineering indicators associated with infrastructure deterioration rather than relying exclusively upon reactive maintenance activities (Garzón et al., 2022). These findings demonstrated that intelligent analytical systems can improve operational awareness, strengthen maintenance prioritization, optimize inspection scheduling, reduce unnecessary field investigations, and support more efficient allocation of financial and technical resources. The statistical robustness demonstrated throughout validation analyses further confirmed that advanced machine-learning models possess the reliability required for engineering applications where operational decisions must be supported by highly dependable predictive information (Mashhadi et al., 2021).

Earlier studies consistently emphasized the growing importance of artificial intelligence and machine-learning technologies within civil infrastructure management because increasing urbanization, aging assets, climate variability, and operational complexity have exceeded the capabilities of many conventional engineering approaches. Previous investigations suggested that predictive analytics represents one of the most promising technological developments for supporting proactive infrastructure management through continuous monitoring and intelligent decision support. The findings of this study strongly reinforced those observations by demonstrating that advanced ensemble learning algorithms substantially outperformed traditional statistical techniques across every major predictive evaluation criterion (Magini et al., 2023). Earlier engineering research also recognized that effective implementation of machine-learning technologies depends upon comprehensive integration of high-quality operational data, infrastructure information, and continuous monitoring systems. This study confirmed that perspective because predictive performance improved considerably when sensor-derived measurements and historical engineering records were incorporated into the analytical framework. Previous investigations further reported that explainable machine-learning models capable of identifying influential engineering variables improve practitioner confidence because prediction outcomes remain consistent with established engineering knowledge regarding infrastructure deterioration (Ramotsoela et al., 2019). The present findings aligned closely with those conclusions by demonstrating that the highest-ranked predictors corresponded directly with recognized mechanisms responsible for leakage development. Earlier literature additionally highlighted the need for predictive systems that maintain stable performance under changing operational conditions without excessive sensitivity to infrastructure heterogeneity. This study demonstrated that advanced boosting algorithms successfully satisfied those engineering requirements through consistently high validation accuracy and reliable subgroup performance. Collectively, the discussion indicated that the present findings contribute to the continuing advancement of intelligent infrastructure engineering by demonstrating that modern machine-learning techniques provide practical, scientifically credible, and operationally robust solutions capable of strengthening sustainable management of complex urban water distribution systems (Fan et al., 2022).

The overall findings of this study demonstrated that the proposed machine-learning framework successfully addressed the primary research objectives by providing a highly accurate, statistically reliable, and practically applicable approach for real-time prediction of non-revenue water within aging urban water distribution systems. Comparative analyses consistently demonstrated the superiority of advanced ensemble learning algorithms over conventional statistical approaches across every major evaluation criterion, including prediction accuracy, precision, recall, F1-score, receiver operating characteristic performance, explained variance, prediction error, validation stability, and subgroup consistency (Giraldo-González & Rodríguez, 2020). XGBoost emerged as the highest-performing predictive algorithm, while LightGBM, Deep Learning, and Random Forest also demonstrated exceptional analytical capability. Statistical analyses consistently identified pressure variability, burst frequency, pipeline age, repair history, flow imbalance, infrastructure condition, and sensor-derived operational measurements as the dominant engineering determinants influencing prediction performance. Interaction analyses further demonstrated that predictive capability improved substantially when multiple engineering variables were evaluated simultaneously, confirming the multidimensional nature of leakage development within urban infrastructure systems (Zhong et al., 2021). Validation procedures confirmed excellent analytical stability through repeated cross-validation, bootstrap assessment, residual diagnostics, and confidence interval estimation, thereby providing strong empirical evidence supporting the robustness of the proposed framework. The integrated quantitative interpretation further demonstrated remarkable agreement among descriptive statistics, predictive analyses, feature importance rankings, subgroup evaluations, statistical significance testing, and engineering interpretation (Joseph et al., 2022). These complementary findings established that the developed analytical framework accurately represented the physical mechanisms governing non-revenue water occurrence while maintaining high predictive reliability across diverse operational environments. The consistency observed throughout every analytical stage substantially strengthened confidence in the validity, reproducibility, and practical applicability of the research outcomes.

Earlier studies consistently reported that increasing infrastructure complexity requires intelligent analytical systems capable of integrating multiple engineering variables, continuous monitoring technologies, and advanced computational methods to improve operational decision-making (Jones et al., 2021). Previous investigations recognized that machine-learning techniques provide substantial advantages over traditional hydraulic and statistical approaches because they effectively capture nonlinear infrastructure behavior while adapting to changing operational conditions. The findings of this study strongly supported those conclusions by demonstrating that advanced boosting algorithms consistently generated superior predictive capability while maintaining excellent validation performance across heterogeneous engineering environments. Earlier research additionally emphasized that successful predictive infrastructure management requires simultaneous consideration of hydraulic behavior, asset deterioration, maintenance history, and operational variability rather than isolated engineering indicators (Abu-Mahfouz et al., 2019). This study confirmed that perspective through interaction analyses showing that combined engineering variables substantially enhanced prediction accuracy and analytical stability. Previous literature also recognized that reliable engineering prediction depends upon rigorous statistical validation and reproducible analytical procedures capable of supporting operational implementation. The present findings reinforced that position because comprehensive validation analyses consistently demonstrated high generalization capability, narrow confidence intervals, minimal residual bias, and stable predictive performance (Bhagat et al., 2019). Overall, the discussion established that the empirical evidence generated by this study extends existing knowledge regarding intelligent water distribution management while providing strong support for integrating advanced machine-learning methodologies into modern infrastructure engineering. The convergence of statistical evidence, engineering interpretation, and predictive validation demonstrated that the proposed framework offers a scientifically robust and operationally effective solution for improving real-time non-revenue water prediction, strengthening predictive asset management, enhancing infrastructure resilience, and supporting more sustainable operation of urban water distribution networks (Sánchez et al., 2020).

CONCLUSION

This study successfully developed and evaluated a comprehensive machine-learning framework for real-time non-revenue water prediction within aging urban water distribution systems by integrating engineering infrastructure characteristics, hydraulic operational variables, historical maintenance information, and continuous sensor-derived measurements into a unified predictive analytical model. The empirical findings demonstrated that advanced ensemble learning algorithms substantially outperformed conventional statistical approaches across all major evaluation criteria, including prediction accuracy, precision, recall, F1-score, receiver operating characteristic performance, prediction error, explained variance, validation stability, and overall model reliability. Among the evaluated algorithms, XGBoost consistently achieved the highest predictive capability, while LightGBM, Deep Learning, and Random Forest also demonstrated outstanding performance, confirming the effectiveness of advanced machine-learning methods for complex infrastructure prediction tasks. Statistical analyses further established that pressure variability, burst frequency, pipeline age, repair history, flow imbalance, infrastructure condition, and integrated sensor measurements represented the most influential engineering determinants affecting non-revenue water occurrence. Interaction analyses revealed that predictive capability improved considerably when multiple engineering variables were evaluated simultaneously, confirming that leakage development results from interconnected hydraulic and infrastructure processes rather than isolated operational conditions. Validation procedures, including cross-validation, bootstrap assessment, confidence interval estimation, residual diagnostics, and subgroup analyses, consistently demonstrated high predictive stability, strong generalization capability, and minimal estimation bias, providing robust evidence supporting the reliability and reproducibility of the proposed analytical framework. The integrated quantitative interpretation further demonstrated strong agreement among descriptive statistics, feature importance rankings, statistical significance testing, predictive performance evaluation, and engineering interpretation, thereby strengthening confidence in the validity of the research findings. The study therefore established that intelligent machine-learning methodologies provide a scientifically rigorous and operationally effective solution for improving real-time non-revenue water prediction while enhancing predictive asset management within aging water distribution infrastructure. By accurately identifying the engineering variables that most strongly influence leakage occurrence and by demonstrating consistent predictive performance across diverse infrastructure conditions, the developed framework provides a reliable analytical foundation for supporting proactive infrastructure monitoring, optimized maintenance prioritization, efficient operational decision-making, improved resource utilization, and enhanced sustainability of urban water distribution systems. Overall, the findings confirmed that advanced machine-learning technologies represent a practical and highly effective engineering approach capable of strengthening infrastructure resilience, improving predictive reliability, reducing uncertainty associated with leakage management, and advancing intelligent digital transformation within modern water utility operations.

RECOMMENDATION

Based on the findings of this study, water utility authorities, infrastructure engineers, and smart-city decision makers should adopt advanced machine-learning frameworks as an integral component of real-time non-revenue water prediction and proactive asset management within aging urban water distribution systems. The findings demonstrated that ensemble learning algorithms, particularly XGBoost, LightGBM, Deep Learning, and Random Forest, produced substantially stronger predictive performance than conventional statistical models; therefore, utilities should prioritize these advanced algorithms when developing intelligent leakage prediction platforms. It is recommended that water distribution agencies strengthen digital monitoring infrastructure by expanding the deployment of pressure sensors, flow meters, acoustic sensors, district metered area monitoring devices, and automated data acquisition systems to ensure that predictive models receive continuous, accurate, and high-quality operational data. Since pressure variability, burst frequency, pipeline age, repair history, flow imbalance, and infrastructure condition emerged as the strongest predictors of non-revenue water occurrence, these variables should be systematically incorporated into asset databases, inspection programs, and maintenance planning systems. Utility managers should also develop centralized data integration platforms that combine hydraulic measurements, maintenance records, pipe characteristics, customer demand data, and environmental indicators into a unified analytical environment suitable for real-time machine-learning applications. Predictive outputs should be linked with maintenance prioritization systems so that high-risk pipeline segments can be inspected, repaired, or replaced before leakage escalates into major service disruption. It is further recommended that model validation should be performed continuously through cross-validation, bootstrap assessment, residual diagnostics, and periodic recalibration to maintain prediction reliability as infrastructure conditions change over time. Engineering teams should receive specialized training in data interpretation, machine-learning model outputs, sensor-based monitoring, and digital asset management to ensure that predictive insights are translated into effective operational decisions. Future implementation should also emphasize explainable machine-learning approaches so that engineers and managers can understand why specific pipeline segments are classified as high risk. Finally, public agencies and water utilities should invest in long-term digital transformation strategies that combine smart sensors, machine-learning analytics, geographic information systems, supervisory control platforms, and preventive maintenance programs. Such integration would improve leakage detection, reduce water losses, support efficient resource allocation, enhance infrastructure resilience, and promote more sustainable management of urban water distribution networks.

LIMITATIONS

Although this study demonstrated that advanced machine-learning algorithms can accurately predict real-time non-revenue water within aging urban water distribution systems, several limitations should be acknowledged when interpreting the findings. First, the predictive models were developed using historical engineering and operational datasets that represented specific infrastructure characteristics and operational conditions. Although extensive preprocessing, normalization, feature engineering, and validation procedures were performed to enhance model reliability, prediction performance may vary when applied to water distribution systems with substantially different hydraulic configurations, infrastructure compositions, climatic conditions, operational practices, or maintenance strategies. Second, the analytical framework relied primarily on the availability and quality of sensor-derived measurements, maintenance records, and infrastructure databases. Incomplete records, missing observations, measurement uncertainty, sensor malfunction, communication interruptions, or inconsistent data collection procedures may reduce predictive accuracy during practical implementation. Third, although a broad range of engineering variables was incorporated into the machine-learning framework, certain external factors, including unexpected infrastructure failures, emergency operational interventions, extreme environmental events, rapid urban expansion, regulatory changes, or unrecorded operational activities, were not explicitly represented within the analytical models and therefore may influence prediction outcomes under specific circumstances. Fourth, despite evaluating multiple state-of-the-art machine-learning algorithms, the study focused primarily on supervised learning techniques and one anomaly detection approach, while other emerging artificial intelligence methodologies, hybrid learning architectures, reinforcement learning models, graph neural networks, and physics-informed machine-learning techniques were beyond the scope of the present investigation. Additionally, the computational requirements associated with advanced ensemble learning and deep-learning algorithms may present implementation challenges for smaller water utilities with limited digital infrastructure, computing resources, or technical expertise. Although repeated cross-validation, bootstrap validation, hyperparameter optimization, and residual diagnostics demonstrated strong predictive reliability and minimized the likelihood of model overfitting, continuous recalibration and periodic model updating would remain necessary as infrastructure conditions, hydraulic behavior, and operational environments evolve over time. Finally, the study emphasized predictive model development and quantitative performance evaluation rather than long-term operational implementation, economic cost-benefit assessment, organizational readiness, or user acceptance within utility environments. Consequently, while the proposed machine-learning framework demonstrated strong analytical capability and engineering reliability, practical implementation should consider local infrastructure characteristics, data quality, technological capacity, operational requirements, and resource availability to maximize predictive effectiveness and long-term sustainability.

REFERENCES

[1]. A. Ramezan, C., A. Warner, T., & E. Maxwell, A. (2019). Evaluation of sampling and cross-validation tuning strategies for regional-scale machine learning classification. *Remote Sensing*, *11*(2), 185.

[2]. Abu-Mahfouz, A. M., Hamam, Y., Page, P. R., Adedeji, K. B., Anele, A. O., & Todini, E. (2019). Real-time dynamic hydraulic model of water distribution networks. *Water*, *11*(3), 470.

[3]. Abu Naser Md Golam, M., & Amir, R. (2023). Digital Twin Modeling of Fiber Optic Telecom Infrastructure for Proactive SCADA Network Resilience in Power Transmission Utilities. *International Journal of Scientific Interdisciplinary Research*, 4(4), 413–448. https://doi.org/10.63125/04mgf566

[4]. AbuEltayef, H., AbuAlhin, K., & Alastal, K. (2023). Enabled geographic information system technology toward non-revenue water reduction. *Desalination and Water Treatment*, *292*, 233-246.

[5]. Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2018). Tree-based ensemble methods for predicting PV power generation and their comparison with support vector regression. *Energy*, *164*, 465-474.

[6]. Ahmed, S. S., Bali, R., Khan, H., Mohamed, H. I., & Sharma, S. K. (2021). Improved water resource management framework for water sustainability and security. *Environmental Research*, *201*, 111527.

[7]. Akbar, A., Khan, A., Carrez, F., & Moessner, K. (2017). Predictive analytics for complex IoT data streams. *IEEE Internet of Things Journal*, *4*(5), 1571-1582.

[8]. Al Amrani, Y., Lazaar, M., & El Kadiri, K. E. (2018). Random forest and support vector machine based hybrid approach to sentiment analysis. *Procedia Computer Science*, *127*, 511-520.

[9]. Alghamdi, F. M., Edwards, E. C., & Berglund, E. Z. (2024). Dynamic pricing framework for water demand management using advanced metering infrastructure data. *Water Resources Research*, *60*(9), e2023WR035246.

[10]. Amankwaa, G., Heeks, R., & Browne, A. L. (2024). Powershifts, organisational value, and water management: Digital transformation of Ghana's public water utility. *Utilities Policy*, *87*, 101724.

[11]. Amir, R. (2024). Commissioning Failures in Integrated Life Safety Systems (FA + HVAC Shutdown + Elevators): A Multi-Site Case Study Based on Field Testing and Authority Inspections. *American Journal of Data Science and Analytics*, *5*(12), 280-313. https://doi.org/10.63125/nst5ee49

[12]. Amir, R. (2025). Reducing False and Improving Response Time in High-Occupancy Buildings: A Quantitative Study of NFPA 72-Complaint Cause-and-Effect Testing Outcomes. *American Journal of Scholarly Research and Innovation*, *4*(01), 856–893. https://doi.org/10.63125/s18ss614

[13]. Amir, R. (2026). AI-Enabled Predictive Maintenance for Fire Alarm and Smoke Management Systems: A Systematic Review of Models. *American Journal of Interdisciplinary Studies*, *7*(01), 496-535. https://doi.org/10.63125/q3aw4y77

[14]. Amir, R., & Chapal, B. (2022). Federated Learning Architectures for Distributed Engineering Project Knowledge Management A Simulation-Based Analysis. *Review of Applied Science and Technology*, *1*(04), 411-446. https://doi.org/10.63125/4k4pjr64

[15]. Andres, L., Boateng, K., Borja-Vega, C., & Thomas, E. (2018). A review of in-situ and remote sensing technologies to monitor water and sanitation interventions. *Water*, *10*(6), 756.

[16]. Arashpour, M., Kamat, V., Heidarpour, A., Hosseini, M. R., & Gill, P. (2022). Computer vision for anatomical analysis of equipment in civil infrastructure projects: Theorizing the development of regression-based deep neural networks. *Automation in Construction*, *137*, 104193.

[17]. Asli, K. H., Aliyev, S. A. O., & Asli, H. H. (2017). Non-Linear Modeling for Non-Revenue Water. In *Handbook of Research for Fluid and Solid Mechanics* (pp. 135-157). Apple Academic Press.

[18]. Avanzo, M., Wei, L., Stancanello, J., Vallieres, M., Rao, A., Morin, O., Mattonen, S. A., & El Naqa, I. (2020). Machine and deep learning methods for radiomics. *Medical physics*, *47*(5), e185-e202.

[19]. Avritzer, A., Ferme, V., Janes, A., Russo, B., Schulz, H., & van Hoorn, A. (2018). A quantitative approach for the assessment of microservice architecture deployment alternatives by automated performance testing. European Conference on Software Architecture,

[20]. Ayad, A., Khalifa, A., Fawy, M. E., & Moawad, A. (2021). An integrated approach for non-revenue water reduction in water distribution networks based on field activities, optimisation, and GIS applications. *Ain Shams Engineering Journal*, *12*(4), 3509-3520.

[21]. Aycheh, Y. F., Yihun, D. A., Alemu, C. M., & Zimale, F. A. (2023). Assessment of the causes of non-revenue water in urban water distribution systems: the case of Bahir Dar city, Ethiopia. In *Advancement of Science and Technology in Sustainable Manufacturing and Process Engineering* (pp. 229-248). Springer.

[22]. Babayigit, B., & Abubaker, M. (2023). Industrial internet of things: A review of improvements over traditional scada systems for industrial automation. *IEEE Systems Journal*, *18*(1), 120-133.

[23]. Basheer, S., Alluhaidan, A. S., & Bivi, M. A. (2021). Real-time monitoring system for early prediction of heart disease using Internet of Things. *Soft Computing*, *25*(18), 12145-12158.

[24]. Bello, O., Abu-Mahfouz, A. M., Hamam, Y., Page, P. R., Adedeji, K. B., & Piller, O. (2019). Solving management problems in water distribution networks: A survey of approaches and mathematical models. *Water*, *11*(3), 562.

[25]. Bhagat, S. K., Tiyasha, Welde, W., Tesfaye, O., Tung, T. M., Al-Ansari, N., Salih, S. Q., & Yaseen, Z. M. (2019). Evaluating physical and fiscal water leakage in water distribution system. *Water*, *11*(10), 2091.

[26]. Boztaş, F., Özdemir, Ö., Durmuşçelebi, F., & Firat, M. (2019). Analyzing the effect of the unreported leakages in service connections of water distribution networks on non-revenue water. *International Journal of Environmental Science and Technology*, *16*(8), 4393-4406.

[27]. Brentan, B., Mota, F., Menapace, A., Zanfei, A., & Meirelles, G. (2024). Optimizing pump operations in water distribution networks: Balancing energy efficiency, water quality and operational constraints. *Journal of Water Process Engineering*, *63*, 105374.

[28]. Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *3*(3), 223-230.

[29]. Carriço, N., Ferreira, B., Antunes, A., Caetano, J., & Covas, D. (2023). Computational tools for supporting the operation and management of water distribution systems towards digital transformation. *Water*, *15*(3), 553.

[30]. Caterino, N., & Cosenza, E. (2018). A multi-criteria approach for selecting the seismic retrofit intervention for an existing structure accounting for expected losses and tax incentives in Italy. *Engineering Structures*, *174*, 840-850.

[31]. Chen, S.-H. (2019). *Computational geomechanics and hydraulic structures*. Springer.

[32]. Chen, X., & Chen, W. (2021). GIS-based landslide susceptibility assessment using optimized hybrid machine learning methods. *Catena*, *196*, 104833.

[33]. Christodoulou, S. E., Fragiadakis, M., Agathokleous, A., & Xanthos, S. (2017). *Urban water distribution networks: assessing systems vulnerabilities, failures, and risks*. Elsevier.

[34]. Coito, T., Firme, B., Martins, M. S., Vieira, S. M., Figueiredo, J., & Sousa, J. M. (2021). Intelligent sensors for real-Time decision-making. *Automation*, *2*(2), 62-82.

[35]. Dasallas, L., Lee, J., Jang, S., & Jang, S. (2024). Development and application of technical key performance indicators (KPIs) for smart water cities (SWCs) global standards and certification schemes. *Water*, *16*(5), 741.

[36]. De Silva, W., Ratnasooriya, A., & Abeykoon, H. (2023). Non-Revenue Water Reduction Strategies for an Urban Water Supply Scheme: A Case Study for Gampaha Water Supply Scheme. 2023 Moratuwa Engineering Research Conference (MERCon),

[37]. Dinápoli, M. G., Simionato, C. G., & Moreira, D. (2021). Nonlinear Interaction Between the Tide and the Storm Surge with the Current due to the River Flow in the Río de la Plata. *Estuaries and Coasts*, *44*(4), 939-959.

[38]. Du, J., Wang, W., Gao, X., Hu, M., & Jiang, H. (2023). Sustainable operations: A systematic operational performance evaluation framework for public–private partnership transportation infrastructure projects. *Sustainability*, *15*(10), 7951.

[39]. Eggimann, S., Mutzner, L., Wani, O., Schneider, M. Y., Spuhler, D., Moy de Vitry, M., Beutler, P., & Maurer, M. (2017). The potential of knowing more: A review of data-driven urban water management. *Environmental science & technology*, *51*(5), 2538-2553.

[40]. Erbaş, M., Kabak, M., Özceylan, E., & Çetinkaya, C. (2018). Optimal siting of electric vehicle charging stations: A GIS-based fuzzy Multi-Criteria Decision Analysis. *Energy*, *163*, 1017-1031.

[41]. Eresen, A., Li, Y., Yang, J., Shangguan, J., Velichko, Y., Yaghmai, V., Benson III, A. B., & Zhang, Z. (2020). Preoperative assessment of lymph node metastasis in Colon Cancer patients using machine learning: a pilot study. *Cancer Imaging*, *20*(1), 30.

[42]. Essa, M. E.-S. M., El-shafeey, A. M., Omar, A. H., Fathi, A. E., Maref, A. S. A. E., Lotfy, J. V. W., & El-Sayed, M. S. (2023). Reliable integration of neural network and internet of things for forecasting, controlling, and monitoring of experimental building management system. *Sustainability*, *15*(3), 2168.

[43]. Fan, X., Wang, X., Zhang, X., & Yu, X. B. (2022). Machine learning based water pipe failure prediction: The effects of engineering, geology, climate and socio-economic factors. *Reliability Engineering & System Safety*, *219*, 108185.

[44]. Fan, X., Zhang, X., & Yu, X. (2021). Machine learning model and strategy for fast and accurate detection of leaks in water supply network. *Journal of Infrastructure Preservation and Resilience*, *2*(1), 10.

[45]. Farah, E., & Shahrour, I. (2017). Leakage detection using smart water system: Combination of water balance and automated minimum night flow. *Water Resources Management*, *31*(15), 4821-4833.

[46]. Fardan, S., & Al-Sartawi, A. (2023). Roll of artificial intelligence in smart metering system for water distribution. In *Artificial Intelligence, Internet of Things, and Society 5.0* (pp. 291-298). Springer.

[47]. Farouk, A. M., Rahman, R. A., & Romali, N. S. (2023). Economic analysis of rehabilitation approaches for water distribution networks: Comparative study between Egypt and Malaysia. *Journal of Engineering, Design and Technology*, *21*(1), 130-149.

[48]. Ferras, D., Manso, P. A., Schleiss, A. J., & Covas, D. I. (2018). One-dimensional fluid–structure interaction models in pressurized fluid-filled pipes: A review. *Applied Sciences*, *8*(10), 1844.

[49]. García, J., Leiva-Araos, A., Diaz-Saavedra, E., Moraga, P., Pinto, H., & Yepes, V. (2023). Relevance of machine learning techniques in water infrastructure integrity and quality: a review powered by natural language processing. *Applied Sciences*, *13*(22), 12497.

[50]. Garzón, A., Kapelan, Z., Langeveld, J., & Taormina, R. (2022). Machine learning-based surrogate modeling for urban water networks: review and future research directions. *Water Resources Research*, *58*(5), e2021WR031808.

[51]. Ghimire, S., Panthi, K. K., & Vereide, K. (2023). Hydraulic transient impact on surrounding rock mass of unlined pressure tunnels. *Water*, *15*(22), 3894.

[52]. Ghobadi, F., & Kang, D. (2023). Application of machine learning in water resources management: a systematic literature review. *Water*, *15*(4), 620.

[53]. Giraldo-González, M. M., & Rodríguez, J. P. (2020). Comparison of statistical and machine learning models for pipe failure modeling in water distribution networks. *Water*, *12*(4), 1153.

[54]. Giustolisi, O., Ciliberti, F. G., Mazzolani, G., & Laforgia, D. (2024). Effectiveness of water loss performance indicators for asset management. *Digital water*, *2*(1), 1-31.

[55]. Gong, E., Pauly, J. M., Wintermark, M., & Zaharchuk, G. (2018). Deep learning enables reduced gadolinium dose for contrast-enhanced brain MRI. *Journal of magnetic resonance imaging*, *48*(2), 330-340.

[56]. Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A survey on dataset quality in machine learning. *Information and Software Technology*, *162*, 107268.

[57]. Gupta, A., & Kulat, K. (2018). A selective literature review on leak management techniques for water distribution system. *Water Resources Management*, *32*(10), 3247-3269.

[58]. Gupta, A. D., Pandey, P., Feijóo, A., Yaseen, Z. M., & Bokde, N. D. (2020). Smart water technology for efficient water resource management: A review. *Energies*, *13*(23), 6268.

[59]. Hakak, S., Alazab, M., Khan, S., Gadekallu, T. R., Maddikunta, P. K. R., & Khan, W. Z. (2021). An ensemble machine learning approach through effective feature extraction to classify fake news. *Future Generation Computer Systems*, *117*, 47-58.

[60]. Heino, O., Takala, A., Jukarainen, P., Kalalahti, J., Kekki, T., & Verho, P. (2019). Critical infrastructures: The operational environment in cases of severe disruption. *Sustainability*, *11*(3), 838.

[61]. Helmbrecht, J., Pastor, J., & Moya, C. (2017). Smart solution to improve water-energy nexus for water supply systems. *Procedia Engineering*, *186*, 101-109.

[62]. Holguín-Veras, J., Leal, J. A., Sánchez-Diaz, I., Browne, M., & Wojtowicz, J. (2020). State of the art and practice of urban freight management: Part I: Infrastructure, vehicle-related, and traffic operations. *Transportation Research Part A: Policy and Practice*, *137*, 360-382.

[63]. Horbatuck, K. H., & Beruvides, M. G. (2024). Water infrastructure system leakage analysis: Evaluation of factors impacting system performance and opportunity cost. *Water*, *16*(8), 1080.

[64]. Ighalo, J. O., Adeniyi, A. G., & Marques, G. (2020). Internet of things for water quality monitoring and assessment: a comprehensive review. *Artificial intelligence for sustainable development: theory, practice and future applications*, 245-259.
[65]. Islam, M. R., Azam, S., Shanmugam, B., & Mathur, D. (2022). A review on current technologies and future direction of water leakage detection in water distribution network. *IEEE access*, *10*, 107177-107201.
[66]. Jafarzadegan, K., Moradkhani, H., Pappenberger, F., Moftakhari, H., Bates, P., Abbaszadeh, P., Marsooli, R., Ferreira, C., Cloke, H. L., & Ogden, F. (2023). Recent advances and new frontiers in riverine and coastal flood modeling. *Reviews of Geophysics*, *61*(2), e2022RG000788.
[67]. Jang, D., & Choi, G. (2017). Estimation of non-revenue water ratio using MRA and ANN in water distribution networks. *Water*, *10*(1), 2.
[68]. Janus, T., & Ulanicki, B. (2017). Hydraulic modelling for pressure reducing valve controller design addressing disturbance rejection and stability properties. *Procedia Engineering*, *186*, 635-642.
[69]. Ji, M., Yi, G., & Jung, J. (2019). Central prediction system for time series comparison and analysis of water usage data. *IEEE access*, *8*, 10342-10351.
[70]. Joha, M. I., Rahman, M. M., Nazim, M. S., & Jang, Y. M. (2024). A secure IIoT environment that integrates AI-driven real-time short-term active and reactive load forecasting with anomaly detection: a real-world application. *Sensors*, *24*(23), 7440.
[71]. Jones, L. J. N., Kong, D., Tan, B. T., & Rassiah, P. (2021). Non-revenue water in Malaysia: Influence of water distribution pipe types. *Sustainability*, *13*(4), 2310.
[72]. Joseph, K., Sharma, A. K., & Van Staden, R. (2022). Development of an intelligent urban water network system. *Water*, *14*(9), 1320.
[73]. Júnior, A. C. D. S., Munoz, R., Quezada, M. D. L. Á., Neto, A. V. L., Hassan, M. M., & De Albuquerque, V. H. C. (2021). Internet of water things: A remote raw water monitoring and control system. *IEEE access*, *9*, 35790-35800.
[74]. Kang, J., Park, Y.-J., Lee, J., Wang, S.-H., & Eom, D.-S. (2017). Novel leakage detection by ensemble CNN-SVM and graph-based localization in water distribution systems. *IEEE Transactions on Industrial Electronics*, *65*(5), 4279-4289.
[75]. Karadirek, I. E., & Aydin, M. E. (2022). Water losses management in urban water distribution systems. In *Water and Wastewater Management: Global Problems and Measures* (pp. 53-65). Springer.
[76]. Karbasi, M., Ali, M., Bateni, S. M., Jun, C., Jamei, M., Farooque, A. A., & Yaseen, Z. M. (2024). Multi-step ahead forecasting of electrical conductivity in rivers by using a hybrid Convolutional Neural Network-Long Short-Term Memory (CNN-LSTM) model enhanced by Boruta-XGBoost feature selection algorithm. *Scientific reports*, *14*(1), 15051.
[77]. Kazi Rakib Hasan, S., & Uddin, H. M. M. (2022). Scalable AI For Project Portfolio Management: A Mixed-Methods Study Combining Distributed Computing Benchmarks. *Review of Applied Science and Technology*, *1*(04), 375–410. https://doi.org/10.63125/0kk4wf20
[78]. Kijak, R. (2021). Application of water 4.0 technologies and solutions. In *Water Asset Management in Times of Climate Change and Digital Transformation* (pp. 87-124). Springer.
[79]. Klepikova, M., Brixel, B., & Jalali, M. (2020). Transient hydraulic tomography approach to characterize main flowpaths and their connectivity in fractured media. *Advances in Water Resources*, *136*, 103500.
[80]. Kombo, O. H., Kumaran, S., Sheikh, Y. H., Bovim, A., & Jayavel, K. (2020). Long-term groundwater level prediction model based on hybrid KNN-RF technique. *Hydrology*, *7*(3), 59.
[81]. Kuhn, M., & Johnson, K. (2019). *Feature engineering and selection: A practical approach for predictive models*. Chapman and Hall/CRC.
[82]. Kyaw, A. C., Nagengast, N., Usma-Mansfield, C., & Fuss, F. K. (2023). A combined reverse engineering and multi-criteria decision-making approach for remanufacturing a classic car part. *Procedia CIRP*, *119*, 222-228.
[83]. Lee, J., Jin, C., & Liu, Z. (2017). Predictive big data analytics and cyber physical systems for TES systems. In *Advances in Through-life Engineering Services* (pp. 97-112). Springer.
[84]. Lee, Y., Ragguett, R.-M., Mansur, R. B., Boutilier, J. J., Rosenblat, J. D., Trevizol, A., Brietzke, E., Lin, K., Pan, Z., & Subramaniapillai, M. (2018). Applications of machine learning algorithms to predict therapeutic outcomes in depression: a meta-analysis and systematic review. *Journal of affective disorders*, *241*, 519-532.
[85]. Liu, D., Li, C., & Malik, O. (2021). Nonlinear modeling and multi-scale damping characteristics of hydro-turbine regulation systems under complex variable hydraulic and electrical network structures. *Applied energy*, *293*, 116949.
[86]. Lloyd Owen, D. A. (2018). Smart water and water megatrend management and mitigation. *Assessing Global Water Megatrends*, 87-104.
[87]. Long, W., Lu, Z., & Cui, L. (2019). Deep learning-based feature engineering for stock price movement prediction. *Knowledge-Based Systems*, *164*, 163-173.
[88]. Magini, R., Moretti, M., Boniforti, M. A., & Guercio, R. (2023). A machine-learning approach for monitoring water distribution networks (WDNs). *Sustainability*, *15*(4), 2981.
[89]. Magzamen, S., Mayer, A. P., Barr, S., Bohren, L., Dunbar, B., Manning, D., Reynolds, S. J., Schaeffer, J. W., Suter, J., & Cross, J. E. (2017). A multidisciplinary research framework on green schools: Infrastructure, social environment, occupant health, and performance. *Journal of School Health*, *87*(5), 376-387.
[90]. Mahmoud, M. A., Md Nasir, N. R., Gurunathan, M., Raj, P., & Mostafa, S. A. (2021). The current state of the art in research on predictive maintenance in smart grid distribution network: Fault's types, causes, and prediction methods—A systematic review. *Energies*, *14*(16), 5078.
[91]. Mahmudul, H., & Sadia, Z. (2026). Safeguarding Financial Records: A Cybersecurity-Driven Management Framework. *International Journal of Scientific Interdisciplinary Research*, *7*(1), 493–526. https://doi.org/10.63125/zch4s169

[92]. Mashhadi, N., Shahrour, I., Attoue, N., El Khattabi, J., & Aljer, A. (2021). Use of machine learning for leak detection and localization in water distribution systems. *Smart Cities*, *4*(4), 1293-1315.

[93]. Md Arif Uz, Z., Sharmin, S., & Md Abdur, R. (2025). Factors Influencing Behavioural Intention to Use he Innovative Open And Distance Learning Systems With The Role Of Continuance Intention. *American Journal of Scholarly Research and Innovation*, *4*(01), 202-220. https://doi.org/10.63125/pbjxp014

[94]. Md Tohidul, I. (2023). ERP-Integrated Financial Analytics Systems for Corporate Financial Reporting, Cash Flow Analysis, and Strategic Investment Planning. *American Journal of Advanced Technology and Engineering Solutions*, *3*(04), 87-128. https://doi.org/10.63125/qhk96h35

[95]. Md. Abdur, R., & Iftekhar, A. (2021). Customer Retention Forecasting in Mobile Wallet Services Using Neural Networks: A Comparative Quantitative Study. *International Journal of Business and Economics Insights*, *1*(4), 70-102. https://doi.org/10.63125/dyrpc387

[96]. Modaresi, F., Araghinejad, S., & Ebrahimi, K. (2018). A comparative assessment of artificial neural network, generalized regression neural network, least-square support vector regression, and K-nearest neighbor regression for monthly streamflow forecasting in linear and nonlinear conditions. *Water Resources Management*, *32*(1), 243-258.

[97]. Mohammad Badrul, A., & Md. Mominul, H. (2023). Machine Learning–Based Assessment of Environmental and Public Health Risks in Sewerage Sludge Management Systems. *American Journal of Advanced Technology and Engineering Solutions*, *3*(02), 33-77. https://doi.org/10.63125/y3yfxa92

[98]. Mohammad, N., & Shahjahan, M. (2019). Design and Development of IoT based Supervisory Control and Data Acquisition System. 2019 5th International Conference on Advances in Electrical Engineering (ICAEE),

[99]. Mohammad Shoriful Hossan, S. (2024). Development of A Predictive Maintenance Model for Industrial Motor Drives and VFD Systems Using PLC/DCS Operational Telemetry. *American Journal of Scholarly Research and Innovation*, *3*(01), 198-242. https://doi.org/10.63125/pv1js496

[100]. Mohammad Shoriful Hossan, S. (2025). Development Of an ARC-Flash Hazard Mitigation Framework for Brownfield Industrial Power Distribution Retrofits in U.S. Manufacturing. *American Journal of Advanced Technology and Engineering Solutions*, *1*(02), 262-303. https://doi.org/10.63125/vhqjgm79

[101]. Mohammad Shoriful Hossan, S., & Adar, C. (2023). SIL-Compliant Safety Instrumented System Design for Oil & Gas Hazardous-Area Operations Using IEC 61508 / IEC 61511 Frameworks. *Journal of Sustainable Development and Policy*, *2*(02), 43-85. https://doi.org/10.63125/cb3d3w81

[102]. Mst Shurovi, A. (2026). AI-Driven Accounts Payable and Receivable Automation for Operational Risk Mitigation in U.S. SMEs: A Systematic Review (2018–2026). *American Journal of Interdisciplinary Studies*, *7*(01), 536-577. https://doi.org/10.63125/x1yzwc45

[103]. Mst Shurovi, A., & Samia Hossain, S. (2022). Difference-in-Differences (DiD) Analysis of Data Driven Accounting Systems and its Impact on Financial Error Rates in SMEs. *Journal of Sustainable Development and Policy*, *1*(04), 117-157. https://doi.org/10.63125/n4nr0d75

[104]. Mubvaruri, F., Hoko, Z., Mhizha, A., & Gumindoga, W. (2022). Investigating trends and components of non-revenue water for Glendale, Zimbabwe. *Physics and Chemistry of the Earth, Parts a/B/C*, *126*, 103145.

[105]. Muhammad Mohiul, I., & Mohammad Badrul, A. (2022). Predictive Analytics for Optimizing Sewerage Sludge Treatment Efficiency and Resource Recovery in Dhaka City. *Journal of Sustainable Development and Policy*, *1*(04), 158-200. https://doi.org/10.63125/t9zdp458

[106]. Mukut Kanti, B. (2025). Lifecycle Carbon Reduction and Net Zero Pathway Modeling Using LEED, EDGE, and SBTI Frameworks in Developing Economy Industrial Buildings. *American Journal of Interdisciplinary Studies*, *6*(3), 306-350. https://doi.org/10.63125/tnneth75

[107]. Mukut Kanti, B., & Rony, S. (2022). Quantitative Risk Assessment and Safety Performance Benchmarking for Structural Integrity Management in Retail Supply Chain Infrastructure. *American Journal of Data Science and Analytics*, *3*(12), 42-88. https://doi.org/10.63125/n88vth90

[108]. Naghibi, S. A., Ahmadi, K., & Daneshi, A. (2017). Application of support vector machine, random forest, and genetic algorithm optimized random forest models in groundwater potential mapping. *Water Resources Management*, *31*(9), 2761-2775.

[109]. Nagpal, K., Foote, D., Liu, Y., Chen, P.-H. C., Wulczyn, E., Tan, F., Olson, N., Smith, J. L., Mohtashamian, A., & Wren, J. H. (2019). Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer. *NPJ digital medicine*, *2*(1), 48.

[110]. Nancy, A. A., Ravindran, D., Raj Vincent, P. D., Srinivasan, K., & Gutierrez Reina, D. (2022). Iot-cloud-based smart healthcare monitoring system for heart disease prediction via deep learning. *Electronics*, *11*(15), 2292.

[111]. Narwaria, M. (2018). Toward better statistical validation of machine learning-based multimedia quality estimators. *IEEE Transactions on Broadcasting*, *64*(2), 446-460.

[112]. Nhu, V.-H., Zandi, D., Shahabi, H., Chapi, K., Shirzadi, A., Al-Ansari, N., Singh, S. K., Dou, J., & Nguyen, H. (2020). Comparison of support vector machine, Bayesian logistic regression, and alternating decision tree algorithms for shallow landslide susceptibility mapping along a mountainous road in the west of Iran. *Applied Sciences*, *10*(15), 5047.

[113]. Nie, X., Fan, T., Wang, B., Li, Z., Shankar, A., & Manickam, A. (2020). Big data analytics and IoT in operation safety management in under water management. *Computer Communications*, *154*, 188-196.

[114]. Novak, P., Guinot, V., Jeffrey, A., & Reeve, D. E. (2018). *Hydraulic modelling: An introduction: Principles, methods and applications*. CRC Press.

[115]. Pang, Q., Lou, D., Li, S., Wang, G., Qiao, B., Dong, S., Ma, L., Gao, C., & Wu, Z. (2020). Smart flexible electronics-integrated wound dressing for real-time monitoring and on-demand treatment of infected wounds. *Advanced Science*, *7*(6), 1902673.

[116]. Pedersen, A. N., Borup, M., Brink-Kjær, A., Christiansen, L. E., & Mikkelsen, P. S. (2021). Living and prototyping digital twins for urban water systems: Towards multi-purpose value creation using models and sensors. *Water*, *13*(5), 592.

[117]. Pinto, F. S., Simões, P., & Marques, R. C. (2017). Water services performance: do operational environment and quality factors count? *Urban Water Journal*, *14*(8), 773-781.

[118]. Popov, A. (2023). Feature engineering methods. In *Advanced methods in biomedical signal processing and analysis* (pp. 1-29). Elsevier.

[119]. Punmiya, R., & Choe, S. (2019). Energy theft detection using gradient boosting theft detector with feature engineering-based preprocessing. *IEEE Transactions on Smart Grid*, *10*(2), 2326-2329.

[120]. Raghunandan, K. (2022). Supervisory control and data acquisition (SCADA). In *Introduction to Wireless Communications and Networks: A Practical Perspective* (pp. 321-337). Springer.

[121]. Rai, R., & Sahu, C. K. (2020). Driven by data or derived through physics? a review of hybrid physics guided machine learning techniques with cyber-physical system (cps) focus. *IEEE access*, *8*, 71050-71073.

[122]. Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific reports*, *14*(1), 6086.

[123]. Ramos, H. M., Kuriqi, A., Besharat, M., Creaco, E., Tasca, E., Coronado-Hernández, O. E., Pienika, R., & Iglesias-Rey, P. (2023). Smart water grids and digital twin for the management of system efficiency in water distribution networks. *Water*, *15*(6), 1129.

[124]. Ramos, H. M., Morani, M. C., Carravetta, A., Fecarrotta, O., Adeyeye, K., López-Jiménez, P. A., & Pérez-Sánchez, M. (2022). New challenges towards smart systems' efficiency by digital twin in water distribution networks. *Water*, *14*(8), 1304.

[125]. Ramotsoela, D. T., Hancke, G. P., & Abu-Mahfouz, A. M. (2019). Attack detection in water distribution systems using machine learning. *Human-centric Computing and Information Sciences*, *9*(1), 13.

[126]. Rediana, R., & Pharmasetiawan, B. (2017). Designing a business model for smart water management system with the smart metering system as a core technology: Case study: Indonesian drinking water utilities. 2017 International Conference on ICT For Smart Society (ICISS),

[127]. Renteria-Mena, J. B., & Giraldo, E. (2024). Predictive modeling of water level in the san juan river using hybrid neural networks integrated with kalman smoothing methods. *Information*, *15*(12), 754.

[128]. Risha, A., & Kazi Mohammad Khalid, A. (2023). A Meta-Analysis of AI-Driven Geospatial Analytics for Predictive Maintenance of Critical Infrastructure in Developing Economies. *International Journal of Scientific Interdisciplinary Research*, *4*(4), 375–412. https://doi.org/10.63125/rayrex49

[129]. RM, S. P., Maddikunta, P. K. R., Koppu, S., Gadekallu, T. R., Chowdhary, C. L., & Alazab, M. (2020). An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection in IoMT architecture. *Computer Communications*, *160*, 139-149.

[130]. Saddiqi, H. A., Javed, Z., Ali, Q. M., Ullah, A., & Ahmad, I. (2024). Modelling and predicting lift force and trans-membrane pressure using linear, KNN, ANN and response surface models during the separation of oil drops from produced water. *Journal of Water Process Engineering*, *66*, 106014.

[131]. Sadia, Z., Mohammad Ariful, I., Sumyta, H., & Hosne Ara, M. (2022). A Review of AI-Powered Data Visualization in Enterprise Reporting: Dashboard Design And Interactive Analytics. *American Journal of Advanced Technology and Engineering Solutions*, *2*(01), 32-54. https://doi.org/10.63125/gabst658

[132]. Samia Hossain, S. (2024). Predictive Financial Volatility Analytics Using Machine Learning Pipelines in ERP-Integrated Enterprises. *American Journal of Interdisciplinary Studies*, *5*(04), 134-181. https://doi.org/10.63125/hz27wz06

[133]. Sánchez, A. S., Oliveira-Esquerre, K. P., dos Reis Nogueira, I. B., de Jong, P., & Filho, A. A. (2020). Water loss management through smart water systems. In *Smart village technology: concepts and developments* (pp. 233-266). Springer.

[134]. Santos, E. (2024). Beyond leakage: non-revenue water loss and economic sustainability. *Urban Science*, *8*(4), 194.

[135]. Saravanan, K., Anusuya, E., Kumar, R., & Son, L. H. (2018). Real-time water quality monitoring using Internet of Things in SCADA. *Environmental monitoring and assessment*, *190*(9), 556.

[136]. Schopper, F., Doetsch, J., Villiger, L., Krietsch, H., Gischig, V. S., Jalali, M., Amann, F., Dutler, N., & Maurer, H. (2020). On the variability of pressure propagation during hydraulic stimulation based on seismic velocity observations. *Journal of Geophysical Research: Solid Earth*, *125*(2), e2019JB018801.

[137]. Sharma, P., Gounder, M., & Kumar, K. (2024). Mastering Feature Engineering: Unlocking the Art of Data Transformation for Enhanced Predictive Modeling with Neural Networks. International Conference on Data Science and Applications,

[138]. Sharma, S., Sharma, K., & Grover, S. (2024). Real-time data analysis with smart sensors. In *Application of artificial intelligence in wastewater treatment* (pp. 127-153). Springer.

[139]. Sheykhmousa, M., Mahdianpari, M., Ghanbari, H., Mohammadimanesh, F., Ghamisi, P., & Homayouni, S. (2020). Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *13*, 6308-6325.

[140]. Shima Ali, S., Musomi, K., & Sonya, G. (2024). Integrated Infectious Disease Preparedness: Multilingual Health Communication Strategies for Public Health Risk Mitigation, Disease Intervention and Prevention for Culturally Diverse Population in United States. *American Journal of Data Science and Analytics*, *5*(10), 01-47. https://doi.org/10.63125/ehk4mt02

[141]. Shima Ali, S., Sonya, G., & Mahbubul, I. (2025). Telemedicine And Remote Patient Monitoring for Preventing Risk and Chronic Disease Management in Underserved Communities. *American Journal of Health and Medical Sciences*, *6*(04), 01–42. https://doi.org/10.63125/y0e9dc38
[142]. Shima Ali, S., Sonya, G., & Mahbubul, I. (2026). STI and HIV Prevention Through Public Health Digital Twins: A Framework for Personalized Prevention and Adaptive Disease Intervention. *American Journal of Scholarly Research and Innovation*, *5*(01), 284–326. https://doi.org/10.63125/rjgna882
[143]. Shuang, Q., Liu, H. J., & Porse, E. (2019). Review of the quantitative resilience methods in water distribution networks. *Water*, *11*(6), 1189.
[144]. Shushu, U. P., Komakech, H. C., Dodoo-Arhin, D., Ferras, D., & Kansal, M. L. (2021). Managing non-revenue water in Mwanza, Tanzania: a fast-growing sub-Saharan African city. *Scientific African*, *12*, e00830.
[145]. Silverii, F., D'Agostino, N., Borsa, A. A., Calcaterra, S., Gambino, P., Giuliani, R., & Mattone, M. (2019). Transient crustal deformation from karst aquifers hydrology in the Apennines (Italy). *Earth and Planetary Science Letters*, *506*, 23-37.
[146]. Şişman, E., & Kızılöz, B. (2020). Trend-risk model for predicting non-revenue water: an application in Turkey. *Utilities Policy*, *67*, 101137.
[147]. Song, L., Wang, L., Wu, J., Liang, J., & Liu, Z. (2021). Integrating physics and data driven cyber-physical system for condition monitoring of critical transmission components in smart production line. *Applied Sciences*, *11*(19), 8967.
[148]. Syed Nurul, I. (2026). A Machine Learning Approach to Intelligent Streetlight Control for Sustainable Energy Management in Smart Cities. *American Journal of Advanced Technology and Engineering Solutions*, *6*(01), 650-689. https://doi.org/10.63125/gxwn1g63
[149]. Taherdoost, H., & Madanchian, M. (2023). Multi-criteria decision making (MCDM) methods and concepts. *Encyclopedia*, *3*(1), 77-87.
[150]. Thakur, S. (2024). Based on digital twin technology, an early warning system and strategy for predicting urban waterlogging. *Simulation Techniques of Digital Twin in Real-Time Applications: Design Modeling and Implementation*, 301-318.
[151]. Thompson, K., Skeens, B., & Liggett, J. (2017). Nonrevenue water. *Internet of things and data analytics handbook*, 467-480.
[152]. Thuraka, B., Pasupuleti, V., Kodete, C. S., Naidu, U. G., Tirumanadham, N. K. M. K., & Shariff, V. (2024). Enhancing predictive model performance through comprehensive pre-processing and hybrid feature selection: A study using SVM. 2024 2nd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS),
[153]. Toubeau, J. F., Iassinovski, S., Jean, E., Parfait, J. Y., Bottieau, J., De Grève, Z., & Vallée, F. (2019). Non-linear hybrid approach for the scheduling of merchant underground pumped hydro energy storage. *IET Generation, Transmission & Distribution*, *13*(21), 4798-4808.
[154]. Urbanowicz, K., Stosiak, M., Towarnicki, K., & Bergant, A. (2021). Theoretical and experimental investigations of transient flow in oil-hydraulic small-diameter pipe system. *Engineering Failure Analysis*, *128*, 105607.
[155]. Varoquaux, G., & Colliot, O. (2023). Evaluating machine learning models and their diagnostic value. *Machine learning for brain disorders*, 601-630.
[156]. Velayudhan, N. K., Pradeep, P., Rao, S. N., Devidas, A. R., & Ramesh, M. V. (2022). IoT-enabled water distribution systems—A comparative technological review. *IEEE access*, *10*, 101042-101070.
[157]. Veriava, A. (2019). Non-revenue water and non-revenue life: A reflection on the making and mitigating of water losses in Johannesburg. *African Studies*, *78*(4), 590-608.
[158]. Villegas-Ch, W., García-Ortiz, J., & Sánchez-Viteri, S. (2024). Toward intelligent monitoring in IoT: AI applications for real-time analysis and prediction. *IEEE access*, *12*, 40368-40386.
[159]. Vita, V., Fotis, G., Chobanov, V., Pavlatos, C., & Mladenov, V. (2023). Predictive maintenance for distribution system operators in increasing transformers' reliability. *Electronics*, *12*(6), 1356.
[160]. Wacker, S., & Noskov, S. Y. (2018). Performance of machine learning algorithms for qualitative and quantitative prediction drug blockade of hERG1 channel. *Computational Toxicology*, *6*, 55-63.
[161]. Xia, B., Lyu, D., Wang, N., Niu, K., Long, G., Shui, B., Xiang, Y., & Liao, Z. (2023). Development of a rapid and quantitative prediction model for assessing the leakage status of water pipeline systems. *Journal of Civil Structural Health Monitoring*, *13*(2), 605-613.
[162]. Xie, T., Chen, L., Yi, B., Li, S., Leng, Z., Gan, X., & Mei, Z. (2024). Application of the improved K-nearest neighbor-based multi-model ensemble method for runoff prediction. *Water*, *16*(1), 69.
[163]. Yacchirema, D. C., Sarabia-Jácome, D., Palau, C. E., & Esteve, M. (2018). A smart system for sleep monitoring by integrating IoT with big data analytics. *IEEE access*, *6*, 35988-36001.
[164]. Yang, H.-Q., Yan, Y., Wei, X., Shen, Z., & Chen, X. (2023). Probabilistic analysis of highly nonlinear models by adaptive sparse polynomial chaos: Transient infiltration in unsaturated soil. *International Journal of Computational Methods*, *20*(08), 2350006.
[165]. Yang, W., & Zhang, J. (2021). Assessing the performance of gray and green strategies for sustainable urban drainage system development: A multi-criteria decision-making analysis. *Journal of Cleaner Production*, *293*, 126191.
[166]. Yang, Y., Li, X., Yang, Z., Wei, Q., Wang, N., & Wang, L. (2018). The application of cyber physical system for thermal power plants: Data-driven modeling. *Energies*, *11*(4), 690.
[167]. Yao, J. (2018). Model-based nonlinear control of hydraulic servo systems: Challenges, developments and perspectives. *Frontiers of Mechanical Engineering*, *13*(2), 179-210.

[168]. Yin, S., Rodriguez-Andina, J. J., & Jiang, Y. (2019). Real-time monitoring and control of industrial cyberphysical systems: With integrated plant-wide monitoring and control framework. *IEEE Industrial Electronics Magazine*, *13*(4), 38-47.

[169]. Young, B. A., Hall, A., Pilon, L., Gupta, P., & Sant, G. (2019). Can the compressive strength of concrete be estimated from knowledge of the mixture proportions?: New insights from statistical analysis and machine learning methods. *Cement and concrete research*, *115*, 379-388.

[170]. Yuan, Y., Tang, X., Zhou, W., Pan, W., Li, X., Zhang, H.-T., Ding, H., & Goncalves, J. (2019). Data driven discovery of cyber physical systems. *Nature communications*, *10*(1), 4894.

[171]. Yun, K. K., Yoon, S. W., & Won, D. (2021). Prediction of stock price direction using a hybrid GA-XGBoost algorithm with a three-stage feature engineering process. *Expert Systems with Applications*, *186*, 115716.

[172]. Zaman, S. (2024). A Systematic Review of ERP And CRM Integration For Sustainable Business And Data Management in Logistics And Supply Chain Industry. *Frontiers in Applied Engineering and Technology*, *1*(01), 204-221. https://doi.org/10.70937/faet.v1i01.36

[173]. Zhong, S., Zhang, K., Bagheri, M., Burken, J. G., Gu, A., Li, B., Ma, X., Marrone, B. L., Ren, Z. J., & Schrier, J. (2021). Machine learning: new ideas and tools in environmental science and engineering. *Environmental science & technology*, *55*(19), 12741-12754.

[174]. Zhuang, D., Gan, V. J., Tekler, Z. D., Chong, A., Tian, S., & Shi, X. (2023). Data-driven predictive control for smart HVAC system in IoT-integrated buildings with time-series forecasting and reinforcement learning. *Applied energy*, *338*, 120936.