

Development of Software and Hardware-Based Methods for the Detection of Leaks and Bursts in the Water Supply Systems for Developing an Intelligent Water Network

Kiran Joseph

Bachelor of Technology (Electrical and Electronics Engineering)

Master of Technology (Energy Systems)

**Thesis submitted for the fulfilment of the requirements for the degree of
Doctor of Philosophy (PhD)**

Victoria University, Australia

Institute for Sustainable Industries & Liveable Cities (ISILC)

November 2025

Abstract

Centralised water supply systems have been utilised globally for more than a century; however, ageing infrastructure, population growth, climate variability, and increasing urbanisation have intensified operational pressures on water utilities. Globally, Non-Revenue Water (NRW) primarily due to physical losses from leaks and bursts is estimated to account for approximately 20–30% of total water supplied in many systems, with significantly higher losses reported in ageing networks. These losses result in substantial economic costs associated with water production, energy consumption, infrastructure repair, and service disruptions, while also contributing to environmental degradation and reduced water security. As of late 2024 and early 2025, global NRW is estimated at 346 million cubic meters per day, equivalent to a loss of approximately USD \$39 billion annually. This research was undertaken to develop and validate software- and hardware-based methodologies for leak and burst detection as part of a broader framework for establishing an Intelligent Water Network (IWN).

This thesis integrates two complementary studies to develop a comprehensive, scalable, and field-validated framework for intelligent monitoring of pressurised water distribution systems. The first study focuses on real-time leak and burst detection using operational Supervisory Control and Data Acquisition (SCADA) data from an Australian pumping main and controlled measurements from a purpose-built experimental testbed. The second study proposes an optimisation-based methodology for identifying the most hydraulically informative sensor locations within a large-scale international water network. Together, these studies establish a unified technical foundation to support the transition toward future IWNs.

The leak and burst detection study was conducted on the 6.24-km, 450-mm Sunbury Pumping Main in Victoria, Australia, which conveys water from the Loemans Road Pump Station to the Shepherds Lane Tank. The system operates under variable pump-speed control, with SCADA continuously logging pressure, flow, and pump-speed measurements. A deterministic hydraulic logic model was developed to analyse pressure, flow, and pump-speed behaviour across start-up, steady-state, and shutdown conditions. Adaptive pressure thresholds were applied to distinguish

minor leaks, major leaks, and bursts, while the Darcy–Weisbach formulation enabled accurate localisation of burst events along the pipeline alignment.

Machine Learning analysis was also applied to the Sunbury Pumping Main to identify abnormal hydraulic behaviour in the absence of labelled leak events. Five unsupervised anomaly-detection algorithms including Local Outlier Factor (LOF), Isolation Forest, One-Class Support Vector Machine (One-Class SVM), Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and the K-Means Clustering Algorithm (K-Means) were trained on normalised SCADA telemetry. LOF demonstrated the highest anomaly-detection performance after class balancing, achieving an F1-score of approximately 0.69 and a Receiver Operating Characteristic – Area Under the Curve (ROC–AUC) of 0.72, owing to its ability to capture subtle, localised deviations in pressure, flow, and pump-speed interactions. The final LOF model was integrated into a real-time web-based monitoring platform for automated leak and burst alerting.

To provide controlled experimental validation, a 111-m long, 110-mm polyethylene (PE) prototype pipeline was constructed in Melton, Victoria, connected to an existing 100-mm supply line. Two flow meters and two pressure sensors were installed at hydraulically informative locations. Controlled leak scenarios were introduced using 20-mm and 25-mm lateral outlets, representing realistic small-to-moderate leak conditions commonly associated with service connections, fittings, and joint failures in distribution systems. Fourteen supervised Machine Learning classifiers were trained on high-resolution hydraulic datasets representing no-leak, minor-leak, and major-leak scenarios. Decision Tree, k-Nearest Neighbours (KNN), and Random Forest algorithms achieved perfect classification accuracy and F1-scores, demonstrating excellent capability under controlled conditions.

The second investigation develops a robust Optimal Sensor Placement (OSP) framework applied to the Hanoi Water Distribution Network (WDN), a widely used benchmark model in hydraulic research. Unlike the Sunbury trunk main and Melton prototype system, the Hanoi network represents a meshed distribution system, enabling evaluation of sensor placement strategies under complex nodal interactions. Leak scenarios were simulated using EPANET by introducing emitter-based leak representations at 32 nodes with equivalent orifice diameters of 10, 50, 100, and 150 mm. These diameters represent a spectrum of leak severities, from small joint-related losses to larger structural failures. Pressure deviations at all nodes were filtered using a 0.5-m sensitivity

threshold to construct a hydraulic sensitivity matrix representing nodal responsiveness to leaks. Four optimisation algorithms including Greedy, Sensitivity-based Top-K, D-Optimal, and a Hybrid Greedy→D-Optimal strategy were benchmarked against a Random baseline. The deterministic algorithms achieved consistently high (>96%) to perfect detection coverage, with the Hybrid approach providing the best balance between detection performance, redundancy reduction, and sensor stability. Sensitivity-evolution analysis identified Nodes 29, 28, 27, 15, and 30 as the most consistently informative locations across all leak magnitudes. Additional robustness, runtime, and threshold-perturbation analyses confirmed the scalability and reliability of the proposed OSP framework for utility-scale implementation. Collectively, the integrated methodologies improve fault detectability, reduce false alarms, enhance localisation accuracy, and optimise sensor deployment strategies across both trunk main and distribution network contexts. The outcomes offer significant benefits to water utilities through reduced non-revenue water, lower operational costs, improved infrastructure resilience, and enhanced service reliability for communities. Environmentally, the framework supports sustainable water-resource management by minimising unnecessary abstraction and treatment energy. From a research perspective, this thesis advances the integration of hydraulic theory with data-driven analytics, contributing a scalable and practically deployable foundation for next-generation Intelligent Water Networks.

Declaration of Authenticity

I, Kiran Joseph, declare that the Doctor of Philosophy (PhD) thesis entitled “Development of Software and Hardware-Based Methods for the Detection of Leaks and Bursts in the Water Supply Systems for Developing an Intelligent Water Network” is no more than 80,000 words in length, including quotes and exclusive of tables, Figures, appendices, bibliography, references, and footnotes. This thesis contains no material that has been submitted previously, in whole or in part, for the award of any other academic degree or diploma. Except where otherwise indicated, this thesis is my own work.

This thesis includes material that has been published in four peer-reviewed journal articles based on this research study. The content included in the thesis is consistent with the published versions of these manuscripts.

“I have conducted my research in alignment with the [Australian Code for the Responsible Conduct of Research](#) and [Victoria University’s Higher Degree by Research Policy and Procedures](#).”

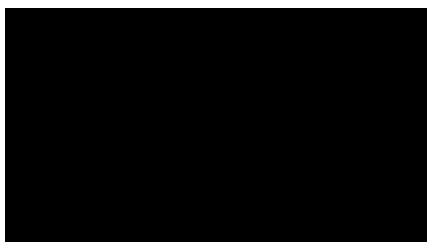
Generative Artificial Intelligence (AI) Declaration

This thesis has been edited for clarity of expression and proofreading using the following AI-assisted tools using Grammarly for grammar and style editing; ProWritingAid for style, consistency, and structure checks; This use complies with Victoria University guidelines on the use of AI in HDR theses.

Funding Declaration

This research was supported by Greater Western Water (GWW) and Victoria University, whose financial assistance and collaboration are gratefully acknowledged.

Signature:



Date: 27.11.2025

Acknowledgements

I would like to express my deepest gratitude to my Principal Supervisor, Dr. Ashok K Sharma, for his invaluable guidance, continuous inspiration, and tireless support, throughout my PhD journey at Victoria University. I am grateful for his trust in my ability and willingness to accept me to be his research student. Dr. Ashok K Sharma not only provided me with vital advice for my academic and career pursuits, but he also shared his in-depth knowledge and experiences, which have significantly contributed to my research and ensured my success.

I would also like to express my appreciation to my associate supervisors, Dr Rudi van Staden, Dr Wasantha P Liyanage, and Dr Jyoti Shetty for their ongoing support and valuable contributions to this research. Their expertise and advice were crucial to the successful completion of the thesis. Their insightful comments and constructive suggestions were crucial in refining my writing of the journal and conference papers on which this thesis is based. I am profoundly grateful to have had the fortune to work with such great supervisors, without their exceptional supervision, guidance and motivation, my PhD thesis would not have been completed in a timely manner.

In addition to my supervisors, I would like to express my gratitude and sincere appreciation to my mentors, Dr. Rajesh Gupta, Dr. Nikita Palod, Dr. Marcelle Cacciattolo, Dr. Paul Joseph for their time and tireless efforts to provide valuable comments and insights on both my research works and personal developments, with their sincere attitude and generous suggestions that guided me to a successful path on my PhD research journey as well as my academic career development at Victoria University.

I gratefully acknowledge the contributions of Greater Western Water staff and engineers, particularly Mr. Jason Cotton, Mr. Grant Powell, Ms. Sharna Small, and Mr. Nathan Benett, for their support with data collection, review, and feedback that strengthened my research outcomes. I also extend my appreciation to the staff of the Institute for Sustainable Industries and Liveable Cities (ISILC) at Victoria University for their research funding support and assistance.

On a personal note, I would also like to thank my fellow colleagues and mentors in the bachelor's, postgraduate, and research programs for sharing their journey and keeping me motivated. My

special thanks to all my friends from other streams who have continuously supported and inspired me in many ways that encouraged me to keep moving forward and made my PhD journey more enjoyable and bearable. Their presence turned struggles into shared strength and reminded me that I was never alone. Above all, my deepest and eternal gratitude goes to my family. Their unconditional love, countless sacrifices, and unwavering belief in me have been the true foundation of this achievement. In times of challenge, their support became my greatest strength, and their encouragement gave me the resolve to keep moving forward. This thesis is not only the culmination of my own efforts but also a testament to their enduring love and dedication. With all my heart, I dedicate this work to them.

List of Publications

1. Joseph, K, Sharma, AK & van Staden, R 2022, 'Development of an Intelligent Urban Water Network System', *MDPI Water*, vol. 14, no. 9, pp. 1-17.
2. Joseph, K, Sharma, AK, van Staden, R, Wasantha, P, Cotton, J & Small, S 2023, 'Application of Software and Hardware-Based Technologies in Leaks and Burst Detection in Water Pipe Networks: A Literature Review', *MDPI Water*, vol. 15, no. 11, p. 2046.
3. Joseph, K, Shetty, J, Patnaik, R, Matthew, NS, Van Staden, R, Liyanage, WP, Powell, G, Bennett, N & Sharma, AK 2025, 'Early Leak and Burst Detection in Water Pipeline Networks Using Machine Learning Approaches', *Journal of Water*, vol. 17, no. 14, p. 2164.
4. Joseph, K, Shetty, J, Sharma, AK, van Staden, R, Wasantha, P, Small, S & Bennett, N 2024, 'Leak and Burst Detection in Water Distribution Network Using Logic-and Machine Learning-Based Approaches', *Journal of Water*, vol. 16, no. 14, p. 1935.

List of Peer Review Contributions

1. SPRINGER NATURE – Signal, Image and Video Processing (2025): Reviewed the manuscript “*Advancing Water Quality Analysis Through Deep Learning Techniques*”, reflecting recognition of my expertise in deep learning, Machine Learning, and signal/image processing for engineering applications.
2. TAYLOR & FRANCIS – Urban Water Journal (2025): Reviewed the manuscript “*Detecting Small-Size Leaks in Water Distribution Pipelines: Optimizing Leak Localization and Mitigation Strategies for Urban Water Systems Using a Dense Hebbian Neural Network Approach*”, demonstrating my domain expertise in Intelligent Water Networks, leakage detection, hydraulic modelling, and Machine Learning-based condition assessment.
3. SPRINGER NATURE – Archives of Computational Methods in Engineering (2026): Reviewed the manuscript “*Toward Intelligent Water Systems: A Systematic Literature Review of Artificial Intelligence-Based Models for Water Quality Monitoring*”, demonstrating my expertise in intelligent water systems, AI-driven water quality monitoring, Machine Learning and deep learning applications in environmental engineering, IoT-based sensing technologies, and systematic literature review methodologies.

Conference Presentations

1. Joseph, K 2025, *Water distribution network management: identification of leaks and bursts for water loss reduction*, paper presented at the 12th IWA International Water Efficiency Conference – Efficient 2025, Victoria University, Melbourne, 16–18 September.
2. Joseph, K & Sharma, A 2023, *Intelligent method for the detection of leaks & bursts in water network system: Australian case study*, poster presented at the World Water Management and Development Conference (WWMDC 2023), Murdoch University, Perth, Western Australia, 3–8 December.
3. Joseph, K 2023, *Themed presentation – Engineering & Science*, ISILC HDR Symposium 2023, Victoria University, Melbourne, October. *KIRAN JOSEPH HDR Symposium 2023*.
4. Joseph, K 2021, *Stream presentation – Engineering & Science*, HDR Student Conference 2021, Victoria University, Melbourne, November. *KIRAN JOSEPH Conference 2021*.

Table of Contents

Table of Contents

Abstract	2
Declaration of Authenticity.....	5
Acknowledgements	6
List of Publications.....	8
List of Peer Review Contributions	9
Conference Presentations	10
Chapter 1	14
1.1 General Remarks	14
1.2 Background and context of the study	15
1.3 Aims of the Study	17
1.4 Scope of the study.....	18
1.5 Research Significance and Innovations	23
1.6 Chapter Summary and Reflections	25
Chapter 2.....	27
2.1 General Remarks	27
2.2 Introduction	27
2.3 Conventional Methods of Leak Detection	31
2.4 Hardware-Based Leak and Burst Detection Methods	35
2.5 Software-Based Leak and Burst Detection Methods.....	39
2.6 Optimal Allocation of Pressure Sensors	49
2.7 Chapter Summary and Reflections	52
Chapter 3	56
3.1 General Remarks	56
3.2 Pumping main system.....	56
3.3 Prototype water supply system at Melton Training Yard.....	59
3.4 Hanoi Water Distribution System as case study to identify preferred locations of pressure sensors	63
3.5 Chapter Summary and Reflections	67
Chapter 4	68

4.1 General Remarks	68
4.2 Leak and Burst Detection Methodology	69
4.3 Framework for Intelligent Leak and Burst Detection	97
4.4 Chapter Summary and Reflections.....	101
Chapter 5	103
5.1 General Remarks	103
5.2 Sunbury Pumping System Data Collection.....	104
5.3 Data Collection at Melton Prototype System.....	118
5.4 Key benefits of collected datasets (Sunbury and Melton).....	124
5.5 Chapter Summary and Reflections	125
Chapter 6	127
6.1 General Remarks	127
6.2 Application of logic-based methodology for leak and burst identification in pumping system.....	128
6.3 Sunbury Case Study: Software and hardware-based methods for pumping main.....	137
6.4 Melton Case Study: Software and Hardware-based method for prototype system.....	159
6.5 Chapter Summary and Reflections.....	179
Chapter 7	181
7.1 General Remarks	181
7.2 Research Methodology	182
7.3 Application of Methodology and Results.....	193
7.4 Chapter Summary and Reflections.....	232
Chapter 8	234
8.1 General Remarks	234
8.2 Evaluation of the Logic-Based Model.....	235
8.3 Evaluation of Supervised and Unsupervised Learning Models	236
8.4 Comparative Analysis of Detection Methods	237
8.5 Real-Time Alert Generation and Operational Performance	239
8.6 Lessons learned and key insights	239
8.7 Chapter Summary and Reflections.....	240
Chapter 9	242
9.1 General Remarks	242
9.2 Context and Motivation.....	243
9.3 Chapter Summary and Reflections.....	249
References	251

Chapter 10	264
10.1 Appendices for Chapter 1	264
10.2 Appendices for Chapter 2	265
10.3 Appendices for Chapter 3	270
10.4 Appendices for Chapter 4	272
10.5 Appendices for Chapter 5	275
10.6 Appendices for Chapter 6	279
10.7 Appendices for Chapter 7	287

Chapter 1

Introduction

1.1 General Remarks

Chapter 1 establishes the foundation of this research by introducing the concept of IWN and highlighting the growing need for intelligent monitoring and management systems within modern water distribution networks. The discussion outlines how urbanisation, population growth, and ageing infrastructure have amplified the challenges of leaks and bursts, which contribute significantly to non-revenue water, operational inefficiencies, and increased economic and environmental costs. The chapter defines the research context by linking hydraulic behaviour, sensor technologies, and artificial intelligence into a unified perspective that emphasises predictive and preventive management approaches.

The aims and objectives of the study were formulated to address critical challenges associated with water distribution systems, including high levels of non-revenue water (NRW), delayed detection of leaks and bursts, and limited real-time monitoring capability. Globally, NRW is estimated to account for approximately 20–30% of total system input volume, resulting in significant economic losses, increased operational costs, and reduced system efficiency (Liemberger & Wyatt, 2019). In addition, conventional leak detection approaches are often reactive, labour-intensive, and unable to provide continuous monitoring, leading to prolonged leak durations and increased infrastructure deterioration. These challenges highlight the need for integrated, data-driven monitoring frameworks that combine sensor technologies, SCADA-based data acquisition, and advanced analytical methods for real-time detection and management of leaks and bursts. So, the main advantage of hardware and software-based methods is to identify leaks and burst reasonability quickly thus minimising water losses from the system.

In summary, Chapter 1 provides a clear rationale and defines research boundaries while underscoring the novelty of integrating Machine Learning with SCADA-based sensing for real-time detection of leaks and bursts. The chapter positioned the study within the broader context of

sustainable water resource management, establishing the significance and innovation of this research for both academic and practical applications in the water industry.

1.2 Background and context of the study

In water networks, intelligence is defined as knowing and anticipating likely events or network behaviours before and after their occurrence, then setting up measures to mitigate potential outcomes or prevent their occurrence (Marney & Sharma, 2012). IWN collaborates with artificial intelligence, hydrologic and hydraulic engineering, and various sensor technologies to understand and solve water supply problems practically, logically, and analytically. It encompasses the ability to identify patterns, plan, invent, think abstractly, make decisions, memorise knowledge, and communicate through language. In general, an IWN is a simulation of human intelligence implemented on systems via reactive machine tools, memory capacity, theory of mind, and consciousness. It is programmed to learn, adapt, and solve or react to issues in the network layer of the water supply system's pipes, pumps, storage reservoirs, and valves (Greater Western Water, 2019). Water is a precious resource that is vital for sustaining life and supporting critical sectors, including agriculture, industry, urban infrastructure, and energy. However, one of the most pervasive and costly challenges faced by water utilities is the challenge of water leakages and bursts within municipal water distribution networks (Gas, M.P., 2021).

Leaks and bursts in municipal water distribution networks represent a major contributor to non-revenue water (NRW), which globally accounts for approximately 20–30% of total system input volume in many utilities (Liemberger & Wyatt, 2019). These losses can result in substantial economic impacts, with global estimates suggesting billions of cubic metres of treated water are lost annually, leading to significant financial burdens for water utilities and increased operational costs (Farley et al., 2008).

Leaks may occur as small, undetected losses over extended periods, often remaining unnoticed while gradually contributing to cumulative water loss and infrastructure deterioration (Puust et al., 2010). In contrast, pipe bursts are typically associated with sudden pipe failure, resulting in rapid water loss, pressure drops, and potential service disruptions to consumers. Depending on the location and severity of the burst, these events can partially or completely interrupt water supply, leading to economic, social, and environmental consequences (Romano et al., 2014).

Typical examples of a pipe leak and a burst are shown in Figure 1.1. Addressing this issue is not just an economic imperative but also a moral obligation, considering the growing global concern over water scarcity and the need to optimise energy consumption. In response to this critical challenge, the adoption of smart systems driven by logic-based and Machine Learning (ML) techniques has emerged as a transformative solution.



Figure 1.1: A typical (a) example of a pipe leakage (modified after R.S. Plumbing, 2024) and (b) example of a pipe burst (modified after Water Leak Detection, 2020).

Hydraulic characteristics of the pipeline, pressure sensors, and flow meters are integral components of a water network. This enables the service provider to plan for and mitigate some of the possible outcomes, or even prevent their occurrence (Gas, M.P., 2021). Consequently, in the case of IWNs, asset management procedures can be planned for these events and mitigate possible practical consequences or prevent them completely (Gas, M.P., 2021).

Information systems help fulfill the leak detection goal of policymakers (Stair & Reynolds, 2012). The application of emerging technology such as sensors, and decision support systems has major and transformative value when applied (Gourbesville, 2011). For instance, a 24/7 smart water-distribution network set up in Pimpri Chinchwad, India, suggested that detection of leaks would be a positive outcome, improved assessments of water quality, even decreased energy (Barate, 2017). Yet such intelligent technologies will not be in place to facilitate the effective running of pumping stations, distribution systems, and storage reservoirs of public water distribution systems unless comprehensive control and management systems run on intelligent computing technologies. Smart systems represent an integrated approach that combines sensor technologies, data analytics, automation, and system hydraulics to enhance the efficiency and sustainability of water distribution networks. This research explores the multifaceted issue of water leakages and bursts and the urgent need for smart systems to mitigate their impacts. It also delves into the application

of logic-based and ML techniques for early leak detection and precise identification, highlighting their pivotal role in optimising water and energy consumption while improving customer services.

1.3 Aims of the Study

The primary aim of this research is to develop a robust and efficient methodology to detect leaks and bursts in water pipeline networks by analysing SCADA data. This research will also encompass the optimal placement of smart pressure sensors within the network, ensuring comprehensive coverage and enhanced accuracy in monitoring. The study will integrate key parameters, including flow rate, pressure variations, pump speed, and timestamp data, to establish a detailed understanding of the pipeline dynamics under various operating conditions.

The specific objectives of this research are;

1. to thoroughly examine both leak and burst scenarios in order to identify unique patterns, thresholds, and indicators associated with these events in the water pipeline network, by exploring the influence of flow, pressure, pump speed, SCADA IoT integrated data;
2. to identify suitable locations for pressure sensor placement for leak detection;
3. to design, by leveraging logic-based analysis, advanced data analytics, Machine Learning techniques, and hydraulic modelling, a system that can predict and localise anomalies;
4. to contribute to reducing water loss, enhancing the efficiency of pipeline operations, and supporting sustainable water resource management.

This research is guided by the following key research questions:

1. How can SCADA data be effectively utilised to detect leaks and bursts in water distribution networks?
2. What hydraulic patterns and indicators distinguish leak and burst conditions from normal operation?

3. How can Machine Learning and logic-based approaches be integrated to improve the accuracy and reliability of anomaly detection?
4. What are the optimal locations for pressure sensor placement to maximise leak detection coverage and sensitivity?
5. How do different leak sizes influence hydraulic sensitivity and sensor placement strategies within a water distribution network?

1.4 Scope of the study

The scope of this study focuses on developing methods to support IWN systems for the efficient monitoring, detection, and management of leaks and bursts in water distribution networks. The study focuses on integrating software and hardware solutions, including SCADA systems, IoT-enabled flow meters, sensors, and hydraulic models, to address existing challenges in water network operations. Key areas of exploration include optimal placement of pressure sensors within water pipelines, utilising system design, historical leak data, and dynamic water demand patterns to enhance network efficiency.

The study encompasses the development of an advanced leak and burst detection model utilising Machine Learning techniques, with a focus on both supervised and unsupervised algorithms. These methodologies are applied to analyse flow, pressure, and pump speed data from water pipeline networks, with the overarching goal of identifying and optimising the most accurate model for detecting leaks and bursts in operational conditions. The choice of Machine Learning models for this research is justified by their ability to process and analyse complex, high-dimensional datasets, which include critical parameters such as flow rates, pressure readings, pump speeds, and timestamps. Unlike traditional methods, Machine Learning algorithms are uniquely equipped to uncover hidden patterns and correlations within live data streams, enabling more nuanced and reliable detection capabilities. They can process large volumes of data quickly and efficiently, reducing the likelihood of undetected leaks or bursts and minimising false positives. Furthermore, their scalability makes them suitable for deployment across diverse water distribution networks, from small rural systems to large urban infrastructures. By leveraging these capabilities, the proposed methodology aims to significantly enhance the accuracy, reliability, and efficiency of

leak and burst detection, ultimately contributing to improved water resource management, reduced water loss, and the sustainability of critical infrastructure systems.

Supervised Machine Learning algorithms are leveraged to classify and predict known leak and burst patterns based on labelled training data, while unsupervised Machine Learning techniques are employed to identify anomalies and detect previously unknown or unexpected events. This combination of these two techniques ensures the model's robustness and adaptability to diverse operating conditions and varying network configurations. Moreover, Machine Learning models can be iteratively trained and refined as more live data is collected, enhancing their predictive performance and maintaining high levels of accuracy over time.

To validate the proposed methodology, experimental simulations are conducted within a controlled environment. Flow meters and pressure sensors are strategically positioned at optimal locations within the network to replicate real-world conditions, including the occurrence of leaks and bursts under predefined scenarios. These simulations enable the calibration and fine-tuning of the detection model to ensure its effectiveness in practical applications. Live data from these experiments is transmitted through a SCADA-based wireless communication system, allowing centralised collection, monitoring, and analysis. This live data serves as the foundation for evaluating the model's detection capabilities and system response under realistic operating conditions.

The integration of Machine Learning into this framework provides several advantages over traditional detection techniques. The conceptual framework of an IWN is illustrated in Figure 1.2. The IWN System diagram highlights three core layers:

1. Water system and Operation Layer – involving the water network system (pipes, pumps, treatment systems) and real-time hydraulic simulations for operational efficiency.
2. Communication and Sensing Layer – integrating SCADA, IoT, flow meters, and sensors for seamless real-time data transmission and monitoring.
3. Application and Prediction Layer – focusing on system management and the development of a predictive model for detecting leaks and bursts based on flow and pressure patterns.

The study was conducted in the Victorian Sunbury region, where SCADA data, including flow,

pressure, and pump speed, were analysed. Experimental testing and validation were carried out in collaboration with water utility's facilities to ensure accuracy and reliability. The findings of this research aim to enhance water conservation efforts, improve infrastructure management, and provide water authorities with a robust tool for addressing leak and burst-related challenges in existing and future water distribution systems.

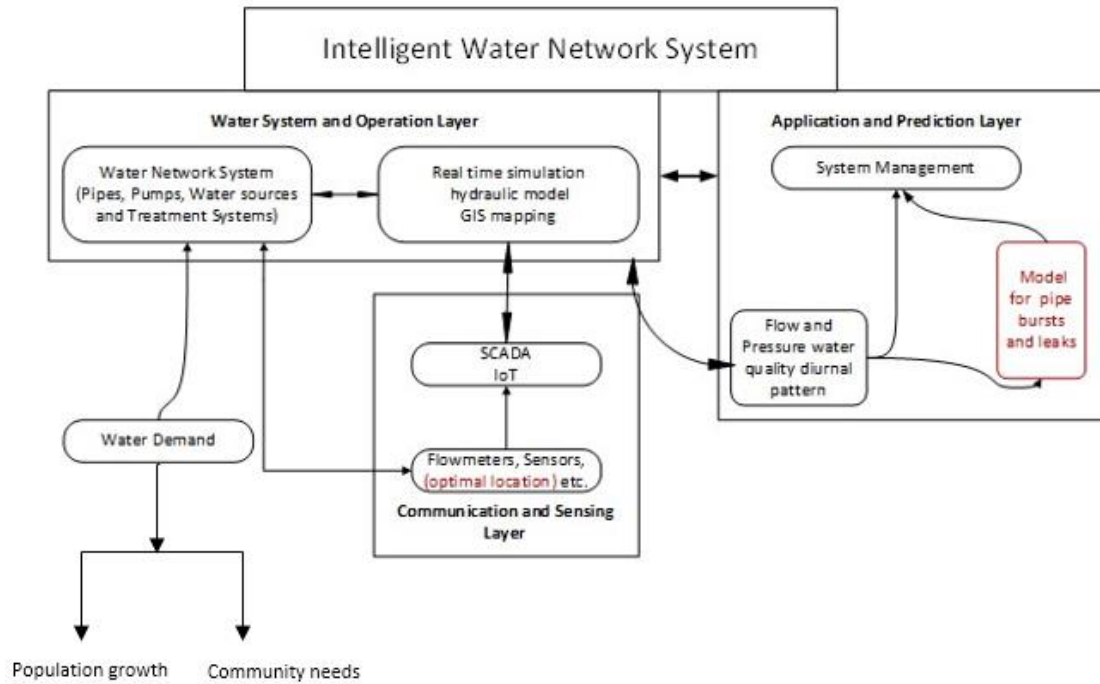


Figure 1.2: A conceptual plan for an IWN system (adopted from Joseph et.al. (2022))

Each layer shown in Figure 1.2 plays a distinct role, integrating advanced technologies, predictive analytics, and live data monitoring to address challenges in water resource management, including leaks, bursts, and operational inefficiencies. The system is structured into three primary layers that collectively ensure efficient and sustainable water network management. The scope of this research is focused on the tasks highlighted in red in Figure 1.2.

1.4.1 Water System and Operation Layer

This layer forms the physical and operational foundation of the water distribution network, focusing on the infrastructure and the processes required to deliver water to consumers efficiently. The key components are as follows:

1.4.1.1 Water Network System

This includes critical infrastructure such as:

- a) Pipes: Responsible for transporting water across the network. Pipe material, size, and layout play a crucial role in ensuring efficient flow and minimising losses.
- b) Pumps: Regulate water pressure and ensure adequate flow, especially in areas with varying elevations or high demand.
- c) Water Sources: Surface water (rivers, lakes) and groundwater sources supply the raw water required for distribution.
- d) Treatment Systems: Ensure water quality by removing impurities and contaminants before distribution to meet safety standards.

1.4.1.2 Real-time simulation hydraulic model and GIS mapping

- a) The hydraulic model simulates water movement, pressure, and flow throughout the network. It predicts how the system will respond under different conditions, such as changes in demand, operational adjustments, or potential failures.
- b) GIS mapping provides a visual representation of the network, including the location of pipes, sensors, and valves, enabling better spatial planning and quicker identification of problem areas.

The functionality of the Water system and Operation layer uses population growth, community needs, and water demand data to dynamically adjust system operations. It ensures a balance between water supply and demand while maintaining system reliability and efficiency. By simulating real-time scenarios, it helps pre-emptively identify vulnerabilities and optimise water resource allocation.

1.4.2 Communication and Sensing Layer

This layer bridges the physical network and the analytical components by collecting, transmitting, and processing live data from the water distribution system. It serves as the nervous system of the IWN, ensuring real-time awareness of network performance. The key components are as follows:

1.4.2.1 SCADA System

SCADA systems collect, monitor, and control live data from sensors and devices distributed

across the network. It provides operators with a centralised interface to view and manage the entire system, making it possible to detect and respond to issues promptly.

1.4.2.2 Internet of Things (IoT) Devices

The IoT devices include,

- a) Flow meters: Measure the rate of water flow in various parts of the network, detecting irregularities indicative of leaks or bursts.
- b) Pressure sensors: Monitor pressure variations to identify anomalies caused by bursts, blockages, or pump malfunctions.
- c) Optimal sensor placement: Sensors are strategically placed based on hydraulic modelling and network analysis to maximise data accuracy and coverage while minimising costs.

The functionality of the communication and sensing layer has a few stages. This layer facilitates live data transmission from the network to centralised systems for analysis, ensuring continuous system awareness. It enables two-way communication, allowing remote adjustments to operational parameters based on insights from the Application and Prediction Layer. IoT-enabled devices enhance the scalability and flexibility of the system, making it adaptable to evolving network requirements.

1.4.3 Application and Prediction Layer

This layer is the brain of the IWN, where live data is analysed and decisions are made. It employs advanced computational techniques, including predictive analytics and Machine Learning, to provide actionable insights and enhance system performance. In terms of system management, centralised control mechanisms process live data to monitor water quality, optimise pump schedules, and address inefficiencies in real time. Within this layer, Machine Learning models analyse flow, pressure, and other operational parameters to predict and classify leaks and bursts. These predictive algorithms combine both supervised learning to classify known patterns and

unsupervised learning to identify previously unknown anomalies, ensuring comprehensive coverage of potential failure scenarios.

The models continuously learn from operational patterns, using flow and pressure data to establish baselines for normal system behaviour, where deviations from this baseline signal potential issues such as leaks, bursts, or contamination. This adaptive capability enables proactive maintenance and faster response to abnormal events. The functionality of this layer includes enhancing predictive accuracy to minimise false alarms, reducing downtime through refined detection algorithms that flag only legitimate issues, and feeding the insights generated by this layer back into the Water system and Operation Layer to optimise overall network performance.

The IWN achieves its objectives through seamless integration and interaction among its three layers, forming a continuous feedback loop that drives system intelligence and efficiency.

1. Demand-driven design ensures that population growth, community needs, and water-demand data dynamically shape the operation of the Water system and Operation layer.
2. Live data collection from the Communication and Sensing Layer provides real-time measurements of flow, pressure, and quality across the network, which are transmitted to the Application and Prediction Layer for advanced analytics and decision-making.
3. Feedback and control mechanisms then relay insights from the analytical layer back to the operational layer, enabling continual optimisation of water distribution, pressure regulation, and resource allocation.

Through this closed-loop framework, the IWN delivers multiple benefits: it enhances leak- and burst-detection accuracy through optimal sensor placement and Machine Learning analysis; improves operational efficiency by reducing energy consumption and response times; promotes long-term sustainability by minimising losses and maximising resource utilisation; and maintains scalability, allowing deployment across both small rural schemes and complex urban infrastructures.

1.5 Research Significance and Innovations

This research contributes to improving the efficiency, resilience, and sustainability of modern water distribution systems in the context of increasing demand, urbanisation, and ageing infrastructure.

The significance of this research can be categorised into economic, societal, environmental, and research contributions. It presents a unified framework that combines real-time monitoring, optimised pressure sensor placement, and data-driven decision-making to strengthen network management and water conservation efforts.

From an economic perspective, the proposed framework reduces operational and maintenance costs by enabling early leak and burst detection, minimising non-revenue water (NRW), and optimising sensor deployment.

From a societal perspective, the research supports reliable water supply, which is essential for public health, service continuity, and urban resilience. Timely detection of system failures enables faster intervention and reduces disruptions to consumers.

From an environmental perspective, the framework contributes to sustainable water management by reducing water losses and energy consumption associated with pumping and treatment processes.

A key innovation of this study is the development of an algorithm for the optimal placement of pressure sensors within water pipelines, designed to maximise network coverage and enhance the accuracy of leak and burst detection. This algorithm is integrated with SCADA-based flow and pressure monitoring devices, enabling real-time data acquisition and continuous assessment of hydraulic behaviour. The resulting data feeds directly into Machine Learning models that analyse operational patterns, predict anomalies, and generate automated alerts for water authorities.

From a research contribution perspective, this study introduces a hybrid approach that combines logic-based methods with Machine Learning techniques. While logic-based methods are effective for simple pipeline systems, they are limited in capturing complex, non-linear patterns in large-scale water distribution networks. Machine Learning models are therefore utilised to address these complexities and improve detection accuracy.

The framework has been validated through on-site experiments and real-world SCADA datasets, demonstrating its scalability and adaptability across diverse network sizes. By integrating advanced sensing technologies with intelligent analytics, the system enhances operational efficiency, reduces costs, and strengthens long-term infrastructure sustainability.

Overall, this research advances water-network management through a proactive, data-driven approach that supports resource optimisation, informed decision-making, and sustainable infrastructure development.

1.6 Chapter Summary and Reflections

This chapter establishes the intellectual and practical foundation of the thesis by contextualising the need for intelligent systems in modern water distribution networks. It provides a comprehensive overview of how digital technologies, including SCADA systems, IoT-enabled sensing, and data-driven analytics, can transform conventional water management into a predictive and adaptive framework (Mounce et al., 2010; Colombo, Lee & Karney, 2009).

A key focus of this chapter is the importance of detecting leaks and bursts at an early stage. These events are not only technical challenges but also have significant quantifiable impacts, including increased non-revenue water (NRW), higher energy consumption due to inefficient pumping, and substantial economic losses associated with infrastructure damage and maintenance (Farley et al., 2008; Puust et al., 2010). Beyond economic implications, undetected leaks and bursts can affect community water reliability, public health, and environmental sustainability, particularly in regions facing water scarcity (WHO, 2017; UNESCO, 2020). Therefore, effective leak and burst detection is positioned as a critical component of sustainable water resource management and infrastructure resilience.

From a broader perspective, the chapter highlights the necessity of interdisciplinary integration in addressing these challenges. Modern water distribution systems are inherently complex, requiring the convergence of hydraulic engineering, artificial intelligence, and system automation. Traditional approaches, which rely primarily on rule-based or isolated methods, are often insufficient for capturing the dynamic and non-linear behaviour of large-scale networks (Romano, Kapelan & Savic, 2014). In contrast, Intelligent Water Networks leverage data-driven models capable of learning, adapting, and optimising operations in real time (Soldevila et al., 2016).

The chapter also reflects on the role of combining logic-based methods with Machine Learning approaches. While logic-based techniques provide simplicity, interpretability, and effectiveness in smaller or less complex systems, they are limited in handling the intricate patterns present in large-scale water distribution networks. Machine Learning methods, on the other hand, offer the

capability to detect complex, non-linear relationships and improve predictive accuracy (Soldevila et al., 2016). This complementary approach forms the conceptual basis for the framework developed in this thesis.

The knowledge and conceptual clarity developed in this chapter provide a strong foundation for the subsequent Literature Review (Chapter 2). The next chapter critically examines existing research in leak and burst detection, sensor placement optimisation, and intelligent monitoring systems, identifying key gaps and limitations that motivate the proposed research framework.

Chapter 2

Literature Review

2.1 General Remarks

This chapter establishes the conceptual and technical foundation for the thesis by surveying the evolution of leak and burst detection in water distribution systems and the emergence of Intelligent Water Networks). It clarifies terminology used throughout the thesis (e.g., "leak" vs "burst", "hardware-based" vs "software-based" methods, "optimal sensor allocation"), and sets out the organising taxonomy adopted here: (i) conventional field practices, (ii) hardware-based approaches, (iii) software-based approaches (model-driven, AI/ML), and (iv) sensor placement and network observability. The literature reviewed prioritises peer-reviewed studies, standards, and utility reports that (a) quantify detection performance, (b) report deployment constraints (cost, data quality, telemetry), and (c) discuss transferability to utility-scale operations. Where appropriate, benchmark networks and real-world case studies are considered to bridge the gap between laboratory setups and operational conditions. Emphasis is placed on how data acquisition (SCADA/AMI/IoT), hydraulic modelling, and analytics integrate into decision support for leak reduction, pressure management, energy recovery, and asset strategies. These general remarks also frame the chapter's role in the overall thesis: to distil principles that guide the design of the case studies and the selection of methods validated in later chapters. Specifically, the literature insights inform (1) which sensors and sampling rates are most informative, (2) how to structure detection logic versus data-driven classifiers, and (3) how to place pressure sensors to maximise observability under realistic budgetary and operational constraints. The remainder of the chapter follows this structure, culminating in cross-cutting reflections and gaps that motivate the experimental and modelling programmes presented next.

2.2 Introduction

Water distribution systems are increasingly challenged by ageing infrastructure, urbanisation, and rising demand, necessitating advanced monitoring and management approaches. Traditional

reactive methods are no longer sufficient, leading to the emergence of Intelligent Water Networks (IWNs) that integrate sensing, data analytics, and automation (Colombo, Lee & Karney, 2009). This chapter reviews existing research on leak and burst detection, sensor placement strategies, and the application of Machine Learning in water systems, identifying key gaps that motivate this study. Water distribution networks' challenges have driven a transition from traditional reactive management approaches toward intelligent and data-driven systems. Recent studies highlight the emergence of smart water systems and Intelligent Water Networks (IWNs), which integrate sensing, communication, and analytics to enable predictive and adaptive management (Meseguer et al. 2023; Pagano et al. 2025). Pipeline networks are vital infrastructure for transporting water to urban and rural communities (Sapkota et al. 2021; Ren et al. 2022; Dong et al. 2022; Kang & Lansey 2009). The increasing demand for reliable and sustainable supply has driven innovations in intelligent water management and leak detection (Joseph et al. 2022). Pipeline failures through bursts or hidden leaks can lead to significant property damage, environmental impacts, and economic loss (Al-Ghamdi 2011; Darweesh & Moghavvemi 2020). The development of Intelligent Water Networks and modern leak-detection technologies has therefore become central to sustainable urban water-system management. Historically, the development of water networks reflects the evolution of civilisation itself. Urban water-supply and sewage systems were already established in Crete, the Aegean Islands, and the Indus Valley civilisations during 3200–1100 BC (Angelakis et al. 2005). The Mesopotamians implemented advanced water-management systems (Angelakis & Koutsoyiannis 2003), and the Minoan, Egyptian and Indus Valley civilisations influenced later Greek and Etruscan water-management technologies (Angelakis et al. 2012). Such advances in hydraulic engineering laid the foundations for modern water, wastewater and storm-water systems (Mays 2010). Mathematical modelling of pressure and flow in pipe networks has long attracted engineers, progressing from graphical and physical analogies to numerical methods such as the Hardy Cross method. In Australia, utilities are seeking innovative strategies to reduce water loss and infrastructure costs amid drought and rising energy prices (Marney & Sharma 2012). These challenges mirror global pressures of urbanisation, population growth and climate variability. Pipeline networks are crucial for moving water from one place to another to serve larger rural and urban communities (Ullah, Faheem & Kim 2017; Chen et al. 2017; Bacco et al. 2017; Hunaidi et al. 2000). Pipelines have also been used since World War II to transport gas and oil (Liu 2021). The study by Wang et al. (2021) described how a supervisory control and data acquisition (SCADA) system was used to collect production and operational data from an oil field

in Western China and outlined trends toward intelligent oil-field transportation networks with big-data analytics, automatic control, and situational awareness. Similar automation principles apply to water distribution systems as they evolve into IWNs. Kraidi et al. (2021) reviewed oil and gas pipelines in Iraq and comparable countries, highlighting risks associated with pipeline failures and mitigation approaches relevant to water supply networks. Failures of water-supply pipelines cause property damage, environmental impacts, and even loss of life (Shaw et al. 2012; Zhang et al. 2013).

Recent leakage incidents have heightened environmental awareness and public concern, while increasing operational costs for utilities through staff time and clean-up efforts. Globally, non-revenue water (NRW) is estimated to account for approximately 30–40% of total water supply in many systems, representing billions of cubic metres of lost water annually (Farley et al., 2008; Puust et al., 2010). Affordable and reliable leak-detection systems are therefore increasingly in demand as regulators in industrialised nations tighten requirements to minimise public water losses (Zhang, 1997). Traditional leak detection methods include acoustic monitoring, pressure-based analysis, and transient modelling techniques. These methods have been widely studied and applied, but they often require manual intervention and are difficult to scale across large networks (Colombo, Lee & Karney 2009; Puust et al. 2010). Transient-based methods, while effective in controlled environments, can be complex and sensitive to noise in real-world applications (Colombo, Lee & Karney 2009). Recent review studies emphasise a shift toward integrated approaches that combine physical modelling with data-driven techniques. Meseguer et al. (2023) highlight that modern leak detection research is increasingly focused on improving robustness, scalability, and real-time applicability rather than relying solely on detection accuracy. The main limitations of traditional supply systems are deterioration and ageing infrastructure leading to water loss through leakages and bursts (Colombo & Karney 2002). Quantifying losses, identifying leak locations and reducing response time remain persistent challenges; hence, water-loss mitigation strategies are essential (Colombo & Karney 2002). Further risks include earthquakes, storms, floods, pollution from point and non-point sources, and hydrological variability induced by climate change (IPCC 2021). Average rehabilitation rates rarely match expanding demand and asset age. The significance of research on leak detection stems from its potential to enable early identification of events such as water leaks and bursts, water theft, low-pressure or high-flow conditions, blockages, and contamination. Intelligent leak-detection and burst-response systems therefore offer socio-economic benefits, including reducing non-revenue water, lowering operational costs, and increasing utility revenue (Joseph, Sharma & van Staden 2022). Real-time data support decision-making and minimise

customer impacts (Thompson et al. 2016). Such systems reflect a shift towards Intelligent Water Networks capable of predicting events and acting proactively (Marney & Sharma 2012).

Leakage in water distribution systems commonly occurs at pipe joints, flanges, primarily as a result of material deterioration, aging, and structural degradation of the pipes. The main limitations of traditional supply systems include deterioration and ageing infrastructure, which significantly contribute to leakage and burst events (Colombo & Karney, 2002). Studies indicate that leakage can account for 15–50% of total system input volume depending on network condition, highlighting the scale of the problem (Puust et al., 2010). Quantifying losses, identifying leak locations, and reducing response time remain persistent challenges; however, effective mitigation strategies can substantially reduce water loss and operational costs (Colombo & Karney, 2002). Additional risks such as climate variability, extreme weather events, and pollution further exacerbate system vulnerability (IPCC, 2021).

Intelligent leak-detection and burst-response systems offer significant socio-economic benefits. For example, utilities implementing advanced monitoring systems have reported reductions in NRW of up to 20–30% and improved operational efficiency (Sharma, Sharma & Staden, 2022). Real-time data-driven systems also enable faster response times, reducing customer disruptions and minimising economic losses (Thompson et al., 2016). These developments reflect a broader shift towards Intelligent Water Networks, which support predictive and proactive system management (Marney & Sharma, 2012).

Smart-city frameworks and the concept of “smart environments” emphasise the efficient management of water, energy, and environmental resources (Ullah, Faheem & Kim, 2017; Chen et al., 2017; Bacco et al., 2017; Kulkarni & Farnham, 2016; Rostirolla et al., 2017). Studies have shown that the integration of IoT and ICT technologies can improve water-use efficiency by 10–25% and significantly enhance real-time monitoring capabilities (Bacco et al., 2017; Chen et al., 2017). Because large volumes of water are transported through complex distribution systems, leaks and bursts can result in substantial economic losses, environmental damage, and public health risks.

To address these challenges, numerous techniques have been developed to detect leak frequency and magnitude. For instance, pressure-based and transient analysis methods can detect leaks with location accuracy typically within 5–10% of pipe length under controlled conditions (Hunaidi et

al., 2000), while recent Machine Learning approaches have demonstrated higher detection accuracy and reduced false alarms compared to traditional methods (Ismail et al., 2019; Hu et al., 2021; Joseph et al., 2023).

The evolution of these techniques is closely linked to the development of Intelligent Water Networks, which integrate sensing, communication, modelling, and automation to enable sustainable and adaptive infrastructure (Sharma et al., 2013; Ramos et al., 2019).

Despite the effectiveness of these methods in isolation, their limitations such as cost inefficiency, difficulty in pinpointing exact leak locations, and their reactive nature highlight the need for integrated, proactive solutions. Software-based (Machine Learning) and hardware-based approaches now play a critical role in optimising leak management by combining hydraulic modelling, sensor data, and economic analysis. This integration allows for real-time monitoring, predictive maintenance, and economically optimal decision-making across leak detection, repair scheduling, and pressure management. By integrating conventional hydraulic models with Machine Learning and optimisation frameworks, it is now possible to simultaneously evaluate the impact of pressure reduction and leak-detection frequency. This holistic approach results in greater accuracy, better resource allocation, and reduced non-revenue-water losses benefits clearly demonstrated in recent case studies such as that by Hammond et al. (2024), which found leak avoidance equivalent to 13–23 % of customer demand and energy recovery equal to 10–11 % of the system's energy needs. An Intelligent Water Network is used to identify and react to potential events or behaviours in water networks prior to their prevalence, to mitigate or prevent their potential consequences (Sharma & Marney 2012). This helps water authorities better understand what is happening in their water systems (Sorbello & Marck 2014). The benefits of IWN include helping to identify and repair leakages quickly, thereby saving water; improving customer service through early problem detection; aiding water authorities in better managing their pipe assets and equipment; and saving money by reducing water waste and avoiding major issues.

2.3 Conventional Methods of Leak Detection

Conventional leak-detection approaches rely primarily on visual inspection, acoustic monitoring, pressure or flow variation analysis, and empirical field techniques (Puust et al. 2010; Colombo & Karney 2002; Hunaidi & Giamou 1998). These techniques have historically formed the backbone

of water-utility maintenance programs and remain valuable in many developing regions where advanced sensing infrastructure is unavailable. Their major advantage lies in the use of simple, low-cost instrumentation and minimal computational requirements; however, they are generally labour-intensive, time-consuming, and reactive rather than predictive (Lambert & Hirner 2000; Li et al. 2015).

Leakage in water distribution systems represents a major global challenge, contributing significantly to non-revenue water (NRW) and operational inefficiencies. Studies estimate that a substantial proportion of treated water is lost due to undetected leaks and bursts, resulting in economic and environmental consequences (Farley et al., 2008).

A wide range of techniques has been developed for leak detection, broadly classified into acoustic methods, pressure-based approaches, and transient analysis techniques (Puust et al., 2010). Among these, transient-based methods have gained attention for their ability to detect leaks through pressure wave propagation, although their practical implementation can be complex (Colombo, Lee & Karney, 2009).

Despite these advancements, conventional approaches often rely on manual inspection or threshold-based logic, which limits their ability to detect subtle or evolving anomalies in large-scale networks.

Manual and Visual Inspection:

Early leak detection depended on manual walk-through inspections, listening rods, ground microphones, and tracer dyes to locate water surfacing or seepage zones (Thornton & Lambert 2006). Operators searched for signs such as damp patches, water pooling, or unexplained vegetation growth above buried pipes (Hunaidi et al. 2000). While inexpensive, such inspections detect only visible leaks and are ineffective for small or underground losses that do not reach the surface (Puust et al. 2010). They also depend heavily on operator experience, environmental conditions, and accessibility of pipelines, resulting in subjective accuracy (Fantozzi et al. 2019). From the reviewed studies, common causes of leaks and bursts in water distribution systems include pipe ageing, corrosion, excessive internal pressure, ground movement, and material deterioration (Puust et al., 2010; Farley et al., 2008). External factors such as traffic loads, temperature variations, and soil conditions (e.g., reactive clay) also contribute to pipe failures. In

terms of location, leaks and bursts are frequently observed at pipe joints, connections, valves, and older pipeline sections where structural integrity is weakened. Additionally, areas subjected to high pressure or transient events, such as near pumping stations or elevation changes, are more prone to failures. These insights highlight the importance of both structural and operational factors in leak occurrence and support the need for continuous monitoring approaches.

Tracer-gas methods introduce inert gases commonly helium or hydrogen into pressurised pipelines; gas escaping through leaks is detected by a gas sensor at the surface (Rathore et al. 2018). These methods provide greater sensitivity than visual inspection and can be used on non-metallic pipes; however, they are costly, require pipe isolation, and are impractical for continuous monitoring.

Acoustic and Correlation Techniques:

Acoustic methods exploit the propagation of leak-induced noise along pipe walls and through water columns (Hunaidi & Giamou 1998; Muggleton et al. 2011). Noise sensors such as hydrophones or accelerometers record the sound spectrum, which is analysed to estimate leak position and severity. Correlators compare signals between two sensors to compute the time delay and hence the leak location (Li et al. 2015). Although acoustic correlation is a standard field method, its accuracy deteriorates with pipe material heterogeneity, background noise, and leak size (Hunaidi & Chu 1999). Plastic pipes, in particular, attenuate high-frequency noise, making leak detection more difficult (Muggleton et al. 2011). Advancements have included digital signal processing and cross-correlation techniques that filter extraneous noise and enhance frequency-band analysis (Kang & Lansey 2009). Yet even modern correlators require experienced operators and can cover limited distances, typically less than 300 m in metallic mains (Rathi & Ingeduld 2013). Moreover, they are reactive: leaks must already exist for detection to occur.

Pressure and Flow Monitoring:

Pressure- and flow-based methods measure variations in hydraulic parameters between neighbouring sections of a network. Abnormal drops in pressure or unexpected increases in flow

may indicate leakage or bursts (Colombo & Karney 2002; Farley et al. 2008). The minimum-night-flow (MNF) technique analyses flow records during low-demand hours to identify deviations from baseline consumption (Farley & Liemberger 2005). Similarly, the water-balance method recommended by the International Water Association (IWA) quantifies “non-revenue water” through systematic accounting of input, billed consumption, and losses (Lambert & Hirner 2000). However, such approaches cannot accurately locate leaks, as pressure and flow readings reflect aggregated system behaviour (Puust et al. 2010). Moreover, they are sensitive to diurnal demand fluctuations, pressure-management changes, and sensor noise (Wu et al. 2016). Thus, conventional hydraulic-parameter analysis serves mainly for system-wide screening rather than pinpoint localisation.

Ground Penetrating Radar and Infrared Thermography:

Ground-penetrating radar (GPR) employs electromagnetic waves to detect subsurface anomalies caused by leaking water (Hunaidi et al. 2000). The difference in dielectric constants between wet and dry soil produces reflections that reveal leakage zones (Ayala-Cabrera et al. 2011). While non-invasive, GPR performance depends on soil conductivity, pipe depth, and surface interference. Similarly, infrared thermography (IRT) senses temperature gradients at the surface due to water seepage; these are effective in shallow pipelines or when leaks cause noticeable cooling (Kong et al. 2018). Both GPR and IRT are valuable supplementary tools but require skilled interpretation and are affected by environmental noise, surface materials, and weather.

Mass-Balance and Comparative Methods:

The mass-balance or inflow-outflow approach measures discrepancies between metered supply and recorded consumption (Lambert & Hirner 2000; Thornton & Lambert 2006). Although conceptually simple, its precision depends on meter accuracy, calibration, and network segmentation (Colombo & Karney 2002). Time-series comparison methods analyse historical data to identify anomalous patterns, but their effectiveness diminishes when operational or seasonal changes alter baseline behaviour (Wu et al. 2016).

Limitations of Conventional Methods:

Despite widespread adoption, conventional methods share critical limitations: (i) low sensitivity to small leaks; (ii) dependence on operator expertise; (iii) discontinuous monitoring; (iv) poor

performance in noisy or complex urban environments; and (v) lack of automation and predictive capacity (Puust et al. 2010; Farley et al. 2008). These shortcomings have driven the transition toward automated, intelligent solutions integrating sensors, telemetry, and data analytics. The emergence of SCADA, Internet of Things (IoT) connectivity, and real-time hydraulic modelling now enable continuous leak surveillance, event classification, and proactive maintenance planning (Joseph et al. 2023; Sharma et al. 2013). The next section therefore explores the technological evolution from traditional inspection and correlation tools to modern hardware-based leak and burst detection systems.

Historically, leak management in water distribution systems has relied on conventional methods such as acoustic detection (e.g. ground microphones, leak noise correlators, and noise loggers), ground-penetrating radar, and thermal infrared imaging. These techniques, while valuable, are often labour-intensive, limited by environmental noise, and less effective in complex or buried infrastructures. Additionally, pressure management using Pressure Reducing Valves (PRVs) has traditionally been used to control leaks by minimising excess system pressure. However, while effective, PRVs dissipate energy that could otherwise be recovered. Recent innovations such as Pumps-as-Turbines (PATs) offer a dual advantage by reducing pressure and recovering energy, enhancing both efficiency and sustainability. The industry has also explored preventative pipe-replacement strategies, but these are often not economically justifiable compared to detection and repair or pressure-reduction methods (Fields et al. 2009). All field instrumentation and laboratory measurements in this study follow the metrological and calibration standards outlined in ISO 4064-1:2014 and ISO/IEC 17025:2017 to ensure traceable accuracy and testing reliability (ISO 4064-1:2014; ISO/IEC 17025:2017).

2.4 Hardware-Based Leak and Burst Detection Methods

Hardware-based leak-detection systems rely on the deployment of physical sensors and field instruments that monitor hydraulic, acoustic, or structural responses of pipeline systems in real time. These methods use transducers, flow meters, acoustic sensors, fibre-optic cables, pressure loggers, and distributed sensor networks to directly capture physical phenomena associated with

leakage or bursts (Hunaidi & Chu 1999; Colombo & Karney 2002; Puust et al. 2010). They represent a critical link between traditional manual techniques and modern intelligent monitoring frameworks that underpin the evolution of Smart Water Networks (SWAN) and Intelligent Water Networks. The primary advantage of hardware-based methods is their ability to provide continuous, quantitative measurements with a high degree of spatial and temporal resolution. However, they often require significant capital investment, careful calibration, and maintenance under varying environmental and hydraulic conditions (Misiunas et al. 2005; Puust et al. 2010). These technologies are now complemented by software and data-driven models that interpret the collected signals, but the foundation of any leak-detection system remains reliable, field-validated hardware.

Pressure- and Flow-Sensor-Based Systems:

Pressure transducers and flow meters are among the most common hardware used for leakage monitoring in water distribution systems (Farley et al. 2008; Wu et al. 2016). Abnormal pressure drops or fluctuations may indicate potential leaks or bursts, while deviations in flow measurements from expected values can reveal unaccounted water losses (Puust et al. 2010). The deployment of high-frequency pressure loggers across a network enables transient-event monitoring, where sharp negative or positive waves suggest pipe ruptures or valve operations (Misiunas 2005; Palau et al. 2012). These devices record data at intervals as low as 1 Hz or faster, which are then transmitted to central databases through SCADA or IoT gateways for real-time analysis. Advanced systems also incorporate piezoelectric sensors or strain gauges attached to critical nodes for continuous stress and vibration monitoring. However, effective placement of pressure and flow sensors remains a technical challenge due to network topology and hydraulic diversity (Sharma et al. 2013). Over- or under-instrumentation can result in suboptimal performance or excessive costs. Therefore, the integration of pressure-based sensors with decision-support algorithms, data-compression techniques, and predictive modelling has been proposed to enhance diagnostic capability (Colombo & Karney 2002; Joseph et al. 2023).

Acoustic Emission and Correlation Hardware:

Acoustic-based hardware devices are widely used because leaks and bursts produce distinctive sound signatures as pressurised water escapes through orifices or cracks (Hunaidi & Chu 1999;

Muggleton et al. 2011). Acoustic emission sensors such as geophones, hydrophones, or accelerometers are installed along pipelines to detect these vibrations. The recorded signals are processed using cross-correlation, spectral analysis, or wavelet transforms to identify the source location (Li et al. 2015). Advanced correlation loggers such as the LeakFinderST and Permalog systems use wireless communication to synchronise time stamps and perform distributed leak detection (Rathi & Ingeduld 2013). The correlation between signals measured at two points determines the leak location based on the velocity of sound in the pipe medium. In practice, these systems provide accurate results for metallic pipes and moderate distances but experience limitations in plastic networks due to high damping of acoustic energy (Muggleton et al. 2011). Environmental noise from pumps, vehicles, and consumer use can also mask leak sounds. Nevertheless, recent improvements in digital signal processing (DSP) hardware and low-power wireless acoustic loggers have significantly enhanced detection performance. For example, large-scale networks can deploy hundreds of permanently installed nodes that communicate via mesh networks and synchronise measurements for continuous leak surveillance (Zhang et al. 2013; Hu et al. 2021). Such distributed systems can identify leak signatures before they become visible, reducing water loss and operational risk.

Fibre-Optic and Distributed Temperature Sensing (DTS) Systems:

Fibre-optic sensing represents one of the most advanced hardware technologies for pipeline monitoring, capable of detecting minute temperature, strain, or acoustic variations along several kilometres of buried pipeline (Miah et al. 2020; Kang et al. 2021). In distributed temperature sensing (DTS) systems, a fibre-optic cable functions as a continuous sensor: laser pulses transmitted along the fibre generate backscattered light that varies with temperature or strain, allowing the detection of leak-induced thermal anomalies (Bao & Chen 2012). Distributed Acoustic Sensing (DAS) applies similar principles to detect vibrations and acoustic emissions associated with leaks or external interference such as digging or vandalism. These methods provide spatially continuous data rather than discrete-point measurements, yielding unparalleled coverage and early-warning capability (Miah et al. 2020). While installation costs and data-management complexity remain significant, the durability and immunity of fibre-optic cables to electromagnetic interference make them ideal for high-risk pipelines such as long-distance transmission mains or critical water crossings (Gao et al. 2022). In addition, modern hybrid systems combine DAS with

DTS to simultaneously detect both temperature and acoustic changes, enabling comprehensive leak and intrusion monitoring (Li et al. 2019).

Smart Wireless Sensor Networks and IoT-Based Devices:

The rapid growth of the Internet of Things (IoT) has revolutionised hardware-based leak detection. Miniaturised pressure, flow, and vibration sensors equipped with low-power microcontrollers can now transmit data wirelessly via Zigbee, LoRa, NB-IoT, or 5G protocols (Ramos et al. 2019; Al-Fuqaha et al. 2015). These nodes form wireless sensor networks (WSNs) capable of real-time communication, self-organisation, and multi-hop data routing across vast water-distribution areas (Ullah, Faheem & Kim 2017). For instance, the work by Shiddiq et al. (2021) proposed an IoT-based monitoring system integrating ultrasonic flow sensors with a cloud database and mobile application for visualising leaks in real time. Similarly, Mutchek and Williams (2014) demonstrated a smart-water-grid prototype using pressure and flow sensors to transmit readings via wireless gateways to SCADA servers. The integration of IoT enables automatic leak alarms, device diagnostics, and energy-efficient operation through duty-cycling and event-based communication (Bacco et al. 2017). However, challenges persist regarding power management, communication reliability, cybersecurity, and sensor calibration. Battery life remains a limiting factor for remote installations, necessitating the use of energy-harvesting techniques such as piezoelectric or solar charging (Ramos et al. 2019).

Vibration, Strain, and Ultrasonic Sensors:

Leakage generates mechanical vibrations that propagate along the pipe and surrounding soil. Vibration sensors (accelerometers or MEMS-based devices) detect these dynamic responses and provide data for structural-health monitoring (Li et al. 2015). Ultrasonic sensors, meanwhile, use the time-of-flight principle to measure flow rates and detect anomalies due to leaks or obstructions (Kim et al. 2018). Strain gauges can also monitor pipe deformation, which often precedes failure. By integrating strain measurements with finite-element models, engineers can predict stress concentrations and schedule maintenance proactively (Zhou et al. 2020). These sensors are often deployed in critical pipeline segments, such as high-pressure mains or bridge crossings, where early detection of stress changes prevent catastrophic bursts.

Satellite and Aerial Monitoring Hardware:

Satellite-based leak detection technologies have recently emerged as a complementary approach. They use synthetic aperture radar (SAR) or hyperspectral imaging to identify soil-moisture anomalies or vegetation stress patterns indicative of leaks (Ochoa-Rodriguez et al. 2019). Airborne and drone-mounted sensors equipped with thermal or multispectral cameras can perform large-scale surveys rapidly, reducing manual inspection time (Kang et al. 2021). Although spatial resolution and data-processing costs remain barriers to widespread adoption, these systems offer an effective macro-level monitoring tool for utilities managing extensive or remote networks. The hardware integration between aerial sensors and ground-based IoT nodes further enhances cross-validation and event localisation.

Advantages and Limitations:

Hardware-based leak detection provides direct, physically grounded measurements of network conditions. Its key strengths include high detection accuracy, real-time capability, and independence from complex simulation models (Hunaidi & Chu 1999; Puust et al. 2010). However, these benefits are counterbalanced by high installation and maintenance costs, data-management burdens, and potential sensor drift or malfunction under harsh conditions. Hardware systems alone cannot infer root causes or optimise sensor placement without software integration (Sharma et al. 2013; Joseph et al. 2023). As utilities transition toward Smart Water Networks, hardware is increasingly viewed as the sensory layer of a multi-tier architecture comprising data acquisition, communication, analytics, and decision-making. The next section therefore explores the software-based leak and burst detection methods that interpret sensor data, perform pattern recognition, and support predictive diagnostics across Intelligent Water Networks. Liu et al. (2019) employed wireless sensors with 4G transmission for remote data collection. They used an SVM classifier for leak identification and demonstrated reduced power consumption compared to conventional methods while maintaining coverage.

2.5 Software-Based Leak and Burst Detection Methods

Software-based leak and burst detection methods use mathematical models, numerical simulations, statistical pattern recognition, and artificial intelligence to interpret hydraulic and sensor data for fault identification and localisation (Puust et al. 2010; Mashford et al. 2009; Mounce & Boxall

2010). These methods form the analytical core of Intelligent Water Networks, converting raw telemetry into actionable knowledge for real-time decision-making (Sharma et al. 2013; Joseph et al. 2023). Unlike purely hardware-based approaches, software models can operate on historical or simulated data, enabling prediction, event classification, and optimal sensor deployment even before full network instrumentation is achieved. They are indispensable for large, complex water distribution systems where physical inspection or exhaustive hardware coverage is impractical. Berardi and Giustolisi (2021) presented a physically based approach for the calibration of water distribution network (WDN) hydraulic models aimed at supporting leakage management plans since the early stages. Their application on two real networks, supported by experience across many real WDNs, demonstrates that pressure and flow monitoring can be used to calibrate a design model for early-stage leakage management. Joseph et al. (2024) described an integrated hardware-and software-framework tailored for an Intelligent Water Network system. In this study, the water system established connectivity among flow meters, pressure sensors, and other monitoring devices through the SCADA system, linking them to a data-analysis centre. Data from flow and pressure sensors were harnessed for calibrating the hydraulic model and for comparison with the real-time simulations of the water network. Anticipated future increases in water demand will require rapid actions to improve resources, reduce demand, and increase treatment and transmission efficiency, further promoting the need for intelligent networks.

Campos, Sousa and Teixeira (2018) suggested an Internet-of-Things (IoT) framework with several layers for water-supply networks. These layers form an Intelligent Water Network framework comprising:

- (i) the sensor layer;
- (ii) the communication layer;
- (iii) the water-system and operation layer; and
- (iv) the application and prediction layer.

SCADA receives data from sensors and flow metres for flow, pressure, and water-quality characteristics. The optimal distribution of pressure sensors and flow metres depends on the topography of the area, the size of the water-delivery system, and historical data on water-quantity changes caused by the ageing infrastructure, environmental conditions, and other factors. The SCADA system connects flow metres, pressure sensors, and other monitoring devices to the data-

analysis centre. Data from these sensors are used to calibrate the hydraulic model and compare with real-time simulation results of the water network.

Fereidooni, Ghaderi, Nazif and Ashoori (2019) proposed a quick hybrid method for detecting leaks and estimating the volume of material lost by combining hydraulic relations with artificial-intelligence algorithms. Their technique uses straightforward, low-cost flow sensors deployed at each pipeline-network junction. By applying hydraulic equations such as Hazen–Williams, Darcy–Weisbach, and pressure drop, influential features for leak identification were produced. Prediction models using decision trees, k-nearest neighbours (KNN), Random Forest, and Bayesian networks were constructed, and based on the pipeline topology, leaks and their pressures were identified. Gorenstein (2022) demonstrated that data-driven algorithms outperform logic-based models in each metric by at least 5 per cent. Furthermore, as algorithms are trained with new data, their predictive accuracy improves, while adding attributes connected to geography does not significantly increase accuracy. Gorenstein suggested that the development of more complex prediction algorithms, such as Bayesian-belief networks and deep neural networks, should be the focus of future research. The development of the logic-based method lies in its structured approach to initial leak and burst detection. Logic-based algorithms efficiently capture straightforward anomalies that align with predefined conditions, offering a clear and interpretable framework for identifying common patterns and deviations within data. However, such models may struggle with complex or evolving patterns outside predefined rules. This limitation highlights the value of integrating Machine Learning techniques, which excel at capturing intricate patterns and outliers not apparent through fixed rules. The relationship between the logic-based and Machine Learning-based approaches is therefore complementary. The logic-based method provides a structured and efficient framework for initial detection, while the Machine Learning-based approach enhances detection by learning from historical data, adapting to changing conditions, and capturing complex patterns beyond fixed rules. Together, they achieve a comprehensive and effective approach to leak and burst detection, leveraging the strengths of each to enhance overall performance and reliability.

Future research should focus on refining these methodologies, developing algorithms with fewer false alarms, and optimising sensor deployment to improve accuracy and reliability in leak and burst detection.

2.5.1 Hydraulic and Model-Based Approaches

The integration of SCADA systems has enabled real-time monitoring of flow and pressure within water distribution networks. These systems provide continuous data streams that support operational decision-making and system control. However, while SCADA systems offer extensive data availability, their full potential is often underutilised due to limited analytical capabilities. Traditional monitoring frameworks typically rely on predefined rules, which are insufficient for capturing complex hydraulic behaviours (Mounce, Boxall & Machell, 2010). This limitation has driven the transition towards Intelligent Water Networks, where real-time data is combined with advanced analytics to enable predictive and adaptive system management. Recent developments have expanded SCADA into broader smart water systems that incorporate IoT technologies, distributed sensors, and advanced communication networks. Pagano et al. (2025) conducted a comprehensive survey of IoT applications in water systems, identifying key challenges such as scalability, interoperability, and energy efficiency. Similarly, Chen et al. (2017) and Bacco et al. (2017) highlight that integrating IoT and ICT technologies enables real-time monitoring, improved system visibility, and enhanced operational decision-making. These systems can improve water-use efficiency by approximately 10–25%, demonstrating their potential for large-scale deployment.

Hydraulic modelling has long been used to simulate water-distribution behaviour under varying demands and boundary conditions. Models such as EPANET enable computation of pressure, flow, and head loss throughout the network using conservation equations (Rossman 2000). Leak or burst events cause measurable deviations between observed and simulated states. By comparing SCADA data with model outputs, residuals can be analysed to infer anomalies (Savic et al. 2009). One classic example is the inverse transient analysis (ITA) approach, which models pressure transients caused by leaks to estimate their location and size (Misiunas et al. 2005; Palau et al. 2012). ITA uses optimisation algorithms to minimise differences between measured and simulated pressure heads over time, yielding leak parameters that best explain the observed data. While effective, the method requires precise knowledge of network topology, material properties, and boundary conditions. Its sensitivity to model uncertainty and computational demand under large-scale systems has led to various hybrid extensions (Covas & Ramos 2010; Kapelan et al. 2003).

Other model-based techniques include residual analysis, sensitivity analysis, and state estimation using Kalman filters (Mounce et al. 2003; Meseguer et al. 2015). These tools rely on the ability of

the hydraulic simulator to reproduce system dynamics accurately. Consequently, the quality of sensor data, calibration, and demand forecasting directly affects detection accuracy (Wu et al. 2016). Model-based leak detection offers strong interpretability and physical grounding. However, the complexity of calibrating real networks and the computational cost of running numerous scenarios limit its scalability (Puust et al. 2010). Therefore, recent research has shifted towards data-driven artificial intelligence and Machine Learning (AI/ML) techniques that learn directly from sensor data.

2.5.2 Artificial Intelligence-Based Leak Detection Methods

Artificial Intelligence (AI) approaches simulate human reasoning to recognise patterns and detect leaks automatically from multi-sensor datasets. Early AI systems used expert rules derived from operator experience (Mashford et al. 2009; Mounce & Boxall 2010). More recent developments combine big data analytics with Machine Learning algorithms to achieve autonomous fault detection and prediction (Wu et al. 2016; Joseph et al. 2023). AI-based detection systems generally follow three stages: (1) data acquisition from SCADA, AMI, or IoT sensors; (2) feature extraction, filtering, and signal transformation; and (3) classification or regression to identify abnormal behaviour (Li et al. 2015; Hu et al. 2021). These models can learn correlations between input variables (e.g., flow, pressure, vibration) and leak conditions, thus generalising across operating scenarios. Knowledge-based systems have been widely applied. For example, expert systems encode heuristic rules such as “if a sudden drop in pressure coincides with increased inflow, a burst may have occurred.” Such logic is implemented through decision trees or fuzzy inference systems (Mounce & Boxall 2010). Although interpretable, rule-based models depend heavily on expert formulation and often underperform in dynamic networks where behaviour varies over time.

The efficacy of techniques for detecting anomalies, particularly related to pipe failures, leaks, and bursts focusing on a case study in the Sunbury region of Victoria, Australia was investigated by Joseph et al. (2024). This study utilised SCADA sensor technology to extract hydraulic characteristics and developed algorithms for leak and burst detection. Both logic-based and Machine Learning approaches were employed. Logic-based algorithms excelled at capturing straightforward anomalies, while Machine Learning techniques enhanced detection by learning from historical data and adapting to changing conditions. Kammoun, Kammoun and Abid (2022) discussed hardware-based methods such as acoustic sensors, ground-penetrating radar, and fibre-

optic sensing, as well as software methods using pressure, flow, and acoustic data analysis. Similarly, Sourabh, Timbadiya and Patel (2023) investigated the use of Machine Learning techniques specifically Artificial Neural Networks (ANN) and Support Vector Machines (SVM) for detecting and locating leaks in water distribution networks. Their study demonstrated the potential of Machine Learning in improving leak detection and management. Zhou, Lau and Wang (2019) proposed a Machine Learning-based leak identification (ML-LI) method for water supply systems. Unlike conventional signal-propagation model (SPM) approaches that rely on deterministic models, this method handles system dynamics and uncertainties. It uses two data flows: (1) raw data fed directly into a deep neural network (DNN), and (2) data processed through a convolutional neural network (CNN) and pooling before DNN. A novel fusion-enhanced stochastic optimisation (FuSO) algorithm was introduced for DNN training, improving performance by extracting stable features from dynamic measurement samples.

2.5.3 Machine Learning Approaches

Machine Learning (ML) provides data-driven methods that automatically learn the mapping between input features and leak states without requiring explicit hydraulic equations. ML algorithms have been successfully applied to classify network states, detect anomalies, and localise leaks using SCADA, flow, and pressure datasets (Wu et al. 2016; Zhou et al. 2020; Joseph et al. 2023). ML-based methods can be divided into supervised and unsupervised categories depending on the availability of labelled training data. Recent research has increasingly focused on the application of Machine Learning techniques for leak detection in water distribution systems. Machine Learning models can identify complex, non-linear relationships in large datasets, making them particularly suitable for analysing SCADA data (Soldevila et al., 2016). For instance, data-driven approaches have been successfully applied to detect anomalies and predict pipe bursts by analysing historical and real-time pressure and flow data (Romano, Kapelan & Savic, 2014). These methods significantly improve detection accuracy compared to traditional rule-based systems.

Recent studies have further advanced this field with deep learning and generative models. For example, Rajabi et al. (2023) proposed a conditional deep convolutional generative adversarial network (GAN) for leak detection.

Graph-based Machine Learning approaches have also gained attention. A convolutional graph neural network model proposed by Wang et al. (2024) incorporates network topology and segment-level features, improving detection performance in systems with limited sensor coverage. Despite these advancements, challenges remain in terms of data availability, model interpretability, and real-world deployment. A recent systematic review by Li et al. (2025) highlights that while Machine Learning models show strong performance in controlled environments, their practical application is still constrained by data quality and system complexity.

Nevertheless, Machine Learning approaches also present challenges, including data quality requirements, model interpretability, and the need for large training datasets.

2.5.3.1 Supervised Machine Learning Approaches

Supervised learning algorithms use labelled datasets where examples of normal and leak conditions are provided. The model learns to distinguish between these classes and predicts new cases accordingly (Li et al. 2015; Mounce & Boxall 2010). Commonly used algorithms include decision trees, random forests, support vector machines (SVM), artificial neural networks (ANN), k-nearest neighbours (k-NN), and ensemble methods. Mashford et al. (2009) proposed one of the first ANN-based leak-detection frameworks for water distribution systems, using pressure and flow data to identify leak locations. The network achieved high classification accuracy but required careful pre-processing and training under multiple operating conditions. Similarly, Wu et al. (2016) applied random-forest models to SCADA pressure data for burst detection, demonstrating robust performance under noisy measurements. SVMs have also been applied to classify leakage events based on residual-pressure patterns (Hu et al. 2021). Their capacity to handle non-linear relationships makes them effective for complex network interactions. Ensemble approaches, such as AdaBoost and gradient boosting, combine multiple weak learners to improve prediction reliability (Joseph et al. 2023). Hybrid approaches integrate hydraulic modelling and Machine Learning, where simulated leak scenarios from EPANET serve as training data (Savic et al. 2009; Wu et al. 2016). Such synthetic datasets expand model robustness to rare or unobserved conditions. Furthermore, deep learning architectures, including convolutional and recurrent neural networks (CNNs, RNNs), have been employed to capture spatial-temporal dependencies in multi-sensor

data streams (Zhou et al. 2020). Deep networks can automatically learn hierarchical features, enabling accurate detection of subtle pressure anomalies. Despite high accuracy, supervised ML models depend on labelled data, which are often scarce in real utilities due to the infrequency of documented leaks. Simulated datasets help but may not capture the full variability of field conditions (Joseph et al. 2023). Overfitting and interpretability also remain major concerns for operational deployment.

Farajzadeh, Sadeghzadeh and Jokar (2024) used Machine Learning to predict when pipes in municipal water distribution networks require repair. Their methodology involved a dataset of 1467 records of burst pipes from Saveh, Iran. Three one-class classification algorithms OC-SVM, Isolation Forest, and Elliptic Envelope were trained on features such as pipe size, type, and location. The models were assessed using standard performance metrics, including accuracy, area under the receiver operating characteristic curve (AUC), and the F1-score. The AUC provides a comprehensive measure of a classifier's ability to discriminate between classes across all decision thresholds, with values closer to 1 indicating superior separability. The F1-score, defined as the harmonic mean of precision and recall, offers a balanced evaluation of a model's performance, particularly in scenarios involving imbalanced class distributions. Furthermore, Predict and Score Sample metrics were applied to both the original and balanced datasets to examine model consistency and robustness across varying data conditions. This approach could assist municipalities in proactively maintaining water infrastructure and reducing water waste.

McMillan, Fayaz and Varga (2023) presented a hybrid Machine Learning framework for detecting and predicting leakage in water distribution systems at the district metered area (DMA) level. It used one year of 15-minute interval flow data for over 2500 DMAs from Yorkshire Water, North Yorkshire, UK. The framework combined long short-term memory (LSTM) and recurrent neural networks (RNNs) for mean-flow forecasting with Kalman filters (KFs) for residual forecasting. For leakage detection, the Isolation Forest algorithm was employed to identify outliers in the flow data, which correlated with known repairs. Fagiani et al. (2016) proposed a statistical framework for detecting leaks in water and natural-gas grids. The authors used Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), and One-Class Support Vector Machines (OC-SVM) to characterise normal consumption patterns and identify anomalies. A set of features was extracted from consumption data, including temporal features, and a modified sequential feature-selection

algorithm was used to select the best features. The framework was tested on residential and building scenarios using two datasets with different time resolutions. Komba, Mathonsi and Owolawi (2024) proposed an algorithm combining SVM and Random Forest (RF) techniques to enhance leak detection and localisation in water distribution networks (WDNs), addressing challenges posed by background leakages. They presented an IoT-based architecture integrating sensors, wireless communication, and cloud computing for leak detection. The authors concluded that the SVM–RF approach offers superior accuracy and recommended further research on scalability and real-world implementation. Other studies have also evaluated various Machine Learning models for leak detection. For instance, Chi et al. (2021) investigated the effectiveness of Random Forest (RF) compared to other ML methods using experimental data from a laboratory pipeline setup. Similarly, Karadirek, Kaya-Basar and Akdeniz (2024) contributed to understanding pipe-failure risks to inform asset-management strategies.

Tijani et al. (2022) developed ML-based leak-detection models for real water distribution networks in Hong Kong using acoustic signals from noise loggers. They found that models trained on features from de-noised signals had higher classification accuracy than those using raw signals. Wu, Ma and Wang (2022) applied the XGBoost algorithm to identify leakage zones and predict leakage levels. Their methodology included simulating leakage scenarios with EPANET, sensor placement via fuzzy c-means clustering, and comparison with a back-propagation neural network (BPNN). XGBoost outperformed BPNN in both accuracy and computational efficiency, showing adaptability to different sensor configurations and a non-linear relationship between prediction accuracy and the number of leak points. Fares et al. (2023), Shen and Cheng (2022), and Yussif and Zayed (2023) developed models based on acoustic-emission data. Shen and Cheng (2022) found that AdaBoost and Random Forest outperformed Decision Trees, suggesting significant potential for automated leak detection. Yussif and Zayed (2023) used single noise loggers to predict leak distance and direction, identifying SVM and a transformed regression ensemble with k-NN as effective tools, particularly k-NN for leak direction. Li et al. (2022) proposed a hybrid model that integrates both data-driven and model-based approaches within the BattLeDIM framework. Their method used STL decomposition, k-means clustering, and stepwise fault diagnosis to detect and localise multiple leaks effectively, especially burst and large-scale events.

2.5.3.2 Unsupervised Machine Learning Approaches

Unsupervised learning algorithms detect anomalies without labelled data by identifying deviations from normal operational patterns (Mounce & Boxall 2010; Zhou et al. 2020). These methods are particularly valuable for large networks where leak events are rare or undocumented. Techniques such as clustering, principal component analysis (PCA), autoencoders, and Gaussian mixture models (GMMs) are used to model baseline behaviour. Deviations from the learned distribution signal potential leaks or bursts (Wu et al. 2016). For example, Mounce et al. (2003) applied self-organising maps (SOMs) to classify flow and pressure data into normal and abnormal categories, achieving real-time detection within SCADA environments. Recent advances in deep unsupervised learning employ variational autoencoders (VAEs) and long short-term memory (LSTM) networks to capture temporal correlations in streaming data (Zhou et al. 2020). These models reconstruct expected sensor values; significant reconstruction errors indicate anomalies. The advantage of such models lies in their adaptability to new data without explicit retraining or labelling. However, unsupervised methods can produce false positives during legitimate operational changes such as pump switching, demand variation, or pressure management adjustments. Therefore, hybrid strategies combining unsupervised detection with rule-based validation have been proposed (Joseph et al. 2023). These approaches balance sensitivity with robustness and are increasingly integrated into smart utility dashboards for real-time monitoring.

2.5.4 Comparative Analysis of Software Approaches

Software-based detection systems provide flexibility, scalability, and predictive capability superior to traditional hardware-only setups. They enable utilities to perform continuous monitoring, scenario testing, and automated alarm generation using SCADA or IoT data (Sharma et al. 2013). However, their reliability depends on data quality, model calibration, and computational efficiency. Model-based approaches offer strong interpretability but suffer from parameter uncertainty; ML-based approaches achieve higher adaptability but may lack transparency. Consequently, hybrid frameworks combining physical models, data-driven learning, and optimisation algorithms are emerging as the most effective solution (Joseph et al. 2023). Such frameworks enable leak detection, localisation, classification, and sensor placement optimisation within a unified digital twin environment which is the core of next-generation Intelligent Water Networks. The latest literature highlights a shift toward digital water systems, where sensing, analytics, modelling, and

automation are integrated into unified platforms. These systems enable real-time decision-making, predictive maintenance, and adaptive control (Pagano et al. 2025). Digital twin technologies are also emerging as a key component of intelligent water systems, providing virtual representations of physical infrastructure for simulation and optimisation. This transition represents a significant step toward sustainable and resilient water management.

2.6 Optimal Allocation of Pressure Sensors

The optimal allocation of pressure sensors is a critical design problem within intelligent water distribution systems. Since installing sensors across all nodes is economically impractical, researchers have sought methods to determine the most informative locations for monitoring leaks, bursts, and pressure anomalies (Sharma et al. 2013; Wu et al. 2016; Joseph et al. 2023). Sensor placement directly affects the accuracy, robustness, and computational efficiency of detection algorithms. Therefore, optimal sensor placement integrates hydraulics, network topology, uncertainty modelling, and information theory to achieve maximal detection performance with minimal instrumentation cost. The literature divides sensor placement methodologies broadly into two categories: optimisation-based approaches and hydraulics/geometry-based approaches (Puust et al. 2010; Kapelan et al. 2003). Optimisation-based methods employ mathematical programming or heuristic search to maximise specific objective functions, while hydraulics-based methods exploit network sensitivity, pressure gradients, or centrality measures derived from hydraulic simulations. Effective sensor placement is critical for ensuring accurate detection and monitoring in water distribution systems. Poorly placed sensors can lead to incomplete coverage and reduced detection performance. Research has explored various optimisation techniques to determine optimal sensor locations, aiming to maximise detection sensitivity while minimising deployment costs (Puust et al., 2010). These approaches consider factors such as network topology, hydraulic behaviour, and uncertainty in demand patterns. Despite progress in this area, many existing methods do not fully integrate real-time data analytics or Machine Learning capabilities, limiting their effectiveness in dynamic network conditions. Recent studies emphasise the importance of integrating sensor placement with data-driven methods. Rodriguez et al. (2025) demonstrate that observability-based optimisation significantly improves leak detection performance, particularly in large and complex networks. These findings suggest that sensor placement should not be treated as an independent problem but rather as a component of an integrated monitoring and detection framework.

2.6.1 Optimisation-Based Approaches

Optimisation-based formulations treat sensor placement as a mathematical problem, where the objective is to minimise detection uncertainty or maximise leak-detection probability under given constraints (Kapelán et al. 2003; Wu et al. 2016). One early example is the single-objective optimisation of leak-detection performance using sensitivity matrices derived from pressure deviations at each node. Kapelán et al. (2003) introduced a genetic-algorithm-based method that maximised the distinguishability of leak events by selecting pressure-monitoring nodes that yield unique hydraulic signatures. The fitness function quantified the mutual information between leak events and sensor responses. Subsequent studies extended this to multi-objective optimisation, balancing competing goals such as detection accuracy, cost, and reliability (Casillas et al. 2013; Wu et al. 2016). Multi-objective evolutionary algorithms (MOEAs), including the Non-dominated Sorting Genetic Algorithm II (NSGA-II) and multi-objective particle swarm optimisation (MOPSO), have been widely applied (Sharma et al. 2013). These algorithms evolve a population of sensor configurations to approximate the Pareto front of trade-offs between performance and cost. Information-theoretic formulations compute metrics such as entropy, mutual information, and Fisher information to quantify the expected gain from each sensor location (Soldevila et al. 2016). The optimal configuration maximises global information coverage while minimising redundancy among sensors. Another key development involves Bayesian optimisation and Markov Chain Monte Carlo (MCMC) simulation to incorporate uncertainty in demands, leak magnitudes, and measurement errors (Wu et al. 2016). Bayesian methods yield probabilistic estimates of detection likelihood rather than deterministic configurations. This probabilistic framework is essential for real-world networks where demand fluctuations and noise can obscure leak signals. The emergence of reinforcement learning (RL) and deep-learning-based optimisation has introduced adaptive sensor placement strategies (Joseph et al. 2023). In RL-based formulations, an agent iteratively learns the most rewarding sensor locations based on leak-detection accuracy feedback, effectively adapting to changes in network hydraulics. Such data-driven optimisation provides flexibility that traditional static models lack. Optimisation approaches can also be hybridised with hydraulic simulation platforms such as EPANET, generating thousands of leak scenarios under varying operating conditions. The resulting pressure response matrix forms the input for search algorithms (Sharma et al. 2013). For example, Joseph et al. (2023) demonstrated an optimisation framework combining sensitivity analysis and degree centrality weighting to rank

candidate nodes, achieving robust sensor configurations with minimal redundancy. Despite their robustness, optimisation-based approaches are computationally demanding and sensitive to model fidelity. The dimensionality of the problem increases exponentially with network size, necessitating heuristic algorithms rather than exhaustive search. Nonetheless, their flexibility and ability to incorporate multi-criteria decision-making make them a cornerstone of modern intelligent water-network design.

2.6.2 Hydraulics and Geometry-Based Approaches

Hydraulics- and geometry-based methods select sensor locations using network-topology characteristics or hydraulic sensitivities without explicit optimisation loops (Wu et al. 2016; Sharma et al. 2013). These methods are computationally efficient and suitable for utilities seeking rapid initial deployment. Hydraulic sensitivity analysis evaluates how pressure at each node responds to a hypothetical leak or burst at another node. Nodes exhibiting the largest pressure deviations across multiple leak scenarios are prioritised for sensor installation (Kapelan et al. 2003). Sensitivity matrices derived from EPANET simulations can quantify detection coverage and inter-node redundancy (Joseph et al. 2023). Graph-theoretic metrics such as degree centrality, betweenness, and closeness are also employed (Yazdani & Jeffrey 2011). High-centrality nodes, located at critical junctions or flow corridors, offer strong network observability and are thus preferred sensor sites. Combining hydraulic sensitivity with topological centrality provides a dual-perspective framework: physical and structural for sensor prioritisation (Sharma et al. 2013). Cluster-based methods partition large networks into smaller zones or district metered areas (DMAs), selecting representative nodes per cluster. These approaches rely on pressure-correlation analysis or k-means clustering of hydraulic responses (Wu et al. 2016). By ensuring that each cluster is represented by at least one high-sensitivity node, utilities achieve balanced spatial coverage with reduced redundancy. Hydraulics-based strategies are simple and interpretable, but they can overlook complex nonlinear interactions among leaks, valves, and boundary conditions. Hence, recent studies integrate them with optimisation or Machine Learning techniques to achieve both computational efficiency and high accuracy (Joseph et al. 2023).

2.6.3 Integrated Intelligent Water Network Perspective

In the broader context of Intelligent Water Networks, optimal sensor allocation forms the physical foundation for data-driven analytics. The sensors act as the eyes and ears of the network, capturing

pressure, flow, and quality data that feed into hydraulic simulators, anomaly-detection algorithms, and decision-support systems (Sharma et al. 2013).

Integrated IWN architecture typically includes four layers:

1. Sensing layer – physical devices such as pressure, flow, and acoustic sensors.
2. Communication layer – wired or wireless protocols (SCADA, LoRaWAN, NB-IoT) that transmit sensor data.
3. Processing layer – cloud or edge computing systems performing hydraulic simulation and ML-based analysis.
4. Application layer – dashboards, alarms, and control strategies for operators. Within this framework, the placement of sensors determines not only the accuracy of leak detection but also the responsiveness of control actions such as valve closure or pump adjustment (Ramos et al. 2019). Smart optimisation ensures that a small number of strategically located sensors yield maximal information coverage, supporting both operational efficiency and long-term asset management.

2.6.4 Summary of Findings

Across the literature, both optimisation- and hydraulics-based sensor placement methods converge on the same principle: a well-chosen subset of pressure nodes can achieve near-complete leak observability (Sharma et al. 2013; Joseph et al. 2023). The integration of data-driven algorithms with hydraulic insight enables robust, scalable sensor networks for real-time anomaly detection. Challenges remain in handling data uncertainty, adapting to network reconfiguration, and balancing cost against reliability. Nonetheless, advances in AI, IoT, and distributed computing are closing the gap between research prototypes and practical utility implementation. The combined evolution of hardware, software, and optimisation techniques underpins the realisation of sustainable, resilient, and intelligent water distribution systems.

2.7 Chapter Summary and Reflections

This chapter has provided a comprehensive examination of the evolution of leak and burst detection in water distribution systems, tracing the technological transition from traditional,

manually intensive practices to modern intelligent and data-driven frameworks. It began by analysing the limitations of conventional techniques such as visual inspections, acoustic correlation, and minimum night-flow assessments, which, while foundational, often lack the sensitivity, automation, and predictive capacity needed for contemporary urban water-management demands. The discussion then expanded to hardware-based innovations, including pressure and flow sensors, acoustic emission loggers, fibre-optic sensing, distributed temperature and acoustic monitoring, wireless sensor networks, and satellite and drone-based systems. These physical instruments represent a fundamental shift from reactive maintenance toward continuous, quantitative, and spatially distributed observation of network health. However, their effectiveness is inseparably linked to calibration, placement, and integration within broader decision-support systems.

Equally significant is the evolution of software-based approaches, which transform the raw signals obtained from sensors into actionable intelligence. Hydraulic modelling, inverse transient analysis, residual and sensitivity analysis, and state estimation frameworks have provided physically grounded interpretations of leaks and bursts, though they are often constrained by computational complexity, model uncertainty, and calibration requirements. The integration of artificial-intelligence and Machine Learning techniques particularly support vector machines, neural networks, ensemble learning, and deep architectures has further advanced the field by enabling adaptive, self-learning systems capable of recognising complex, nonlinear patterns in hydraulic and acoustic data. By combining hydraulic-model interpretability with Machine Learning adaptability, hybrid frameworks now offer utilities both precision and flexibility for real-time decision-making. Such integration forms the analytical core of Intelligent Water Networks, converting diverse telemetry into predictive insights that support early intervention and asset optimisation.

A central conclusion arising from this review is that neither hardware nor software alone can ensure effective leak and burst management. Reliable detection and localisation require the convergence of physically based sensing, mathematically sound modelling, and data-driven inference, all supported by robust communication infrastructure that can withstand sensor noise, latency, and operational variability. The synergy between these components rather than their independent performance determines the overall efficiency of an intelligent water-network system.

Conventional methods remain indispensable for field confirmation and low-cost screening, yet they must be complemented by automated and predictive technologies to achieve comprehensive network visibility.

Furthermore, this chapter emphasised the crucial role of optimal pressure-sensor allocation in maximising detection accuracy and network observability. A well-selected subset of nodes can achieve near-complete leak coverage when guided by optimisation-based, information-theoretic, and hydraulics- or geometry-based frameworks. The literature consistently demonstrates that strategic placement, informed by sensitivity analysis, centrality metrics, or evolutionary algorithms, yields substantial efficiency gains by reducing redundancy while preserving diagnostic reliability. Consequently, sensor placement emerges not merely as a technical configuration problem but as a determinant of system-wide performance, influencing the responsiveness of control actions such as valve operations, pump scheduling, and emergency isolation.

From an operational perspective, several overarching reflections can be drawn. Uncertainty and robustness remain critical concerns, as leak-detection performance is highly sensitive to variations in water demand, sensor drift, and parameter estimation. Therefore, probabilistic approaches that incorporate Bayesian inference or adaptive thresholds are increasingly preferred over deterministic methods. Comparability across studies also poses an ongoing challenge, given the heterogeneity of datasets, metrics, and experimental conditions. The use of benchmark networks provides partial standardisation but cannot replace real-utility trials that expose algorithms to complex field realities. Integration and interoperability are equally vital: the greatest value of intelligent networks lies in their seamless connection between sensing hardware, communication protocols, analytical models, and operational decision support, enabling automated alarm generation, fault triage, and maintenance scheduling.

Economic realism constitutes another layer of reflection. The deployment of advanced sensors and analytics must be justified not only by technical performance but also by lifecycle costs and the potential for synergistic benefits such as pressure management and energy recovery. The literature shows that leak reduction and energy optimisation are interdependent; properly designed systems can simultaneously achieve non-revenue-water reduction and energy savings through combined strategies like Pumps-as-Turbines. Data governance, cybersecurity, and telemetry reliability also demand attention, as the increasing digitalisation of water infrastructure introduces new

vulnerabilities and ethical responsibilities regarding data ownership and system resilience.

In summary, this chapter underscores that the path toward sustainable, resilient, and intelligent water distribution systems depends on integrating conventional, hardware, software, and optimisation-based methods into a cohesive framework. Intelligent Water Networks represent not a single technology but an evolving ecosystem of interlinked layers sensing, communication, analytics, and decision-making that collectively enhance system awareness, predictive capability, and operational efficiency. Despite the progress outlined here, significant challenges persist in handling uncertainty, reducing false alarms, and achieving economically scalable implementation across diverse network typologies.

The findings and reflections presented in this chapter directly inform the experimental design and validation environments introduced in the next part of this thesis. Chapter 3: Case Study Sites and Description therefore build upon these conceptual and methodological foundations by presenting three complementary systems: the real-world Sunbury pumping main, the Melton experimental prototype, and the benchmark Hanoi Water Distribution Network (WDN). These case studies provide the empirical basis for evaluating the proposed detection and optimisation methodologies under realistic hydraulic, operational, and topographical conditions, thereby linking the theoretical insights of Chapter 2 to practical implementation and validation in the chapters that follow.

Chapter 3

Case Study Sites and Description

3.1 General Remarks

This chapter presents the detailed description of the case study sites and network systems used in this research for the development and validation of the proposed leak and burst detection methodologies. The selection of these sites was guided by both practical relevance and data availability. Three systems were examined: (i) a real-world pumping main system in the Sunbury region, (ii) an experimental prototype water- supply setup located at the Melton training yard, and (iii) the benchmark Hanoi water distribution network (WDN) model. Together, these systems provided a comprehensive platform that captures the hydraulic, operational, and topographical diversity encountered in real and simulated urban water networks. The Sunbury pumping main system facilitated the investigation of leak and burst detection under real operational conditions using Supervisory Control and Data Acquisition (SCADA) data, whereas the Melton experimental setup enabled controlled replication of leaks and bursts under various hydraulic states. The Hanoi WDN was employed for simulation-based studies on pressure sensitivity and optimal sensor placement, thereby linking the experimental and numerical aspects of the research. The collective use of these three case studies ensures that the developed methodologies are robust, scalable, and transferable across both field and laboratory conditions.

3.2 Pumping main system

The Sunbury region, located approximately 40 km northwest of Melbourne’s Central Business District, lies within the Maribyrnong River catchment, which originates from the Macedon Ranges and drains into Port Phillip Bay (Bureau of Meteorology, 2023; Melbourne Water, 2021). Sunbury is characterised by a semi-urban environment with increasing residential development and a strong reliance on water supply infrastructure managed by local water utilities (Hume City Council, 2022).

The catchment comprises a mix of rural and suburban land uses, with elevations ranging between approximately 200 m and 500 m above mean sea level (AMSL) (DELWP, 2022).

Due to population growth and urban expansion, the Sunbury region has experienced increasing demand for water supply services and associated challenges related to water losses from leaks and bursts in ageing pipeline systems (Greater Western Water, 2021; Liemberger & Wyatt, 2019). A water supply pumping main system in the Sunbury region, Victoria, Australia, has therefore been selected as the case study system, considering its varying elevation profile, pipeline configuration, and availability of operational data for the development and validation of leak and burst detection methodologies. Figure 3.1 illustrates the geographical location of the case study area within the Sunbury region, Victoria, Australia.

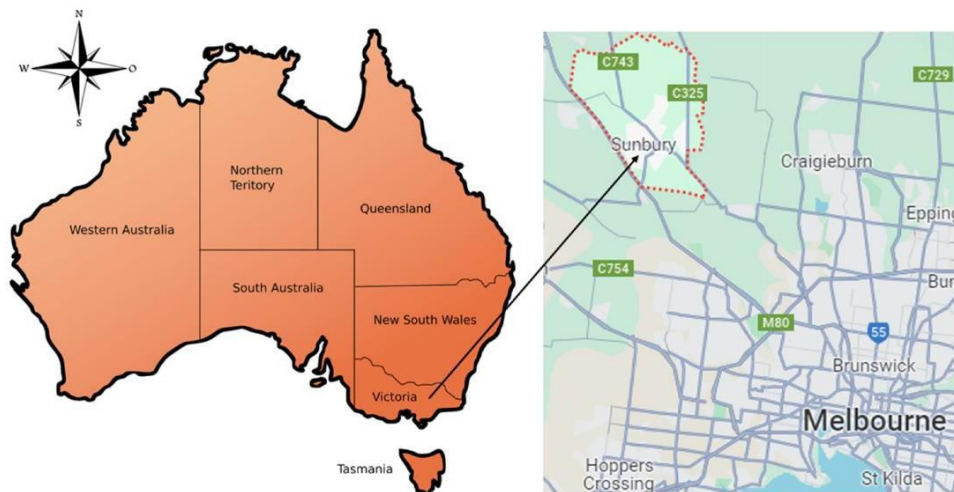


Figure 3.1: Map of the location of the case study site in the Sunbury region, Victoria, Australia. Adapted from Australia map (Pixabay, n.d.) and Google Maps (Google, n.d.).

Water is taken from the Melbourne water tank to the Shepherds Lane tank through two pipelines of 450 mm diameter and 375 mm diameter. There is a discharge pressure sensor and flow meter sensor in each pipeline. A divide valve separates the two pipelines.

In this research, two methodologies are developed. One is the logic-based method, and the other is the Machine Learning-based approach. The overall study is shown in the flow chart below in Figure 3.2.

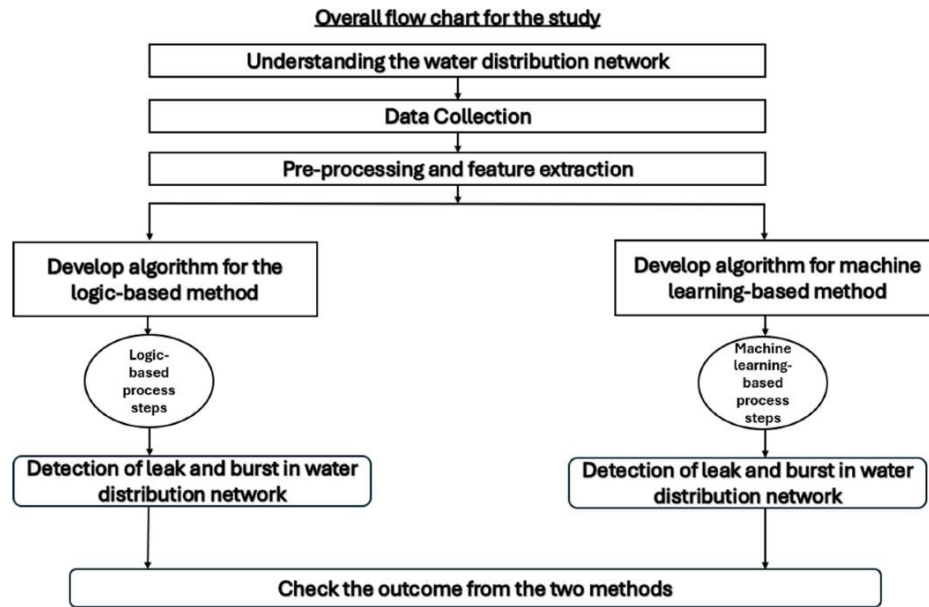


Figure 3.2: An overall flow diagram of the leak and burst detection in the water distribution network.

Figure 3.3 presents the simplified detailed diagram of the case study pumping water supply system with a 450 mm pipeline to water tank. Water is pumped directly from the trunk water main to a tank through a pipeline of 450 mm diameter. There is a pressure sensor and flow meter sensor in the pipeline. The distance between the pumping station and the tank is 6.24 km. The research methodology was developed using data from flow meter, pressure sensor, and pump speed recorder. The pumping station continuously monitors the water flow, pressure, and pump speed and the capacity of the tank is 0.8ML. The maximum water level in the tank is 3.72m (From the extracted SCADA information). The elevations at the pumping station and tank are 59m. Figure 3.4 presents the Panoramic photo of the Loemans road pumping station and Shepherds Lane tank. In this research, two complementary methodologies, namely logic-based and Machine Learning-based approaches, are adopted for leak and burst detection in water distribution systems. The logic-based method is grounded in hydraulic principles and employs threshold-based rules derived from pressure, flow, and operational behaviour. This approach offers high interpretability and computational efficiency, making it well suited for systems where hydraulic behaviour can be reliably characterised. This method is more suitable for single pumping and gravity transmission mains fitted with flow and pressure sensors. However, real-world water distribution systems often

exhibit complex and non-linear interactions due to variations in demand patterns, operational conditions, and measurement noise. To address these challenges, Machine Learning techniques are incorporated to identify subtle anomalies and patterns that may not be effectively captured using deterministic, rule-based approaches.

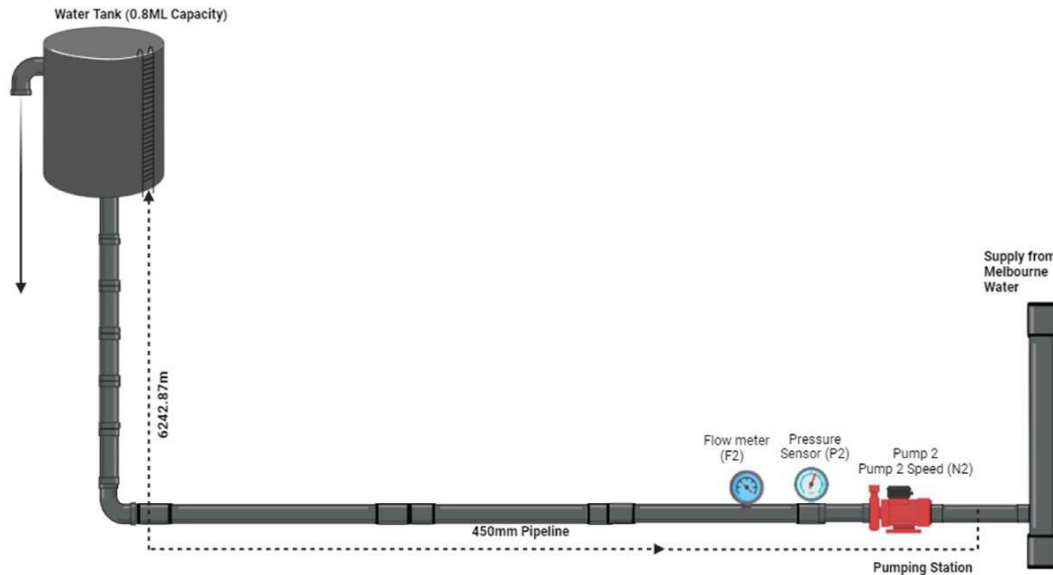


Figure 3.3: Simplified detailed Schematic diagram of the 450 mm pipeline.



Figure 3.4: Panoramic Photo of the Loemans Road Pump Station (Left) and Shepherds Lane tank (Right)

3.3 Prototype water supply system at Melton Training Yard

The Melton region, located approximately 35 km west of Melbourne, is a rapidly growing urban area within the Werribee River catchment (Bureau of Meteorology, 2023; Melbourne Water, 2021). Melton's water infrastructure is managed by local water utilities, supporting increasing residential and peri-urban development (City of Melton, 2022). The catchment comprises urban and peri-urban

land uses, with elevations ranging from approximately 100 m to 250 m above mean sea level (AMSL) (DELWP, 2022). The Werribee River originates from the Wombat State Forest and flows into Port Phillip Bay (Bureau of Meteorology, 2023).

The Melton region experiences a temperate climate with annual rainfall typically ranging between 500–600 mm, predominantly occurring during winter months (Bureau of Meteorology, 2023). Rapid urbanisation in Melton has led to increased water demand and additional pressure on existing water distribution infrastructure (City of Melton, 2022).

Common challenges in the region include pipeline leaks and bursts associated with soil movement, particularly in reactive clay soils characteristic of the area, as well as stresses induced by urban infrastructure expansion (DELWP, 2022; Puust et al., 2010). These factors contribute to non-revenue water losses in the distribution system (Liemberger & Wyatt, 2019).

An experimental prototype water supply system setup is located in the Butlers Road, Mount Cottrell, Melton, Victoria, Australia as shown in Figure 3.5.

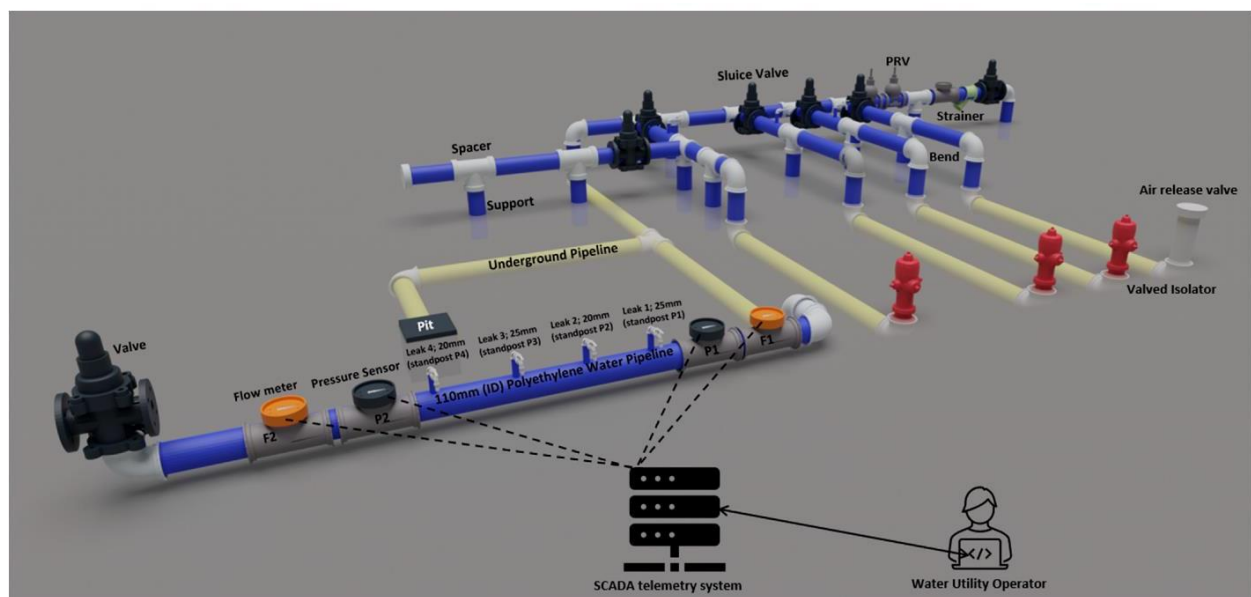


Figure 3.5: Simplified diagram of the Melton experimental site, Butlers Road, Mount Cottrell, Melton region.

A 110 mm internal diameter, 110 m length polyethylene pipe was equipped with two flow meters

and two pressure sensors. The pipe is connected to a 100 mm water pipeline. Leaks were simulated using various pipe outlets of sizes 20 mm and 25 mm from 110 mm pipe (Figure 3.5). The main pipe size of 110 mm was selected considering the available flow at the experimental site. For 110 mm pipe size, 20 mm and 25 mm pipe outlets were considered sufficient enough to provide the leak conditions. The flow meters and pressure sensors were integrated into a SCADA wireless system to monitor and record the experimental data. In Figure 3.5, the inlet flow meter is F1, the outlet flow meter is F2, the inlet pressure sensor is P1, and the outlet pressure sensor is P2.

The study integrates the key parameters such as flow rate, pressure variations, addition of the change in pressure, and flow variations as pressure and flow acceleration, respectively. Timestamp data are also included to establish a detailed understanding of the pipeline dynamics under different operating conditions, and finally, the best performance models are found based on the data collected, which are then used to detect leaks and bursts in the water pipeline.

The following instruments were used in the prototype water system:

(a) Pressure Transducer

A pressure transducer (Model: MX232, Measurex Pty Ltd., Melbourne, Australia; Serial No: 3115262003) was employed to monitor pressure within the pipeline system. The device features a measurement range of 0–16 bar (0–163m) and a manufacturer-specified accuracy of $\pm 0.25\%$ of Full Scale (FS). It was installed using PTFE (polytetrafluoroethylene) thread sealing and securely interfaced with the data acquisition system to ensure reliable measurements. The sensor was factory calibrated prior to installation, and functional validation was conducted through comparison with a certified reference pressure gauge. The observed measurement deviations remained within the stated accuracy range, thereby confirming the sensor's reliability under field conditions.

(b) Pressure Transmitter

In addition to the pressure transducer, an ABB 261GSU SMART pressure transmitter (Type: 261GSU/VT1/N21) was utilised for high-precision pressure monitoring. The instrument is equipped with a silicon oil-filled Hastelloy C276 diaphragm and has a configurable measurement span of 150 kPa. The manufacturer is ABB Ltd., Zürich, Switzerland. The transmitter operates on a 4–20 mA loop-powered output with a supply voltage range of 10.5–32 V DC. According to the manufacturer's specifications, the device offers a standard accuracy of $\pm 0.1\%$ of span, with an

optional high-accuracy mode of up to $\pm 0.04\%$ of span. Prior to deployment, the transmitter underwent comprehensive field calibration using a certified pressure calibrator traceable to NIST standards.

(c) Mechanical Flow Meter

A volumetric multijet dry dial water meter (Serial No: 2404013684), compliant with International Organisation for Standardisation (ISO) 4064 standards (ISO 4064-1:2014), was used to measure volumetric flow during the experiment. The ISO organisation is based in Geneva, Switzerland. The manufacturer is Zenner International GmbH & Co. KG, Saarbrücken, Germany. The device is equipped with a pulse output to facilitate continuous data logging. It provides an accuracy of $\pm 2\%$ for flow rates above the transitional threshold and $\pm 5\%$ below it. The meter was delivered factory-calibrated and accompanied by a compliance certificate. It was installed with tamper-proof sealing to maintain data integrity. A preliminary verification test was conducted to confirm operational performance and the consistency of the pulse signal under test conditions.

(d) Electromagnetic Flow Meter

To complement the mechanical measurements, a Siemens SITRANS FM MAG 8000 electromagnetic flow meter was deployed for continuous volumetric flow monitoring. The manufacturer is Siemens AG, Munich, Germany. This battery-powered inline device is factory-calibrated to ISO/IEC 17025 standards [24] and provides a measurement accuracy of $\pm 0.4\%$ of the actual flow rate. For enhanced accuracy, certain model variants support precision up to $\pm 0.2\%$. The meter incorporates SENSORPROM™ technology, which ensures the preservation of calibration and configuration data across installations. The device was deployed in its original calibrated state. Real-time flow data were retrieved via the onboard display and transmitted through either pulse output or an integrated SCADA telemetry system. The SCADA data were recorded at a fixed interval of one minute, corresponding to a sampling time of 60 s and a sampling frequency of 0.0167 Hz.

Mechanical flow meters were used to check the flow from four leaks as shown in Figure 3.5. Electromagnetic flow meters were used to measure flow at F1 and F2. Pressure transducers were used to measure pressure P1 and P2, and a pressure transmitter was used to validate the pressure transducers.

3.4 Hanoi Water Distribution System as case study to identify preferred locations of pressure sensors

The Hanoi water supply system was as the case study for pressure- sensor placement analysis because of the availability of benchmark network data. Urban water distribution networks play a pivotal role in ensuring continuous, safe, and reliable water supply to metropolitan populations. However, ageing infrastructure, rapid urbanisation, and increased water demand impose significant stress on these systems, particularly in developing cities like Hanoi, Vietnam. Water losses due to leaks and pipe bursts collectively referred to as non-revenue water (NRW) present critical operational challenges. The capital city Hanoi experiences a humid subtropical climate with annual rainfall ranging between 1,700 mm and 1,800 mm. The monsoon season (May to October) contributes approximately 80% of the annual rainfall. These climatic conditions, along with high population density, accelerated urbanisation, and industrial growth, compound the pressure on the city's water infrastructure. Much of the existing pipeline network is aged and vulnerable to corrosion, fractures, and mechanical failure. Leak detection and early warning mechanisms have thus become imperative. Traditional leak detection methods, though useful, are often reactive and labour-intensive. Advanced pressure monitoring using strategically placed sensors offers a more proactive solution. However, due to economic constraints and infrastructure limitations, sensors must be placed at the most critical locations nodes that exhibit the highest-pressure sensitivity to leak scenarios. Due to economic considerations, it is not possible to place advanced pressure sensors at all the nodes (junctions) of a water distribution system for pipe leak and burst detection. This research introduces a systematic approach to pressure sensitivity analysis using US EPANET (U. S. Environmental Protection Agency) for simulating leaks of various sizes at all the nodes of the network. The study aims to identify the most responsive nodes under simulated leak conditions, providing a sensitivity matrix that aids in sensor prioritisation (Fujiwara & Khang, 1990; Perelman, Krapivka & Ostfeld 2013).

A 32 node Hanoi water network model in EPANET shown below Figure 3.6. (Perelman, Krapivka & Ostfeld 2013). The Hanoi Water Distribution Network (WDN), first introduced by Fujiwara and Khang (1990), is one of the most widely used benchmark systems for testing water distribution optimisation methods. It is a gravity-driven network with three loops, comprising 34 pipes and 32

nodes, and receives supply from a reservoir situated at an elevation of 100 m. The system has a total pipe length of 39.420 km, with available commercial pipe sizes ranging between 304.8 mm and 1,016 mm, and corresponding unit costs from \$45.73/m up to \$278.30/m. Each demand node in the network is assigned a water requirement (uniform in the benchmark specification), while a minimum pressure head of 30 m is enforced throughout.

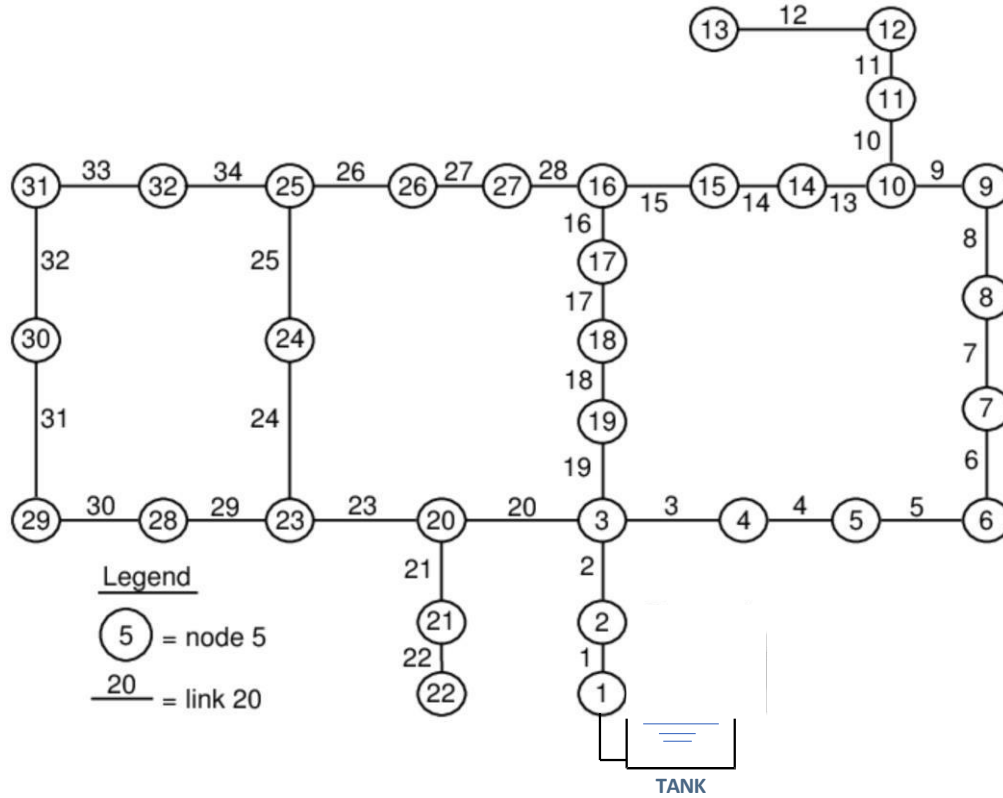


Figure 3.6: Hanoi Case study water network in EPANET (supply point at node1) (Perelman, Krapivka & Ostfeld 2013).

Table 3.1 summarises the principal characteristics of the Hanoi WDN, which has been widely utilised as a benchmark model for evaluating the hydraulic and economic performance of optimisation methodologies in water distribution system design. Uniform nodal demands are assigned at all 32 consumer nodes, in accordance with the original benchmark configuration developed by Fujiwara and Khang (1990). Due to its well-defined structure, simplicity, and documented parameter set, the Hanoi WDN has been widely adopted as a benchmark model in water distribution system research for evaluating hydraulic analysis methods and optimisation algorithms (Fujiwara & Khang, 1990; Perelman, Krapivka & Ostfeld, 2013; Ostfeld et al., 2008).

Table 3.1: Summary of the Hanoi water distribution network (WDN) characteristics (Fujiwara & Khang, 1990; Perelman, Krapivka & Ostfeld 2013)

Parameter	Value / Description
Total number of pipes	34
Total number of nodes	32
Reservoir (source node)	1 (located at an elevation of 100 m)
Number of loops	3
Total pipe length	39.420 km
Available pipe diameters	304.8, 406.4, 508.0, 609.6, 762.0, 1,016.0 mm
Pipe cost range	\$45.73/m – \$278.30/m
Minimum pressure head	30 m
Nodal demands	Assigned at all 32 demand nodes (uniform in the benchmark dataset)

Table 3.2 presents the nodal elevation and base demand data used in the hydraulic modelling of the Hanoi WDN (Fujiwara & Khang, 1990; Palod, Prasad & Khare, 2020). This uniform demand distribution facilitates comparative analysis and model validation across different optimisation studies. The total system demand is therefore 5613.22 L/s, which is supplied by the reservoir under a constant head boundary condition. The data for the ground levels are essential for any water supply network to estimate the terminal pressure at the nodes. The last two columns are essential to modelling and thus have not been removed. In this case study example, the input node (source node) has an elevation of 100 m and all other nodes are on a flat topography at considered 0 m elevation. Thus, the elevation difference between source node and demand nodes is 100 m for hydraulic calculations.

Table 3.2: Nodal elevation and base demand data for the Hanoi water distribution network (WDN)

Pipe No.	Node 1	Node 2	Length (m)	Flow (L/s)	Node	Elevation(m)
1	1	2	1000	5613.22	1	100
2	2	3	800	5366.00	2	0
3	3	4	970	2274.05	3	0
4	4	5	450	2237.94	4	0
5	5	6	950	1602.00	5	0
6	6	7	870	1441.50	6	0

7	7	8	1300	1039.80	7	0
8	2	9	1030	4720.00	8	0
9	9	10	980	3750.25	9	0
10	10	11	1200	2810.40	10	0
11	11	12	350	2149.90	11	0
12	12	13	800	1430.30	12	0
13	13	14	1170	789.60	13	0
14	9	15	780	2920.50	14	0
15	15	16	730	2220.40	15	0
16	16	17	860	1520.60	16	0
17	17	18	800	820.10	17	0
18	3	19	830	3230.00	18	0
19	19	20	600	2420.00	19	0
20	20	21	780	1670.50	20	0
21	21	22	920	990.20	21	0
22	22	23	640	430.10	22	0
23	19	24	990	1900.00	23	0
24	24	25	780	1220.00	24	0
25	25	26	910	690.50	25	0
26	26	27	530	450.30	26	0
27	24	28	720	950.00	27	0
28	28	29	450	520.00	28	0
29	29	30	720	380.00	29	0
30	19	31	860	1090.00	30	0
31	31	32	450	600.00	31	0
32	8	14	1230	790.00	32	0
33	14	22	840	720.00		0
34	7	23	910	690.00		0

From a methodological standpoint, the sensor placement methodology and different algorithms are described in Chapter 7. These collectively define the optimisation framework used for the subsequent analyses. They provide a robust comparative basis for assessing trade-offs between accuracy, coverage, and computational efficiency. Their practical implementation and performance benchmarking within the Hanoi WDN are detailed in Chapter 7, where results are quantitatively compared under noise-perturbed and cross-validated conditions.

3.4.1 Computational Environment

All simulations, benchmarking experiments and sensitivity analyses were executed on a dedicated desktop workstation. The system was equipped with an AMD Ryzen 9 3900X 12-core processor (3.80 GHz base frequency), 32 GB RAM, and a 1 TB Sabrent Rocket 4.0 NVMe SSD for fast sequential read/write performance. Graphical hardware consisted of an NVIDIA GeForce RTX 3070 Ti (8 GB VRAM); however, all sensor-selection algorithms were computed exclusively on the CPU with no GPU acceleration. The operating system was a 64-bit Windows 10 installation running on an x64-based architecture. This configuration ensured stable, reproducible runtime measurements and provided sufficient computational resources for executing Python-based hydraulic simulations, data pre-processing and optimisation routines. In terms of computational performance, individual node EPANET simulation runs were completed within 5-10 seconds per scenario.

3.5 Chapter Summary and Reflections

This chapter introduced and characterised the case study systems forming the empirical foundation of this research. The Sunbury case study described a 6.24 km, 450 mm-diameter pumping main equipped with flow, pressure, and pump-speed sensors used for extracting operational data. The Melton prototype system, consisting of a 110 mm polyethylene pipe instrumented with flow meters and pressure sensors, provided a controlled environment for simulating leaks of varying magnitudes. The third case study, the 32-node, 34-pipe Hanoi WDN, was adopted to perform hydraulic simulations, sensitivity analyses, and optimal pressure-sensor placement using EPANET. Collectively, the data obtained from these systems underpins the subsequent methodological development presented in Chapter 4, where both logic-based and Machine Learning approaches are implemented. The combination of real-world, experimental, and simulated environments provides a rigorous framework for model calibration, validation, and performance benchmarking of the proposed leak and burst detection algorithms.

Chapter 4

Research Methodology

4.1 General Remarks

This chapter outlines the research methodology adopted for the development and validation of the proposed leak and burst detection framework. The methodology was demonstrated through two complementary case studies designed to assess the system's performance under both real-world and controlled experimental conditions.

The first case study involved a field-scale implementation within a SCADA enabled water distribution network operated by Greater Western Water (GWW). This network provided real-time, high-frequency hydraulic data that facilitated the evaluation of the logic-driven leak and burst detection algorithm under actual operating scenarios (Figure 3.3). The second part of the first case study is the software-based leak detection system.

The second case study comprised a prototype water distribution system constructed to replicate essential hydraulic characteristics of municipal pipelines under controlled conditions (Figure 3.5). This setup enabled the testing of both software and hardware elements of the proposed framework by introducing artificial leaks and bursts to validate detection accuracy and system responsiveness.

Together, these two case studies established a dual-validation platform, allowing a rigorous examination of the developed methodologies in terms of detection precision, latency, and operational reliability. Comparative analysis between the field-scale and prototype-scale results provides a comprehensive understanding of system behaviour, scalability, and adaptability across diverse network environments.

The structure of this chapter was therefore designed to present the integrated methodological approach from logic-based detection to Machine Learning-based anomaly analysis culminating in the formulation of a hybrid framework for intelligent leak and burst detection. The subsequent Chapter 5 (Data Collection) presents detailed procedures and datasets acquired

from both case studies, while Chapter 6 (Application of Methodology) demonstrates the practical implementation, performance evaluation, and analytical outcomes of the proposed approach.

4.2 Leak and Burst Detection Methodology

This section introduces the foundational framework for detecting leaks and bursts within water distribution systems, responding to the growing need for intelligent, infrastructure management. Prior research has extensively investigated leak and burst detection in water distribution systems using a range of approaches, including acoustic methods, hydraulic modelling, and data-driven techniques. Conventional methods such as acoustic leak detection are widely used but are often labour-intensive and reactive, limiting their effectiveness for real-time monitoring (Puust et al., 2010). More recent studies have explored the use of SCADA-based data and Machine Learning algorithms to detect anomalies in pressure and flow signals, enabling improved detection performance under varying operational conditions (Romano et al., 2014; Soldevila et al., 2016). Despite these advancements, challenges remain in distinguishing between normal demand variations and actual leak or burst events, particularly in complex water distribution systems. These limitations highlight the need for an integrated framework that combines hydraulic knowledge with data-driven techniques for reliable and scalable leak and burst detection.

The proposed methodology integrates experimental validation, numerical modelling, and data analytics, forming a hybrid approach that bridges both physical and digital domains. Central to this framework is the utilisation of SCADA systems, which serve as the backbone for acquiring high-resolution operational data. The methodology incorporates a logic-driven decision layer to interpret hydraulic anomalies, complemented by advanced Machine Learning techniques to improve the classification accuracy of transient events. Together, these components constitute a multi-layered detection system designed to enhance the responsiveness, reliability, and predictive capability of smart water networks. In general, a leak can be defined as an unintended, typically small opening or defect in a pipe that allows water to escape gradually, often remaining undetected for extended periods. In contrast, a burst refers to a sudden and significant pipe failure that results in rapid water loss, pressure drops, and potential disruption to water supply services (Puust et al., 2010; Romano et al., 2014). Such events can have considerable impacts on system operation and may cause inconvenience to the public, including service interruptions and traffic disruptions. A small amount of water is still dripping through the pipe opening in such a situation and burst in a

pipe can be defined as one that has a significant opening due to rupture of pipe that allows water to easily flow out of it in a significant amount sometimes disrupting the water services. It can be a nuisance for local public, commuters, and traffic. The study considered multiple parameters including pressure, flow, timestamp, pipe size, pump speed, diurnal water demand patterns, and roughness coefficients. Temperature was not included because the pipelines are buried at depth and temperature variations were considered to have negligible influence on the hydraulic behaviour investigated in this research. Soil properties and soil–water interactions were also excluded, as the focus of the study was on leak and burst detection using hydraulic network parameters rather than pipe deterioration processes.

4.2.1 Logic based method for the Pumping Main System

Logic-based methodologies form the foundational layer of anomaly detection within water distribution networks, primarily focusing on identifying deviations from normal operational behaviour. These methods employ logic-based algorithms grounded in hydraulic theory, operational thresholds, and system-specific parameters to detect inconsistencies indicative of potential leak or burst events. By leveraging preconfigured logical constructs, such as flow and pressure bounds, and applying domain expertise, the system is capable of rapidly flagging abnormalities that warrant further investigation. This approach offers transparency and interpretability, serving as a reliable first line of defense prior to more computationally intensive detection techniques. Figure 4.1 illustrates the variation of pressure head (m) with respect to time (minutes) for the 450 mm diameter Asbestos–Cement pumping main located in the Sunbury region, Victoria.

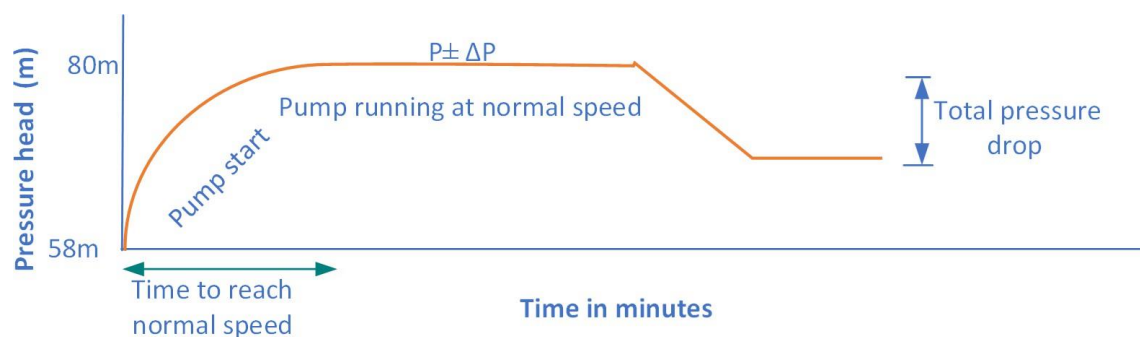


Figure 4.1: Variation of pressure head (m) with respect to time (minutes) for the 450 mm diameter.

Figure 4.2 illustrates the SCADA data profile extracted at one-minute intervals from the pressure transducer installed at the pumping station. The data capture the typical hydraulic behaviour of the pumping main during start-up, steady-state operation, and shutdown phases. This profile forms the

basis for establishing the threshold pressure variation (ΔP) applied in the logic-based leak and burst detection methodology (Joseph et al., 2024). Figure 4.2 is introduced to demonstrate the estimation for the delta P and delta Q. There is a slight variation in the pressure and flow during the pumping over time and that has been captured in Fig 4.2.

At the pump start-up stage, the pressure head rises non-linearly from approximately 58 m to 80 m as the pump accelerates from rest to its normal operating speed of 1500 rpm. This stage reflects the transient hydraulic response required to overcome both the static head and the frictional head loss along the 6.24 km main. The duration of this initial rise defines the time required for the pump to reach normal speed, governed by the impeller inertia and dynamic system resistance.

It should be noted that this analysis is based on a pumping main system with relatively stable demand, where hydraulic variations are primarily governed by pump operation rather than diurnal consumption patterns. Once the pump achieves its normal operating speed, the pressure head stabilises around $80 \text{ m} \pm \Delta P$, where ΔP denotes small pressure oscillations associated with minor system fluctuations and operational disturbances within the network. This period, marked as *Pump running at normal speed*, indicates a steady-state condition, resulting in stable pressure and flow behaviour. Based on the characteristic curve of the pumps, the flow in the pipe can change if the system head changes that is recorded by the flow meters installed in the system. The system head reduces during the burst/ leak condition in the pipeline based on its location. Toward the end of the pumping cycle, the curve shows a gradual but distinct decline in pressure head, corresponding to the pump shutdown phase. As the impeller decelerates and the flow rate decreases, the energy imparted to the water column is reduced, causing the head to fall from the steady-state value back toward 58m. It is due to the fact that there is a not return valve at the start of the pipeline and the elevation difference between pump station and reservoir location is 58m. Under normal operating conditions, the system pressure is almost a constant line; however, a sudden or excessive drop beyond this gradient signal a burst or major leak within the main (Figure 4.1). As can be seen from Figure 4.2 (a) the variation in pressure P at a time will be with ΔP and similarly, Figure 4.2 (b) the variation in flow Q will be within a range of ΔQ .

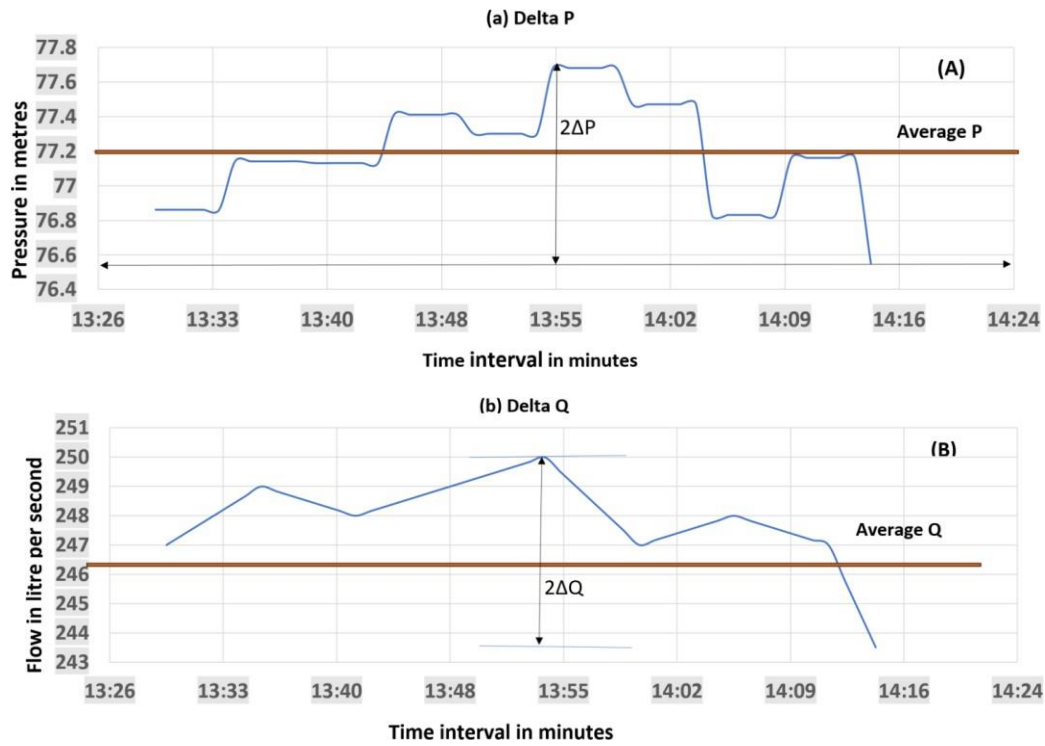


Figure 4.2: (a) Estimation of delta pressure (ΔP) and (b) delta (ΔQ) for case study system

This pressure–time profile defines the baseline hydraulic signature of the 450 mm pipeline under standard operating conditions. It serves as a reference for:

1. Establishing normal pressure variation limits (ΔP , ΔQ) for anomaly detection (Refer Figure 4.2 (a), and Figure 4.2 (b)).
2. Differentiating routine pump transitions from abnormal transients.
3. Supporting calibration and validation of logic-based algorithm used for event classification.

4.2.1.1 Integration of Hydraulic Behaviour into the Logic-Based Algorithm

The empirical pressure–time behaviour shown in Figure 4.2 (a) and Figure 4.2 (b) provides the fundamental basis for the logic-based leak and burst detection algorithm developed in this research. Understanding these transient and steady-state characteristics enables the definition of precise operational thresholds (ΔP and ΔQ) that govern the *if* and *else-if* decision logic used for event

detection. During implementation, variations in pump speed, time, pressure and flow readings from SCADA are continuously compared against these thresholds. If the deviation in pressure exceeds twice the baseline ΔP while the pump operates at normal speed, the algorithm flags a minor or major leak depending on the percentage deviation. When the pressure drop exceeds three times ΔP within a short time window, the condition is classified as a burst. These rules are embedded within the automated workflow described in Figure 4.3 (Logic-Based Algorithm Flowchart), ensuring consistent classification and alert generation.

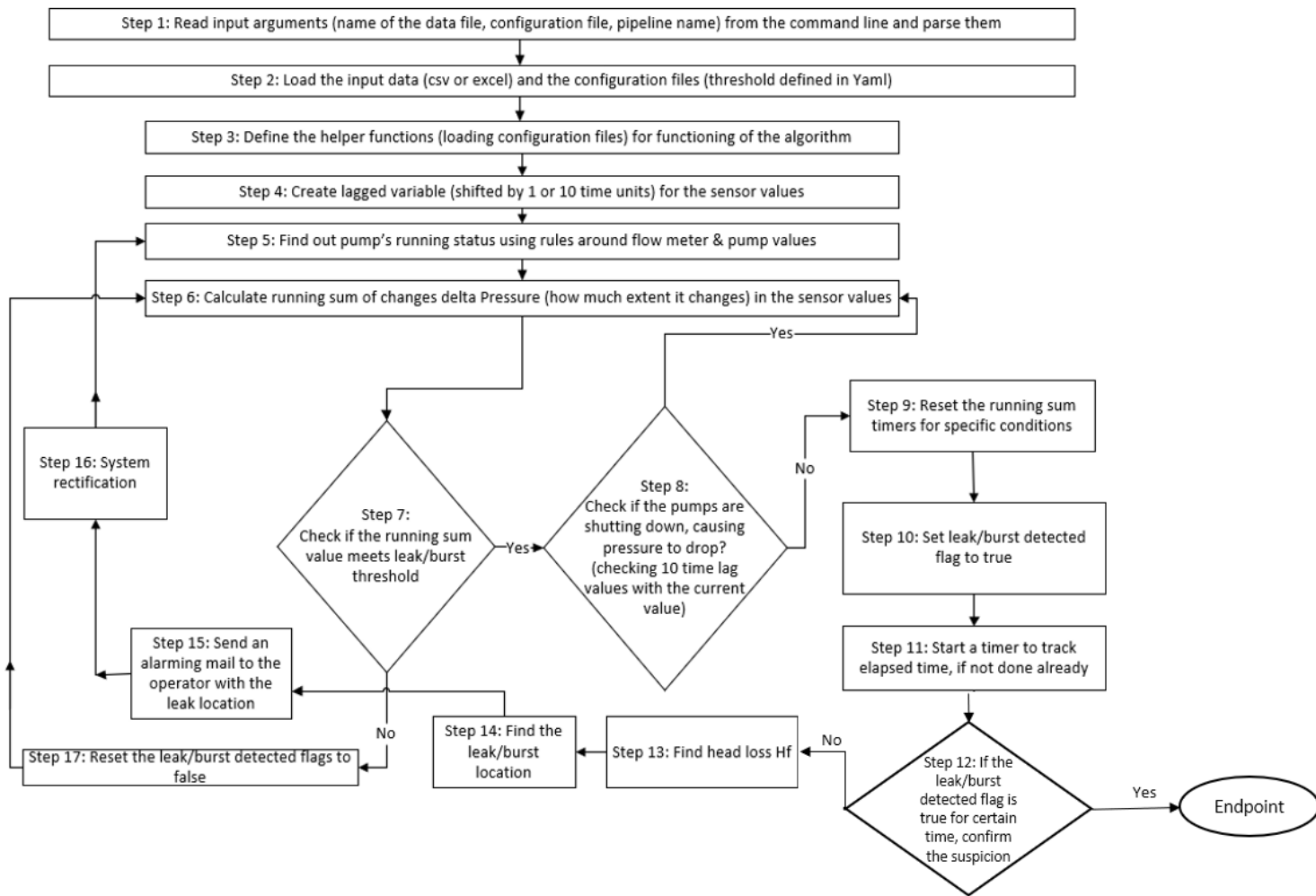


Figure 4.3: A flow diagram of the leak and burst detection algorithm.

Furthermore, integrating the measured pressure profile allows the algorithm to discriminate between genuine faults and normal operational transients such as pump shutdowns. The characteristic rise and decay patterns from the Sunbury dataset were used to model expected

transient responses, enabling the suppression of false alarms. This coupling of field-measured hydraulic signatures with rule-based decision logic establishes a data-driven yet interpretable detection layer, improving reliability and supporting immediate operator response in real-world conditions.

4.2.1.2 Underlying Principles of the Logic-Based Framework

The logic-based leak and burst detection methodology is underpinned by a multi-layered analytical framework that incorporates baseline benchmarking, threshold-based anomaly detection, and hydraulic principles.

Baseline Benchmarking:

- i. Establishing normal operational ranges for flow, pressure, and pump speeds based on historical SCADA data.
- ii. Accounting for pump speed variations to differentiate anomalies from regular usage fluctuations.

Threshold-Based Anomaly Detection:

- i. Setting thresholds for pressure drops, flow variations, and pump speed changes as indicators of leaks or bursts.
- ii. Flagging anomalies when deviations exceed predefined benchmarks.

Hydraulic Analysis:

- i. Employing equations like Darcy-Weisbach to calculate head loss and segment pipelines for precise anomaly localisation.
- ii. Estimating leak distances using pressure variations along segmented pipelines considering ground elevations along the pipeline.

The developed system supports a range of diagnostic functionalities for network monitoring

- i. Leak Detection: Identifying gradual leaks through persistent deviations in pressure or flow.
- ii. Burst Identification: Detecting significant and abrupt changes indicative of pipeline ruptures.

- iii. Pump Monitoring: Analysing pump transitions to differentiate operational fluctuations from true anomalies.

The integrated Python-based logic engine processes high-frequency SCADA input enabling the system to issue actionable alerts with minimal delay. This framework enhances decision-making by minimising false positives and ensuring timely maintenance interventions across the water distribution network.

4.2.1.3 Algorithm design for the Logic-based method

Logic-based methods involve defining explicit logic or conditions that, when met, indicate the presence of specific events or conditions, which are associated with a leak or burst in a pipeline.

The pressure drop logic is defined as if a sudden and significant pressure drop is detected in the pipeline over a short period of time, it may indicate a burst, or if the pressure drop is small over a long period of time in excess to the normal pressure variation during system operation. The logic could trigger an alarm when the pressure drop exceeds a predefined threshold. Pressure anomalies are unusual Pressure patterns, such as unexpected increases or decreases in flow rates, which can be indicative of leaks. Logic-based systems can flag such anomalies for further investigation. Pump speed variation occurs when the pump speed gradually increases from 0 to a normal operating speed once started and similarly decreases from the normal operating speed over a certain period under shutdown conditions.

Logic-based systems can analyse historical data to establish patterns of normal behaviour. Deviations from these patterns can be flagged as potential leaks or bursts. The development of a logic-based algorithm using Python programming language for the detection of leaks and bursts in water pipelines is rooted in its interpretability, adaptability, and domain-specific applicability, which can help in addressing the challenges posed by leak and burst detection.

Figure 4.3 details the steps involved in the development of the algorithm for leak and burst identification. The following are the main considerations involved in the development of the algorithm: (a) Collect SCADA data for flow in pumping main, pressure, and pump speed, (b) estimate variation in pressure and flow (delta ΔP and ΔQ), while the pump is running on its normal speed for assessment of threshold values for leaks and burst conditions for algorithm development, (c) estimate pump speed variation during pump start to full speed and then decrease in pump speed

during pump shutdown condition (identify pump start and shutdown stages), (d) develop an algorithm (coded in Python) for identification of leak and burst conditions and to notify the operator for any leak and burst condition and (e) in case of burst condition estimation of approximate burst location.

The following identifications are essential for algorithm development:

Pressure Anomalies: Unusual Pressure patterns, such as unexpected increases or decreases in Pressure rates, can be indicative of leaks and burst. Logic-based systems can flag such anomalies for further investigation.

Historical Data Analysis for Algorithm Development: Logic-based systems can analyse historical data to establish patterns of normal behaviour. Deviations from these patterns can be flagged as potential leaks or bursts.

It is hoped that the algorithm developed herein will be used on pumping systems using live data to help water systems operators with leak and burst identification in pipe networks quickly.

The main tasks in the diagram are from Step 6 to Step 10. The given specified algorithm outlines a series of steps for detecting leaks and bursts in a water distribution system. The process begins by reading per-minute input arguments from the command line for flow, pressure, and pump speed sensor values. Subsequently, threshold values for ΔP (Delta Pressure), ΔQ (Delta Flow), and ΔS (Delta Speed of the pump) are estimated using the recorded data; however, these values can be updated periodically.

The following main tasks are involved in the development of the algorithm for the identification of leaks and bursts.:

Task 1 (Step 1-5): Collect SCADA data for flow in the pumping main, pressure, and pump speed.

Step 1: Read input arguments (name of the data file, configuration file, pipeline name) from the command line and parse them. First, the program collects information about the pressure sensor, flow meter, and pump in the water pipeline. This data includes the name of the data file, a configuration file, and the name of the pipeline.

Step 2: Load the input data (CSV or Excel) and the configuration files (threshold defined in Yaml).

Where YAML is a human-readable data format commonly used for configuration files and data exchange between languages with different data structures. This step verifies the integrity and structure of the input data files, including checking for missing values, correct formatting, consistent timestamps, and valid sensor readings (pressure, flow, and pump speed) in CSV or Excel format. In addition, the configuration file is validated to ensure that predefined threshold values (e.g., pressure drop (in meters) thresholds with 5, 15, 20m (which are for minor leak, major leak, and burst) are correctly specified in the YAML file for subsequent analysis.

Step 3: Define helper functions for loading and processing configuration files and sensor data. These functions are designed to perform specific tasks within the leak and burst detection framework, including parsing threshold values from the YAML configuration file, extracting relevant SCADA parameters (pressure, flow, and pump speed), and preparing the data for anomaly detection. These modular functions improve the efficiency, readability, and maintainability of the detection algorithm.

Step 4: Create lagged variables (pressure and pump speed) for the sensor data to capture temporal dependencies in the system behaviours. In this study, lagged variables are generated by shifting the time-series data by 1 to 10-time steps, where each time step corresponds to a 1-minute interval in the SCADA data. Therefore, the lagged variables represent delays of 1 to 10 minutes (e.g. checking current pressure and speed data with previous up to 10 minutes data of similar variables). These lag features are used to incorporate historical system behaviours into the detection model, enabling improved identification of transient anomalies associated with leaks and bursts.

Step 5: The final stage involves determining the pump's running status using rules around the flow meter and pump values.

Task 2 (Step 6–8): Estimate variations in pressure (ΔP), flow (ΔQ), and pump speed (ΔS) under normal operating conditions when the pump is running within its typical operational range. These variations are used to establish statistical thresholds for identifying leak and burst events. In this study, anomaly detection is based on statistical deviation from normal operating conditions. A data point exceeding three times the standard deviation (3σ) from the mean is considered indicative of a

burst event, consistent with established statistical approaches for anomaly detection (Montgomery, 2019; Lewis, RE, 2024; Chandola et al., 2009). For leak detection, smaller but persistent deviations from normal operating conditions are considered. Unlike burst events, leak conditions are characterised by gradual changes in pressure and flow, and therefore do not exhibit abrupt deviations exceeding the 3 times threshold. Instead, leak detection is based on identifying sustained deviations from baseline behaviour over time.

Step 6: Calculate the running sum of changes in sensor values (pressure and flow) over a defined time window during normal pump operation. In this study, normal operation refers to periods when the pump is actively running within its standard operating range (i.e., not during start-up or shutdown phases). The running sum is computed by aggregating the differences between consecutive pressure readings (at 1-minute intervals) to identify deviations from stable operating conditions.

Step 7: Compare the calculated running sum values against predefined burst thresholds. A burst condition is identified when the deviation exceeds 3 times of the baseline variation, representing a significant and abrupt change in system behaviour. For leak detection, instead of using a fixed statistical threshold, the analysis focuses on identifying sustained deviations from baseline pressure and flow behaviour over a defined time period. These persistent deviations are indicative of gradual system changes associated with leak conditions. If the burst threshold is exceeded or sustained deviations are observed, proceed to Step 8; otherwise, continue monitoring.

Step 8: Check whether the observed pressure drop is associated with pump operational changes, such as pump shutdown or speed reduction, by analysing pump speed data. This step ensures that pressure variations caused by operational conditions are not misclassified as leak or burst events.

Task 3 (Step 9): Estimate pump speed variation during pump start to full speed and then decrease in pump speed during pump shutdown condition (identify pump start and shutdown stages).

Step 9: Reset the running sum timers for specific condition (if the pump is on or off).

Task 4 (Step 10-12): The identification of leak and burst conditions and to notify the operator of any leak and burst conditions.

Step 10: If the leak/burst thresholds are met, set the leak/burst detected flag to true.

Step 11: Start a counter to track time for Step 10.

Step 12: If the burst detected flag is true for a certain time, confirm the suspicion.

Task 5 (Step 13-17): Estimate the appropriate leak location along the pressure water main

Step 13: Find head loss H_f (shown in equation 4.1) the pressure main based on the current pressure sensor readings under burst condition and topography considerations.

Step 14: Find the burst location using hydraulic equations (shown in equation 4.1 to 4.5)

Step 15: Send an alarming mail to the operator with more information.

Step 16: Burst alerts at the start and end of the event.

Step 17: If the running sum value does not meet the burst threshold, reset the burst threshold values to false.

In case of live data availability, the process will check ongoing data to identify leaks or bursts based on system capabilities in terms of pressure/flow data availability across the pressure main.

4.2.1.4 Burst location identification in a pumping main (case study system)

The algorithm can identify the approximate location of a leak or burst in pumping or gravity main systems, provided significant pressure and flow sensors are placed at regular intervals. In the absence of such an arrangement, still, the approximate location of a pipe burst can still be estimated with limited pressure and flow sensors. Using the change (drop) in pressure head reading at the pumping end, the location of burst L_X can be estimated using any pipe head loss equation.

Darcy Weisbach's head loss equation for flow in pipes can be written as in below equation 4.1:

$$h_f = \frac{8fLQ^2}{\pi^2 g D^5} \quad (4.1)$$

where f is the friction factor in the pipe, L is the pipe length in meters, D is the pipe diameter in meters, Q is the fluid flow in cubic meter/sec at the time of burst, and g is the gravitational constant.

The flow rate Q will slightly change from steady state condition to leak condition due to variation in system head and pump characteristics.

The friction factor f can be estimated using Swamee equation (Swamee, 1993) [34] in the equation 4.2:

$$f = 1.325 \left\{ \ln \left[\frac{\varepsilon}{3.7D} + 4.618 \left(\frac{\nu D}{Q} \right)^{0.9} \right] \right\}^{-2} \quad (4.2)$$

where ε is the average height of roughness projection of the pipe wall and ν is the kinematic viscosity of the fluid. Kinematic viscosity can be obtained using Equation 4.3 (Swamee, 2004) [35]:

$$\nu = 1.792 \times 10^{-6} \left[1 + \left(\frac{T}{25} \right)^{1.165} \right]^{-1} \quad (4.3)$$

where T is the water temperature in degree centigrade.

For known head loss in the pipeline, pipe length and pipe diameter, friction factor f can be calibrated using Equation 4.1, and then ε the pipe wall roughness height can be calculated using Eq. 4.2.

In the case of a pipe burst, the approximate location of the burst from the pumping station along the pipe can be estimated using the following equation.

$$P_X = \frac{E_d}{L}L_x + \frac{8fQ^2}{\pi^2gD^5}L_x \quad (4.4)$$

where P_x is the pressure at the sensor at the time of the burst, L_x is the length of the leak location along the pumping main alignment, L is the length of the pumping main, E_d is the elevation difference between the pumping main and the service tank. The topography of the pipe alignment has a uniform gradient in the case study; however, Equation 4.4 can be modified to incorporate various gradients at different lengths across the pumping main.

A simple system diagram for leak location detection is shown in Figure 4.4 below:

Pipe length for burst location L_x can be estimated as in equation 4.5:

$$L_X = \frac{P_x}{\frac{E_d}{L} + \frac{8fQ^2}{\pi^2gD^5}} \quad (4.5)$$

As the pump system head (static head and head loss in pipe) will change due to pipe burst and thus in some pumps the pump characteristics will also change, which can also be incorporated in the algorithm specific to pump conditions. In Figure 4.4 below, a pipeline with a diameter of 0.45m extends over a length of 6280m, connecting the pump to a water tank. The pump is situated at an elevation of RL 138.3m, while the tank stands at a reduced level of 196.3m.

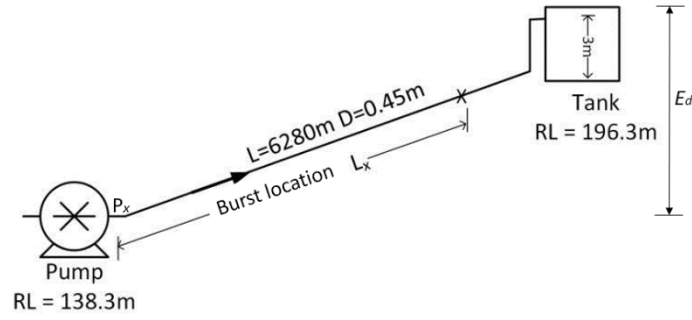


Figure 4.4: Burst Location at X

These equations were incorporated in the software (Python code) to estimate L_x based on flow and pressure.

It should be noted that Equations (4.1)–(4.5) are derived based on an idealised single-pipe system and therefore provide an approximate approach for burst localisation in practical networks with limited instrumentation. This analytical formulation is specifically applicable to the Sunbury pumping main dataset due to its simplified pipeline configuration. In contrast, the Melton prototype system, which involves multiple measurement points and controlled burst scenarios, is analysed using data-driven and comparative methods rather than direct analytical localisation. For more complex water distribution systems with varying pipe diameters and branching, the method is best applied in a segment-wise manner between sensor locations or integrated with hydraulic models to improve accuracy.

4.2.1.5 Machine Learning–Based Detection Methods

Although the logic-based system provides interpretability and quick decision-making, its reliance on fixed thresholds limits adaptability to complex, noisy, or evolving operational conditions. Real SCADA datasets exhibit nonlinear behaviour, stochastic fluctuations, and sensor uncertainties that often blur the distinction between normal and anomalous states. To overcome these limitations and enhance the scalability of leak and burst detection, the research incorporates Machine Learning (ML) algorithms as a complementary analytical layer.

The parameters derived from the hydraulic profile are namely pressure variation (ΔP), flow variation (ΔQ), and pump speed (ΔS) which were then transformed into input features for

unsupervised ML models. Approximately 20,000 data points representing normal, leak, and burst scenarios were used to train and validate multiple algorithms, including Local Outlier Factor (LOF), DBSCAN, Isolation Forest, K-Means, and One-Class SVM.

Unlike deterministic thresholding, ML models infer probabilistic boundaries of normal operation and autonomously identify deviations indicative of leaks or bursts. By learning from historical pressure–flow–speed relationships, the models adapt to dynamic system behaviour, thereby minimising false positives and improving sensitivity to subtle anomalies. The combination of both approaches is logic-based detection for transparency and ML-based detection for adaptability which forms a hybrid intelligent framework capable of anomaly detection, continuous learning, and scalable deployment across diverse water distribution networks.

Machine Learning is a subfield of artificial intelligence (AI) that focuses on the development of algorithms and models that enable computers to learn patterns from data. Machine Learning algorithms are suitable for detecting leaks and bursts in water pipelines due to their ability to provide early detection, analyse large datasets comprehensively, and monitor systems. These algorithms contribute to reducing false positives, adapting to changing conditions, and optimising maintenance strategies by predicting potential issues and prioritising areas at risk. The integration of Machine Learning with Internet of Things (IoT) devices and sensor networks enhances accuracy and efficiency, leading to significant cost savings through proactive intervention and efficient resource allocation. Overall, Machine Learning can play a crucial role in ensuring the sustainable management of water resources by improving the reliability and effectiveness of leak and burst detection in water pipeline systems.

Machine Learning (ML) methods serve as a pivot role in complementing logic-based frameworks by leveraging advanced algorithmic intelligence and data-driven insights to enhance anomaly detection. This section evaluates multiple Machine Learning models, focusing on their performance in detecting leaks and bursts in water networks.

In recent years, ML models have emerged as effective tools for detecting leaks in water pipelines. These models leverage historical data and advanced algorithms to identify anomalies indicative of

leaks, offering a more precise and scalable solution compared to traditional methods such as manual inspections or basic statistical analyses. The methodology, evaluation metrics, and experimental results are outlined to present a clear understanding of each model's capabilities and limitations. Furthermore, this analysis considers the real-world implications of ML-based anomaly detection for practical deployment in smart water systems.

Methodology:

The dataset used in this study was sourced from experimental and real-world data of water distribution networks. The pre-processing phase included several critical steps to ensure the dataset was clean, balanced, and ready for model training and evaluation. These steps are detailed as follows:

1. **Reading Data:** The dataset was imported from an Excel file, which served as the primary repository for historical pipeline data. Column names were standardised for uniform access, ensuring consistency across various pre-processing tasks. This standardisation is particularly important when dealing with large datasets from multiple sources, as it minimises errors during data manipulation.
2. **Data Cleaning:** Missing values were imputed using statistical methods such as mean, median, or mode imputation, depending on the nature of the data. Categorical data refers to variables that represent discrete categories rather than continuous numerical values. In water distribution systems, examples of categorical data include pump operational status (e.g., ON/OFF), valve status (OPEN/CLOSED), and event labels such as normal, leak, or burst conditions. For numerical variables, missing values were imputed using either the mean or median depending on the data distribution. The mean was used when the data were approximately normally distributed and free from significant outliers, as it preserves the overall statistical properties of the dataset. In contrast, the median was used when the data exhibited skewness or contained outliers, as it provides a more robust measure of central tendency. The choice of imputation method was independent of the length of missing data segments; however, longer gaps were treated with caution as they may introduce higher uncertainty in the imputed values.
3. **Feature Standardisation:** Numerical features were normalised to ensure a uniform scale, which is crucial for models sensitive to feature magnitudes. Standardisation involved

transforming the data to have a mean of zero and a standard deviation of one. This step not only improved model performance but also accelerated convergence during training.

4. **Binary Conversion:** The target variable, indicating the presence or absence of leaks and burst, was encoded in binary format-assigning to leak events and non-leak events. Although these labels were not used during the unsupervised model training, this encoding was introduced to simplify interpretation of the model's output. By defining as non-leak and as leak and burst, we can clearly map predictions to real-world leak and burst states and construct an accurate classification error matrix for evaluation.
5. **Dataset Splitting:** The dataset was divided into training and test sets using stratified sampling. This method preserved the proportion of leak and no-leak instances across both subsets, ensuring a balanced representation of the target variable. Such a balanced split is essential for preventing model bias and achieving reliable performance metrics.

The research framework is strategically designed to address the growing demand for advanced leak and burst detection in modern water distribution systems. It integrates experimental validation, numerical modelling, and data analysis to establish an intelligent, scalable, and adaptive detection methodology. The overarching aim of this framework is to leverage the capabilities of SCADA systems in conjunction with logic-based and Machine Learning techniques to enhance the efficiency and reliability of water network management. By combining multiple layers of analytical intelligence ranging from data acquisition to algorithmic processing and validation the framework provides a holistic approach to predictive and responsive network monitoring.

The framework comprises four major components. The first component, SCADA-enabled data acquisition, involves continuous monitoring of flow and pressure parameters through strategically placed sensors across critical network locations. Data collected from both experimental setups and field installations are transmitted to a centralised analysis system, ensuring uninterrupted observation and rapid event detection. The algorithms are designed with features like threshold detection, time-lag computation, and logic-based classification schemes to ensure reliable detection under varying hydraulic conditions.

The second component, Machine Learning integration, enhances detection accuracy and robustness by training supervised and unsupervised models using historical and experimental datasets. Key hydraulic parameters such as flow rate, pressure, pump speed, and diurnal demand patterns serve

as predictive indicators in the classification and anomaly detection processes. The integration of Machine Learning enables the system to adaptively learn from evolving network behaviours, reducing false alarms and improving overall responsiveness. The third component, experimental and numerical validation, strengthens the framework’s reliability by combining controlled

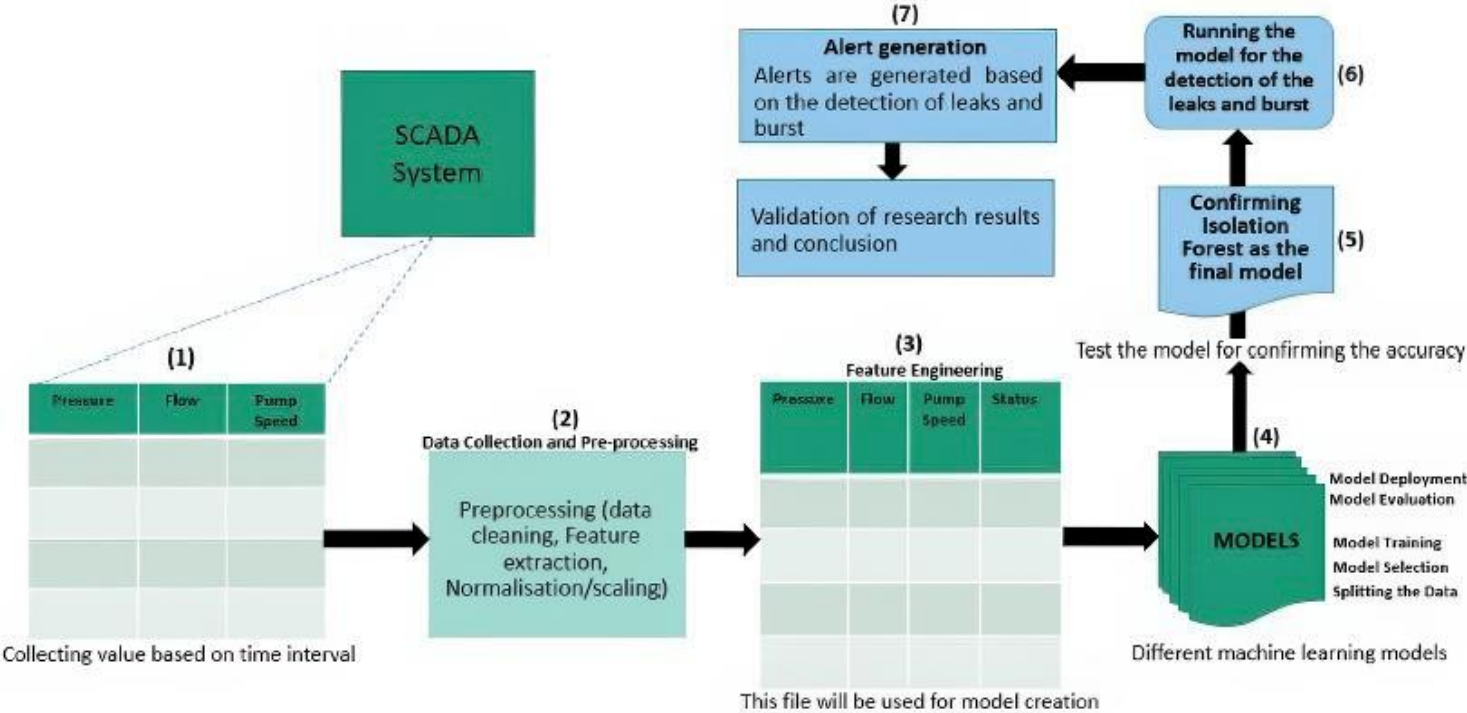


Figure 4.5: Developed Machine Learning approach for leak and burst detection.

prototype experiments with numerical hydraulic modelling. Simulated leak scenarios are generated under predefined conditions to test detection accuracy, while numerical models validate hydraulic responses and confirm the algorithm’s predictive performance. The Machine Learning technique selection and application process is shown in Figure 4.5.

The steps (marked in Figure 4.5) in the Machine Learning approach for leak and burst detection are provided below:

Step 1: Collect the data from the existing SCADA for Pressure sensors, flow meters, and pump speeds in the time interval of 1 minute for 20,000 data points.

Step 2: Preprocess the SCADA data through systematic data cleaning and feature engineering.

Data cleaning involves identifying and handling missing values using appropriate imputation methods (mean based data distribution), removing inconsistent or erroneous records, and ensuring uniform time intervals. Feature engineering includes generating derived variables such as pressure differences (ΔP), flow variations (ΔQ), and lagged features to capture temporal dependencies in the system behaviour.

Step 3: The prepared data will be used for the Machine Learning algorithms.

Step 4: Different Machine Learning models will be used for training and testing.

Step 5: Selection of the highest performing Machine Learning model with the Machine Learning parametric indices (Anomaly F1 score and ROC curve -explained below).

Step 6: Implementing the selected model for the detection of leaks and bursts.

Step 7: Alert generation for the leaks and bursts in the pipeline.

The F1 score is the harmonic mean of precision and recall and is often used in binary classification and anomaly detection problems. In this study, anomalies correspond to leak and burst events within the water distribution system. These anomalous conditions are identified based on deviations from normal operating behaviour using the proposed detection framework. Specifically, burst events are characterised by abrupt and significant changes in pressure and flow (e.g., exceeding predefined statistical thresholds), while leak events are associated with sustained deviations from baseline behaviours over time.

These conditions are used as the ground truth for evaluating the performance of the Machine Learning models. F1 score is calculated using the following equation 4.6 (Hastie et al., 2009; James et al., 2013):

$$F1 = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4.6)$$

Precision is the proportion of data points identified as anomalies by the model that are anomalies and is calculated using equation 4.7. True positive is the proportion of actual anomalies that are correctly identified as anomaly. False positive is the proportion of non-anomalous data points that are incorrectly classified as anomalies. Recall (Sensitivity) is the proportion of actual anomalies that are correctly identified by the model and are calculated using equation 4.8.

$$Precision = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \quad (4.7)$$

$$Recall (Sensitivity) = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \quad (4.8)$$

In evaluating the predictive models for leakage within the pipeline network, it's crucial to consider performance metrics such as the true positive rate and false positive rate. The true positive rate, also known as sensitivity or recall, measures the proportion of actual positive cases that are correctly identified by the model as positive. In the context of leak detection, it signifies the accuracy of the model in correctly identifying instances of leakage, ensuring timely intervention and maintenance. Conversely, the false positive rate measures the proportion of negative cases that are incorrectly classified as positive by the model. In leak detection, a high false positive rate could lead to unnecessary resource allocation and operational disruptions. Balancing these rates is essential for optimising the efficiency and reliability of the predictive model in detecting and managing pipeline integrity issues.

The Receiver Operating Characteristic (ROC) curve is a graphical representation that illustrates the trade-off between the true positive rate (sensitivity) and the false positive rate across different thresholds of a predictive model. The Area Under the ROC Curve (ROC AUC) quantifies the overall performance of the model across all possible threshold values. It represents the probability that the model will rank a randomly chosen positive instance higher than a randomly chosen negative instance. Mathematically, ROC AUC is calculated as the integral of the true positive rate with respect to the false positive rate, ranging from 0 to 1. This integral captures the entire area under the ROC curve, providing a single scalar value that summarises the model's discriminatory power. A higher ROC AUC value indicates better overall performance of the model in distinguishing between positive and negative instances, with a value of 1 representing a perfect classifier.

The ROC AUC (Receiver Operating Characteristic Area Under the Curve) is a measure of the area

under the ROC curve, which plots the true positive rate (sensitivity) against the false positive rate (1 - specificity) at various thresholds. The ROC AUC is calculated using equation 4.9 (Hand & Till, 2001; Lewis, 2024):

$$\text{ROC AUC} = \int_0^1 \text{True Positive Rate (Sensitivity)} d(\text{False Positive Rate}) \quad (4.9)$$

To address the issue of data imbalance, oversampling techniques such as ADASYN and SMOTE were applied to improve the performance of the Machine Learning models in leak detection. In scenarios where positive instances (leaks) are significantly outnumbered by negative instances (non-leaks), these techniques generate synthetic samples for the minority class. By creating a more balanced training dataset, ADASYN and SMOTE help reduce bias, minimise overfitting, and improve the generalisation capability of the models for accurate leak detection in water distribution systems. These methods contribute to a more representative decision boundary, increased sensitivity to anomalies, and improved performance metrics, making them essential tools for building accurate and robust leak detection models.

4.2.1.5.1 Supervised Machine Learning Techniques for the Prototype System

Supervised Machine Learning requires labelled data, where examples of both normal and anomalous (leak/burst) conditions are used to train the model (Hastie et al., 2009; James et al., 2013). However, obtaining labelled data for leak detection can be challenging and costly, as leak events are relatively rare and not always well documented (Romano et al., 2014). These techniques were therefore employed in this study for leak detection in the prototype system. Supervised Machine Learning relies on labelled datasets where instances of leaks and normal pipeline conditions are predefined. Algorithms such as decision trees, support vector machines (SVMs), and neural networks are widely used for classification and anomaly detection tasks (Hastie et al., 2009; Bishop, 2006). By learning patterns from training data, supervised methods can achieve high detection accuracy when sufficient labelled data are available.

Supervised Machine Learning Models

Supervised learning involves training models on labelled datasets, where each input is associated with a corresponding output. The following supervised models were developed and evaluated for

their effectiveness in leak detection based on established Machine Learning methodologies (Hastie et al., 2009; James et al., 2013). The following supervised models were developed and evaluated for their effectiveness in leak detection:

Decision Tree: The decision tree algorithm is a rule-based classifier that splits the dataset into branches using conditions derived from feature values, ultimately assigning class labels at terminal nodes (Quinlan, 1986). In leak detection, it identifies patterns between hydraulic variables such as pressure and flow, forming clear decision pathways that distinguish between normal, minor leak, and burst states. Its main advantage lies in its interpretability and ease of visualisation. However, decision trees are sensitive to noise and small changes in data, often leading to overfitting and reduced generalisation in complex systems.

AdaBoost: AdaBoost, or Adaptive Boosting, combines multiple weak learners, typically decision trees, into a strong ensemble model by assigning higher weights to misclassified samples during each iteration (Freund & Schapire, 1997). In leak detection, AdaBoost improves sensitivity towards minority leak events by sequentially correcting errors from previous iterations, making it particularly effective for imbalanced datasets.

Linear Discriminant Analysis (LDA): LDA is a statistical classifier that projects multidimensional data onto a lower-dimensional space while maximising the separation between classes, under the assumption of normally distributed data with equal covariance matrices (Hastie et al., 2009; Bishop, 2006). When applied to leak detection, LDA identifies linear relationships between hydraulic features such as flow deviations and pressure drops, helping separate normal operation from leak states. The method's simplicity, efficiency, and interpretability make it suitable for preliminary classification. However, its performance diminishes when data exhibits nonlinear behaviour or when class covariances differ significantly, limiting its applicability in complex water networks.

Quadratic Discriminant Analysis (QDA): QDA extends LDA by allowing each class to have its own covariance matrix, thereby introducing nonlinear decision boundaries (Hastie et al., 2009; Bishop, 2006). This flexibility makes QDA suitable for modelling hydraulic responses that vary across leak severities. It can capture differences in variability and correlation among features such as pressure, flow, and velocity. However, QDA requires accurate covariance estimation and is

more prone to overfitting when training data are limited or noisy, particularly in high-dimensional settings.

Gaussian Naïve Bayes (GNB): The Gaussian Naïve Bayes algorithm is based on Bayes' theorem and assumes conditional independence between features along with Gaussian-distributed inputs (Bishop, 2006; Murphy, 2012). In leak detection, it estimates the probability of a leak condition given observed hydraulic measurements, enabling fast and computationally efficient classification. While effective for small datasets, its independence assumption may not hold for correlated hydraulic parameters such as pressure and flow, which can reduce classification accuracy. Therefore, GNB is often used as a baseline model rather than a standalone solution.

Ridge Classifier: The Ridge Classifier is a linear model that applies L2 regularisation to penalise large coefficients, thereby reducing overfitting and stabilising predictions (Hastie et al., 2009; Bishop, 2006). In leak detection, it captures linear relationships between sensor features and leak classes, providing a balance between bias and variance. The algorithm is particularly effective in handling multicollinearity among features, such as correlated sensor readings across nearby nodes. However, it may struggle to represent nonlinear or complex hydraulic relationships and requires careful tuning of the regularisation parameter for optimal performance.

Balanced Random Forest (BRF): The Balanced Random Forest extends the traditional Random Forest by training each decision tree on a balanced subset of the data, ensuring equal representation of each class (Chen et al., 2004; Breiman, 2001). In leak detection, this approach helps mitigate the dominance of normal operation data and improves the detection of rare leak or burst events. The algorithm is robust to noise and high-dimensional data; however, its ensemble nature increases computational cost and reduces interpretability.

RUSBoost: RUSBoost combines Random Under-Sampling (RUS) with boosting techniques such as AdaBoost to improve classification performance on imbalanced datasets (Seiffert et al., 2010). In leak detection applications, it enhances the model's ability to identify minority leak events by reducing the influence of the majority class during training. While effective in improving recall for rare events, it may discard useful information during undersampling, potentially affecting overall model stability.

Voting Classifier (Ensemble): The Voting Classifier is an ensemble approach that aggregates

predictions from multiple base models, such as Random Forest, Support Vector Classifier (SVC), and Multi-Layer Perceptron (MLP), to produce consensus output (Dietterich, 2000). In leak detection, it leverages the complementary strengths of different algorithms to generate robust and generalised classifications that are less sensitive to individual model biases. The key advantage of this approach is its ability to improve stability and accuracy across varying leak scenarios. However, the ensemble's complexity can make it computationally expensive and difficult to interpret, as it obscures the contribution of individual models in the final decision.

K-Nearest Neighbours (KNN): KNN is a non-parametric algorithm that classifies samples based on the majority label among their k nearest neighbours in the feature space (Cover & Hart, 1967; Hastie et al., 2009). In leak detection, it identifies leak conditions by comparing new hydraulic measurements with historical operational data, making it effective for datasets with distinct clusters or patterns. Its simplicity and adaptability make it suitable for quick deployment. However, KNN can be computationally expensive for large datasets due to repeated distance calculations, and its performance may degrade in the presence of irrelevant or noisy features.

Random Forest (RF): Random Forest is an ensemble learning method composed of multiple decision trees, where each tree is trained on a random subset of data and features, and final predictions are made through majority voting (Breiman, 2001). In leak detection, it captures complex nonlinear relationships between pressure, flow, and leak severity while maintaining robustness against overfitting. It also provides feature importance measures, offering insights into the influence of different variables. However, its main limitation is increased computational cost and reduced interpretability compared to single-tree models.

Support Vector Classifier (SVC): SVC constructs optimal hyperplanes to separate classes in feature space by maximising the margin between data points of different labels (Cortes & Vapnik, 1995; Hastie et al., 2009). By using kernel functions, it can transform nonlinear data into higher-dimensional representations for improved separation. In leak detection, SVC effectively captures nonlinear relationships among hydraulic variables and identifies subtle leak patterns. However, it requires careful selection of kernel parameters and can be computationally intensive for large

datasets.

Logistic Regression (LR): Logistic Regression predicts categorical outcomes using a logistic function that estimates the probability of class membership (Hastie et al., 2009; Bishop, 2006). In leak detection, it provides probabilistic predictions indicating the likelihood of normal or leak states, aiding operational decision-making. The model's simplicity and interpretability make it a suitable baseline method. However, since Logistic Regression assumes a linear relationship between inputs and outputs, it may fail to capture nonlinear interactions in hydraulic data, leading to reduced performance in complex or irregular network conditions.

Multi-Layer Perceptron (MLP): The Multi-Layer Perceptron is a feed-forward artificial neural network composed of interconnected layers that learn nonlinear relationships through backpropagation (Bishop, 2006; Goodfellow et al., 2016). In leak detection, MLP models complex dependencies between hydraulic variables and class labels, enabling the identification of intricate leak signatures. It can capture hidden patterns that traditional algorithms may overlook, making it suitable for advanced detection tasks. However, MLP requires substantial training data, computational resources, and careful hyperparameter tuning. Overfitting and long training times remain key limitations in practical applications.

4.2.1.5.2 Unsupervised Machine Learning Techniques for the Pumping Main System

Unsupervised learning involves identifying patterns in unlabelled data without predefined outputs (Hastie et al., 2009; Murphy, 2012). These algorithms discover structures, correlations, or anomalies within the data, making them suitable for tasks such as anomaly detection. In the context of leak detection, unsupervised methods are particularly useful when labelled data are limited or unavailable. They detect irregularities in key pipeline parameters such as pressure, flow, and acoustic signals without prior knowledge of fault conditions. Consequently, they provide a data-driven approach for identifying hidden anomalies that may not be captured by traditional threshold- or rule-based methods.

Prediction and Classification: Machine Learning models can be used for prediction (e.g., forecasting future values) and classification (e.g., assigning data into predefined classes) (Hastie et al., 2009; James et al., 2013). These tasks are typically performed using supervised learning techniques, which rely on labelled data. In

contrast, unsupervised learning focuses on identifying patterns and anomalies in unlabelled data without predefined output labels (Hastie et al., 2009; Murphy, 2012). In the context of water distribution systems, it detects deviations from normal pipeline behaviour using parameters such as pressure and flow, enabling the identification of potential leak and burst events.

1. Unsupervised learning has several applications across various domains, including:
2. Anomaly detection in fraud detection or network security
3. Data pre-processing and feature engineering before supervised learning.

The two primary types of unsupervised learning techniques are:

- a. Clustering: Clustering algorithms group similar data points together based on their inherent characteristics or proximity in the feature space. Common clustering algorithms include K-means, hierarchical clustering, and DBSCAN.
- b. Dimensionality Reduction: Dimensionality reduction techniques aim to reduce the number of features or variables in the dataset while preserving the most important information. This is useful for visualising high-dimensional data or reducing computational complexity. Techniques such as Principal Component Analysis (PCA) and t-SNE (t-Distributed Stochastic Neighbor Embedding) are commonly used for dimensionality reduction.

Unsupervised learning models are well-suited for detecting leaks and bursts as they do not rely on labelled data (Hastie et al., 2009; Chandola et al., 2009). These methods identify patterns in normal operational data and detect anomalies as deviations from expected behaviour. In water distribution systems, such deviations may be observed in parameters such as pressure, flow, or acoustic signals, indicating potential leak or burst events.

Data Pre-processing: Normalisation and standardisation were performed on both training and test data to ensure consistency in feature scaling.

Evaluation: Models were evaluated using F1 score as a metric, considering both normal and anomaly classes.

K-Means Clustering:

K-Means is an unsupervised clustering algorithm that partitions data into k clusters based on similarity (MacQueen, 1967; Hastie et al., 2009). Each data point is assigned to the cluster with the

nearest centroid, and the algorithm iteratively updates cluster centres until convergence. The objective is to minimise the variance within each cluster, ensuring that points in the same cluster are as similar as possible. In the context of leak detection, K-Means can be applied to group hydraulic or pressure data into clusters representing normal and abnormal operational states. Data points that do not align well with any cluster, particularly those with larger distances from cluster centroids, can be identified as potential leaks or anomalies. This approach is useful for analysing long-term operational trends and distinguishing between normal variations and leak-related deviations. However, K-Means has limitations. The number of clusters (k) must be predefined and may not always reflect the true data structure. The algorithm is also sensitive to initial centroid selection and feature scaling. Despite these limitations, K-Means can be enhanced by incorporating distance-based anomaly detection measures, such as evaluating residual distances from cluster centroids, to improve leak detection performance.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

DBSCAN is a density-based clustering algorithm that groups together data points that are closely packed and labels points in sparse regions as noise or anomalies (Ester et al., 1996; Schubert et al., 2017). It depends on two key parameters: ϵ (epsilon), which defines the neighbourhood radius, and $minPts$, which specifies the minimum number of points required to form a dense region. Unlike K-Means, DBSCAN does not require prior knowledge of the number of clusters, making it suitable for irregular or non-spherical data distributions. In leak detection, DBSCAN can identify abnormal hydraulic behaviours such as sudden pressure drops or flow deviations as outliers that do not belong to any dense cluster. This makes it particularly effective for detecting localised anomalies within water distribution networks. It can also be applied to spatio-temporal sensor data to detect regions exhibiting irregular or abrupt pressure changes. However, DBSCAN's performance depends on the appropriate selection of ϵ and $minPts$, which may vary across different network conditions. It may also struggle with datasets containing clusters of varying densities. To address these limitations, extensions such as Hierarchical DBSCAN (HDBSCAN) have been proposed to improve clustering performance in complex datasets (Campello et al., 2013).

One-Class SVM (Support Vector Machine):

One-Class SVM is an anomaly detection algorithm that learns a decision boundary enclosing normal data, assuming that most training samples represent typical system behaviour (Schölkopf et al., 2001; Bishop, 2006). Data points that fall outside this boundary are classified as anomalies. The model uses kernel functions (e.g., linear, polynomial, or radial basis function) to capture nonlinear relationships within the data. In leak detection applications, One-Class SVM can be used to monitor water distribution systems by analysing sensor data such as pressure and flow. Deviations from the learned normal behaviour are flagged as potential leak or burst events. However, the model's performance is sensitive to kernel selection and hyperparameter tuning, which influence its ability to distinguish between normal and anomalous conditions. To adapt to evolving network behaviour, periodic retraining may be required.

Isolation Forest:

Isolation Forest is a tree-based ensemble method designed to detect anomalies by isolating observations through random data partitioning (Liu et al., 2008; Liu et al., 2012). The key principle is that anomalies are rare and distinct, making them easier to isolate compared to normal data points. The algorithm constructs multiple random binary trees, and data points that require fewer splits to be isolated are assigned higher anomaly scores. In leak detection, Isolation Forest is effective in identifying rare events such as leaks or bursts within large-scale datasets from water distribution systems. Its ability to handle high-dimensional data and large datasets makes it suitable for analysing sensor data across multiple network locations. However, its performance may be affected when normal operational data exhibit high variability, potentially leading to false positives. To maintain detection accuracy, periodic model updates may be required to account for seasonal and operational changes in the system.

Local Outlier Factor (LOF):

The Local Outlier Factor (LOF) algorithm identifies anomalies by measuring the local density deviation of a data point relative to its neighbours (Breunig et al., 2000). It assigns an outlier score based on how isolated a point is compared to its surrounding data points, where points with significantly lower density than their neighbours are considered anomalies. This density-based approach enables the detection of subtle deviations in local data structure that may not be captured by global models. In leak detection, LOF is particularly useful for identifying localised anomalies within water distribution systems, such as leaks confined to specific zones or branches of the network. Its sensitivity to spatial variations makes it effective for analysing data from

multiple sensors distributed across the network. However, the performance of LOF depends on the appropriate selection of the neighbourhood size (k), which controls the scale of anomaly detection. Improper tuning may result in missed detections or false alarms. To enhance performance, LOF can be combined with spatial or temporal analysis techniques to improve leak localisation and detection reliability (Chandola et al., 2009).

4.3 Framework for Intelligent Leak and Burst Detection

The research methodology adopted in this study follows a structured and sequential workflow that integrates hydraulic modelling, SCADA data analytics, and Machine Learning-based leak detection. The framework, illustrated in Figure 4.6, consists of six interconnected stages including data acquisition, pre-processing, logic-based detection, Machine Learning modelling with class balancing, model validation, and decision-support integration. Each stage builds upon the preceding one to ensure that the developed system is scientifically rigorous, computationally efficient, and operationally deployable in real-world water-utility environments.

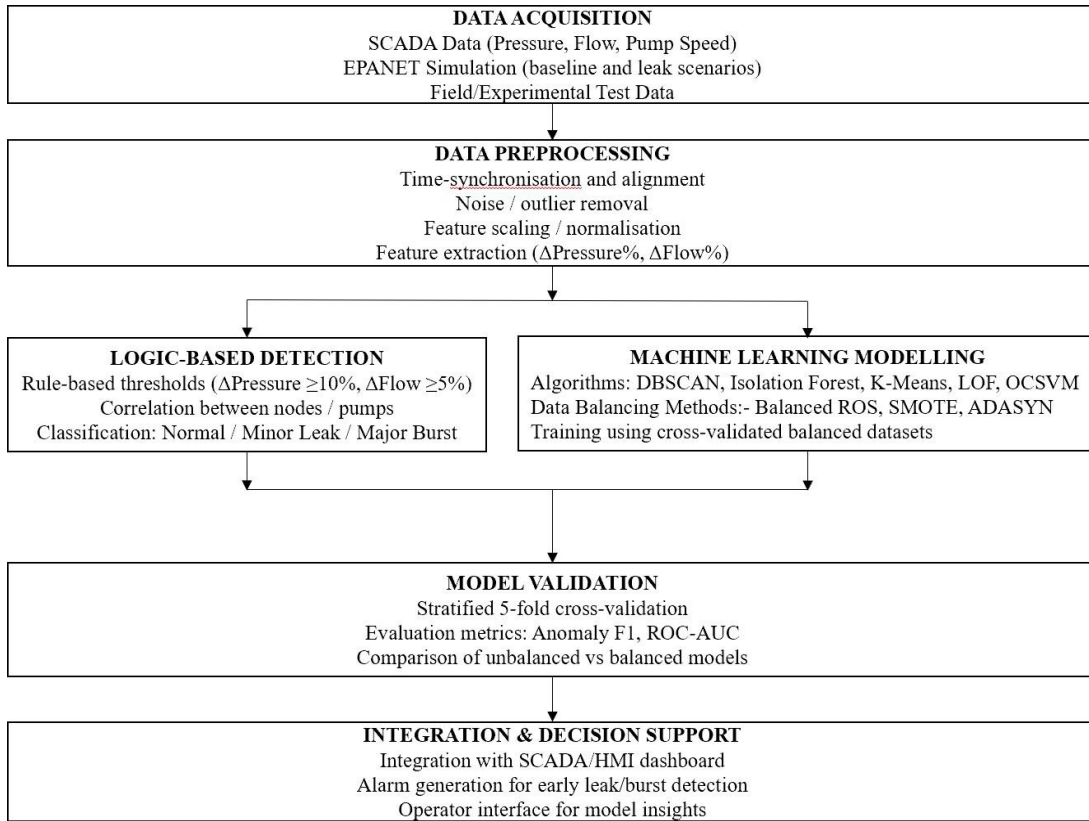


Figure 4.6: Research methodology framework integrating data acquisition, pre-processing, logic-based detection, & Machine Learning analysis (Balanced ROS, SMOTE, ADASYN) with validation & real-time decision-support stages for intelligent leak & burst detection.

The process begins with data acquisition, which serves as the foundation for all subsequent analyses. Continuous time-series data were obtained from SCADA systems installed across the monitored water network. These datasets include flow, pressure, and pump-speed readings sampled at high temporal resolution. Additional experimental or field-verified datasets were incorporated, where available, to enhance model validation and ensure representativeness across varying operational scenarios.

The next stage encompasses Machine Learning modelling, where the pre-processed data are used to train unsupervised algorithms for automatic anomaly detection. Models such as DBSCAN, Isolation Forest, K-Means, Local Outlier Factor (LOF), and One-Class SVM were selected due to their ability to operate without explicit labels, which is advantageous in water networks where leak events are rare and difficult to annotate.

Within this stage, the dataset's inherent class imbalance is addressed using data balancing techniques to improve model sensitivity. Balanced Random Oversampling (ROS) replicates minority-class samples to equalise class proportions, serving as a baseline approach. Synthetic Minority Oversampling Technique (SMOTE) generates synthetic samples by interpolating between minority-class neighbours, expanding the feature space and reducing bias toward normal operating conditions. Adaptive Synthetic Sampling (ADASYN) further refines this process by generating more synthetic samples in sparse or difficult-to-learn regions, allowing the models to better capture subtle anomaly patterns.

These techniques collectively enhance the learning capability of the models, improving their ability to detect rare but critical leak and burst events. This logic-based methodology is applicable to any pumping/ transmission main in water supply systems and EPANET model was used for the identification of optimal location of Pressure Sensors.

Class Imbalance Handling:

Leak events are inherently rare in water distribution systems, as networks predominantly operate under normal conditions with only occasional abnormal events such as leaks or bursts (Romano et al., 2014). Consequently, the dataset used in this study is highly imbalanced, with significantly fewer leak observations compared to normal operating conditions. This class imbalance can bias

Machine Learning models toward the dominant “normal” class, thereby reducing their ability to accurately detect rare leak events.

Balanced Random Oversampling (ROS):

Random Over Sampling balances the dataset by duplicating existing minority class samples until a more uniform class distribution is achieved (He & Garcia, 2009). This method is computationally simple and preserves the original data distribution, making it a baseline approach for comparison with more advanced techniques. However, excessive duplication may lead to overfitting and reduced generalisation.

Synthetic Minority Oversampling Technique (SMOTE):

SMOTE generates synthetic samples by interpolating between minority class instances and their nearest neighbours, thereby expanding the minority decision region (Chawla et al., 2002). This reduces bias toward the majority class and improves classification performance in imbalanced datasets. In this study, SMOTE improved model performance, as reflected in evaluation metrics such as F1-score and ROC-AUC.

Adaptive Synthetic Sampling (ADASYN):

ADASYN extends SMOTE by generating synthetic samples adaptively, focusing more on difficult-to-learn minority instances (He et al., 2008). This approach improves learning for complex decision boundaries but may introduce noise in highly imbalanced or noisy datasets.

Implementation Strategy:

The oversampling methods were incorporated within each training fold of a stratified 5-fold cross-validation scheme. Resampling was applied exclusively to the training subset to prevent information leakage into the test data. The models were then retrained and evaluated on untouched validation sets using *Anomaly F1* and *ROC-AUC* metrics.

Once the models are trained using balanced datasets, the model-validation stage quantifies their predictive performance through rigorous cross-validation. Stratified five-fold cross-validation is

employed to maintain class-distribution consistency across all training and testing partitions. For each fold, oversampling is applied exclusively to the training subset to prevent data leakage, after which models are evaluated on untouched validation sets. Two performance metrics, Anomaly F1 Score and Receiver Operating Characteristic Area Under Curve (ROC-AUC) are computed to assess detection accuracy and discriminative capability. The comparison between imbalanced and balanced datasets demonstrates the positive impact of ROS, SMOTE, and ADASYN on minority-event recognition, with ADASYN delivering the highest overall gains in F1 and AUC for the LOF algorithm. This empirical validation confirms that class-balancing strategies significantly improve leak-detection reliability without compromising generalisation.

Performance evaluation followed two principal metrics. The *Anomaly F1 score* captures the harmonic mean of precision and recall for the minority (leak) class, penalising false negatives, while *ROC-AUC* measures discriminative capacity independent of threshold. Stratified 5-fold cross-validation ensured statistical robustness. The results demonstrated substantial performance gains: the LOF model, for example, improved from $F1 = 0.12$ ($AUC = 0.77$) on imbalanced data to $F1 = 0.69$ ($AUC = 0.72$) after ADASYN balancing. Isolation Forest also exhibited stable accuracy under all three balancing techniques, confirming generalisation reliability.

The final component of the framework is integration and decision support, which translates the analytical results into actionable operational intelligence. The validated models are conceptually integrated within a SCADA or Human-Machine Interface (HMI) environment, where real-time sensor data are continuously processed by the trained anomaly-detection system. Detected anomalies trigger automated alerts or alarms, providing operators with early warnings of potential leaks or bursts. The output interface presents model insights such as location likelihood and anomaly severity in a form accessible to maintenance personnel. This decision-support layer bridges the gap between data analytics and operational response, ensuring that the research outcomes contribute directly to practical water-utility management.

In summary, the research methodology framework represents a comprehensive, multi-layered approach to intelligent water-network monitoring. Beginning with reliable data acquisition and progressing through systematic pre-processing, rule-based assessment, data-driven modelling, and robust validation, it ensures a balance between engineering interpretability and computational intelligence. The integration of Balanced ROS, SMOTE, and ADASYN within the Machine

Learning stage constitutes a key innovation of this work, directly addressing one of the principal challenges in hydraulic anomaly detection class imbalance. Together, these stages establish a reproducible, scalable, and scientifically sound pathway from raw SCADA data to real-time leak and burst decision support, thereby advancing the state of intelligent water-distribution management.

Machine Learning significantly enhances the detection of leaks and bursts in water pipelines through advanced pattern recognition. Using supervised or unsupervised learning techniques, these systems can detect leaks and bursts at an early stage. This reduces water loss, enhances maintenance efficiency, and provides scalable, adaptable solutions for modern pipeline networks.

Alert generation:

As the system operates, the algorithm continuously monitors time-pressure data from the sensors. It compares the current pressure readings with the established baseline to identify any deviations. The algorithm employs logic-based rules to identify patterns or anomalies in the pressure data that may indicate a leak. Common logic includes looking for sudden pressure drops, fluctuations outside expected ranges, or patterns indicative of leaks. When the algorithm detects a deviation beyond the predefined thresholds or violates the established logic-based rules, it triggers an alert. The alert may include information about the burst occurrence time, approximate distance of the burst location, and potential cause of the detected anomaly including the pressure values, flow meter reading, and pump speed. The alerts of leaks and bursts are shown in Figure 6.1, Chapter 6.

4.4 Chapter Summary and Reflections

This chapter presented a comprehensive methodology for the intelligent detection of leaks and bursts within water distribution networks. The approach combined three core elements which are logic-based algorithms, Machine Learning models, and experimental validation to form a hybrid detection framework capable of real-time, adaptive, and scalable operation.

The logic-based component provided interpretability and quick decision-making using deterministic rules derived from hydraulic theory and SCADA data. It served as the first line of anomaly detection, identifying deviations in flow, pump speed, and pressure patterns. In parallel,

the integration of Machine Learning algorithms, including both supervised and unsupervised models, addressed the limitations of static thresholding by learning complex, nonlinear relationships between hydraulic parameters. The hybrid combination of these two approaches ensured improved detection sensitivity and reduced false alarms.

Furthermore, the methodology incorporated data pre-processing, class-balancing techniques (ROS, SMOTE, ADASYN), and cross-validation frameworks to ensure robustness and generalisability of the results. By bridging experimental, numerical, and analytical dimensions, the research demonstrated a cohesive pathway from raw sensor data to actionable insights for operational decision-making.

While the proposed methodology is designed to enhance leak and burst detection, several practical limitations should be acknowledged. These include the dependency on high-quality and high-frequency SCADA data, the computational requirements associated with real-time Machine Learning implementation, and the need for periodic model retraining as network conditions evolve. These considerations highlight areas for further refinement and practical implementation in real-world water distribution systems.

Chapter 5 (Data Collection) presents the datasets and data acquisition protocols employed in this study, including sensor configurations, sampling intervals, and pre-processing procedures. Chapter 6 (Application of Methodology) applies the developed logic-based and Machine Learning frameworks to these datasets for performance evaluation and validation.

Chapter 5

Data Collection

5.1 General Remarks

This chapter presents the complete set of hydraulic datasets used in this research and establishes the empirical basis for the subsequent analytical and detection frameworks. As introduced in Chapter 3, two complementary data environments were utilised: the Sunbury pumping main, which provides long-term real-world SCADA measurements from an operational trunk main, and the Melton prototype system, which offers high-resolution, spatially distributed measurements from a fully controlled experimental pipeline. These two systems collectively capture both the complexity of live network behaviour and the clarity of controlled hydraulic conditions.

The Sunbury dataset represents authentic operational patterns, including pump cycling, demand-driven variations, transient behaviour, and naturally occurring leak and burst events. In contrast, the Melton prototype system provides a noise-minimised, repeatable environment where pressure and flow responses can be observed under stable baseline conditions. Data from both systems were recorded through field instrumentation including pressure sensors, flow meters, and pump-speed telemetry integrated with supervisory control systems to ensure consistent and continuous collection over their respective monitoring periods.

By combining real-world SCADA data with controlled experimental measurements, this chapter documents the full empirical foundation required for analysing hydraulic behaviour, distinguishing baseline and abnormal states, and supporting the detection approaches presented in later chapters. The datasets described herein form a robust and reproducible benchmark for understanding water distribution system dynamics across both operational and experimental contexts.

5.2 Sunbury Pumping System Data Collection

The Sunbury dataset provides the principal real-world hydraulic observations for this research, consisting of long-duration SCADA measurements that capture natural operational variability within an actively managed rising main. The extended monitoring period includes a rich spectrum of behaviours arising from pump operations, demand fluctuations, and transient events, enabling the identification of both stable and abnormal hydraulic states. In addition to the full long-term dataset, a shorter five-day segment was isolated to assess dataset consistency and support independent evaluation. Together, these data windows offer a comprehensive view of the system's dynamic performance and form a critical empirical basis for distinguishing leak and burst patterns, validating hydraulic relationships, and supporting the analytical stages that follow.

5.2.1 SCADA Data and Monitoring Periods

The Sunbury pumping main dataset was obtained from the SCADA system operating along the 6.24 km rising main that transfers water from the pump station to the storage tank. The SCADA system continuously records key hydraulic variables, including pressure at the pump running condition, flow through the rising main, and pump rotational speed. Two monitoring windows were used in this study. The first is a long-term dataset spanning 139 days from 1 May 2022 to 17 September 2022, comprising 200,161 samples and capturing a wide range of operational conditions. The second is a short-term testing dataset covering 1 May to 6 May 2022, consisting of 7,201 samples, selected to provide an independently assessable segment of data. Both datasets include a binary label indicating leak, burst, and non-leak states, enabling direct comparison of hydraulic behaviour under different operating conditions. These SCADA records form the primary real-world empirical basis for the analyses presented in the subsequent sections.

SCADA serves as a centralised platform for collecting real-time data from various sensors and devices installed in the pipeline network, enabling effective decision-making and efficient system management. The setup is tailored for training purposes, simulating real-world scenarios to equip operators with the skills needed to handle operational complexities in water distribution systems. From the sensor data, noise and outliers are removed before being used as inputs to the model. A sample of the historical data collected for analysis and development of the algorithm is shown in Table 5.1.

Table 5.1: Extracted SCADA data from the flow meter, pressure sensor, and the pump speed for the 450 mm pipeline.

Date and time	Pressure (m)	Flow (LPS)	Pump Speed (rpm)
01/05/2022 08:05	77.46	252.67	1500
01/05/2022 08:06	77.46	253	1500
01/05/2022 08:07	77.46	252.83	1500
01/05/2022 08:08	77.46	252.67	1500
01/05/2022 08:09	77.2	252.5	1500
01/05/2022 08:10	77.2	252.33	1500
01/05/2022 08:11	77.2	252.17	1500

With the monitoring periods and the structure of the SCADA records established, the next step is to summarise the key statistical characteristics of the training dataset to provide an initial understanding of the system’s hydraulic behaviour under both leak, burst, and non-leak conditions.

5.2.2 Training Dataset Summary Statistics

The long-term SCADA training dataset from the Sunbury pumping main consists of 200,161 observations collected over a 139-day period. The variables recorded include pressure (P), flow (F), pump speed (N), and an associated leak label indicating the presence or absence of leak conditions. Of the total samples, 91 observations (0.05%) were labelled as leak events, reflecting the rarity of leak conditions in real operational systems. Summary statistics for each variable under leak and non-leak states are presented in Table 5.2, illustrating the distinct hydraulic behaviours characterising normal and abnormal operating periods. Table 5.2, the data collected are under non-leak conditions, including pump start-up and shutdown events, as well as normal pump running conditions. The leak data, however, was collected only during the pump running condition.

Table 5.2: Summary statistics for the Sunbury training dataset

Variable	Condition	Mean	Std Dev	Min	25%	50%	75%	Max
Pressure (m)	Non-leak	63.86	7.26	50.16	59.51	59.94	68.99	92.20
Pressure (m)	Leak	62.13	1.85	58.32	61.05	61.88	62.77	67.93
Flow (L/s)	Non-leak	59.49	98.06	0	1	1	159	272
Flow (L/s)	Leak	238	20.98	197	245	246.5	251	255
Pump Speed (rpm)	Non-leak	349.34	588.79	0	0	0	1039.2	1500
Pump Speed (rpm)	Leak	1429.61	129.43	1178.0	1500	1500	1500	1500

Low flow values (0–1 L/s) and zero pump-speed readings are characteristic of the non-leak operating regime, as the pump remains idle for extended periods during tank-filling cycles and off-peak demand conditions. These idle intervals result in minimal or no flow through the rising main and account for a substantial portion of the non-leak measurements in the Sunbury dataset. Clear differences were observed between leak and non-leak periods. Leak events were characterised by substantially higher flow rates (typically above 200 L/s) and pump speeds consistently near the operational maximum (1,450–1,500 rpm). In contrast, non-leak behaviour was dominated by extended intervals of low or zero flow, corresponding to tank-filling cycles and pump idle states. Pressure exhibited comparatively smaller separation between conditions, remaining within a narrow operational band, but with slightly elevated values during confirmed leak events. These statistics highlight the strong hydraulic signatures associated with leaks and provide a quantitative foundation for the distributional and relational analyses that follow.

To expand upon the statistical profiles presented above, the next step examines the probability distributions of pressure, flow, and pump-speed measurements to understand how these variables behave under both leak and non-leak conditions.

5.2.3 Training Dataset Distribution Analysis

Kernel Density Estimation (KDE) was used to analyse the distributional behaviour of the hydraulic variables in the Sunbury training dataset. KDE is a non-parametric smoothing technique that estimates the probability density function of a continuous variable, enabling visualisation of how measurements are distributed across different operating conditions. The figure for this section contains three KDE plots (see Figure 5.1): the top plot illustrates the pressure distribution for leak and non-leak states, the middle plot shows the corresponding flow distributions, and the bottom plot depicts the distribution of pump-speed values. Together, these three plots provide a comprehensive view of how each hydraulic variable behaves under normal and leak scenarios.

The pressure KDE shows a dominant and narrow peak around 60 m during non-leak periods, reflecting the stable discharge pressure maintained during regular pump operations. In contrast, leak periods exhibit a broader and slightly elevated distribution, with values extending above 62 m, indicating increased hydraulic loading during sustained high-flow events. The flow KDE reveals the strongest separation between classes: non-leak data are heavily skewed toward very low flow values (0–1 L/s) due to prolonged pump idle periods, with a secondary spread across 150–250 L/s during active pumping cycles. Leak conditions, however, form a tight and distinct peak between 230 and 255 L/s, representing the consistently elevated flow produced during confirmed leakage. The pump-speed KDE follows a similar pattern, with non-leak data exhibiting a dual-mode distribution frequent 0 rpm readings when the pump is off and a broad range up to approximately 1500 rpm during normal operation while leak, burst events generate a sharp, narrow peak concentrated near the maximum operational speed of 1450–1500 rpm.

Collectively, these KDE plots highlight clear distributional contrasts between leak, burst, and non-leak states. Pressure shows moderate separation, while flow and pump speed display highly distinctive signatures, reflecting the substantial hydraulic differences caused by leakage. These patterns provide an essential foundation for the relational and anomaly-based analyses presented in the following sections.

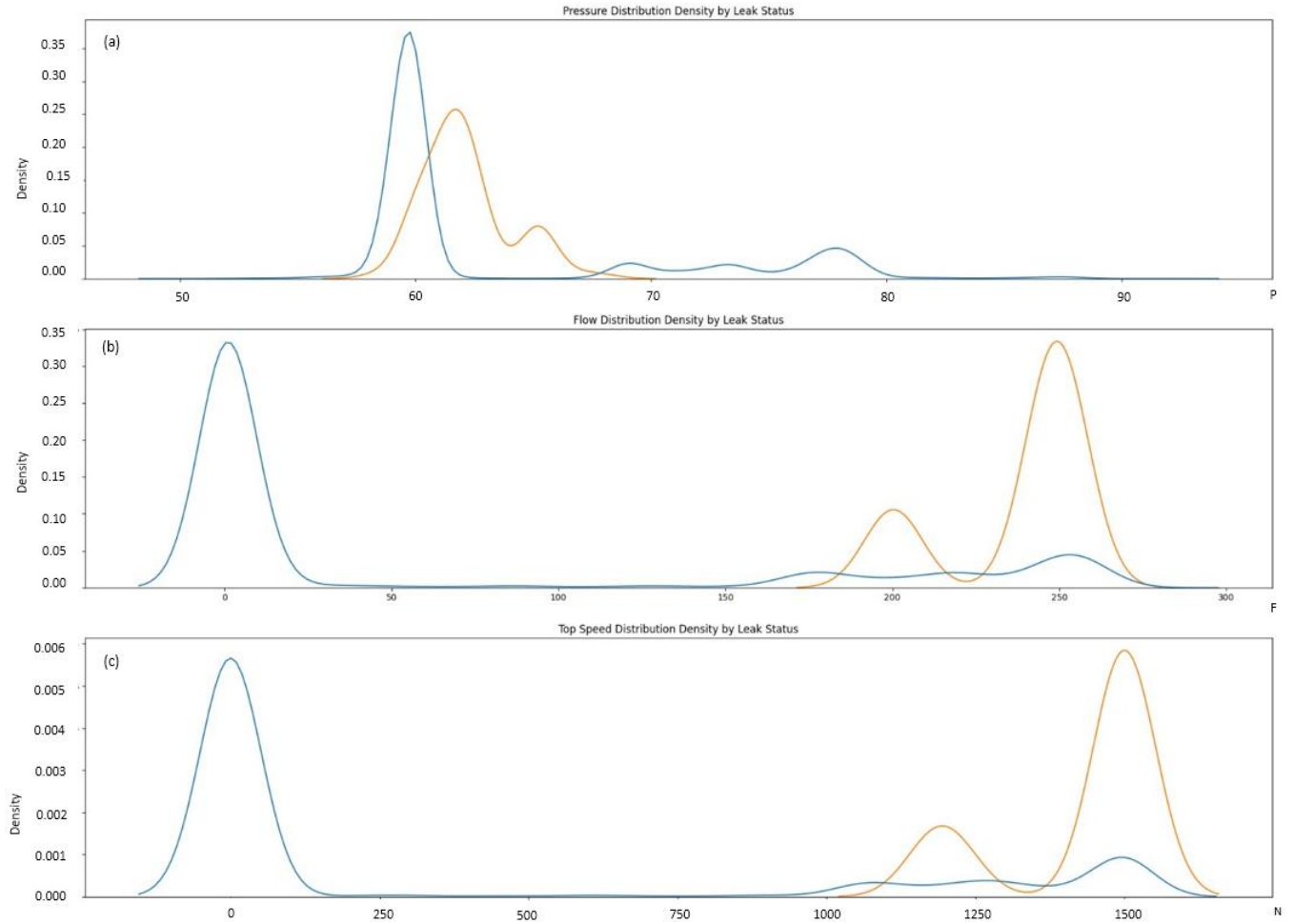


Figure 5.1: Kernel Density Estimates (KDE) of (a) pressure, (b) flow, and (c) pump-speed for leak and non-leak conditions in the Sunbury training dataset.

While distributional patterns offer insight into individual variables, analysing how these variables interact is equally important. The following section explores the pairwise relationships between pressure, flow, and pump speed to reveal their hydraulic coupling.

5.2.4 Training Dataset Relationship Analysis

Scatterplots and correlation analysis were used to examine the relationships between pressure, flow, and pump speed in the Sunbury training dataset (See figure 5.2). The three pairwise scatterplots are Pressure vs Flow, Pressure vs Pump Speed, and Flow vs Pump Speed provide a detailed view of how these hydraulic variables interact under leak and non-leak conditions. A

corresponding correlation heatmap summarises the strength of these relationships across the full dataset.

The Pressure–Flow relationship shows two distinct behavioural regions. Non-leak data form a broad, dispersed cloud with large clusters near zero flow, reflecting extended pump idle periods, and a more scattered spread at higher flow values during pumping cycles. Leak, and burst events occupy a narrow, elevated band of high flows (230–255 L/s) paired with moderately increased pressures, demonstrating the hydraulic loading produced by sustained leakage. This separation indicates that flow is highly sensitive to leak status, while pressure responds more subtly but still exhibits measurable elevation during leaks.

The Pressure–Pump Speed relationship reveals a similarly structured pattern. Non-leak measurements show a dual distribution, with a dense cluster near 0 rpm corresponding to pump-off periods and a widespread up to approximately 1500 rpm when the pump is operating. Pressure rises as pump speed increases, producing a characteristic diagonal trend. Leak events are tightly concentrated at the upper end of the pump-speed range (1450–1500 rpm), paired with slightly elevated pressures, reflecting the pump’s continuous high-speed operation during confirmed leak conditions.

The Flow–Pump Speed relationship displays the clearest separation between leak and non-leak states. Non-leak behaviour again shows a dense cluster near zero flow at 0 rpm, with a wide vertical distribution representing varying flow outputs at intermediate pump speeds. Leak events form a highly compact cluster of high flow and near-maximum pump speed, creating an isolated region distinct from the non-leak distribution. This strong separation demonstrates that combined flow and pump-speed measurements form the most discriminative pair of variables for identifying leakage in the Sunbury system.

The correlation heatmap supports these visual patterns, showing very strong positive correlations between flow and pump speed ($r \approx 0.95$) and between pressure and flow ($r \approx 0.88$), and a slightly weaker but still substantial correlation between pressure and pump speed ($r \approx 0.83$). These values reflect the underlying hydraulic coupling of the pumping main: pump operation directly drives flow, which in turn influences system pressure. Collectively, the scatterplots and correlation matrix

confirm that leak events produce tightly clustered and hydraulically consistent signals, while non-leak periods exhibit far greater variability due to normal operational cycling.

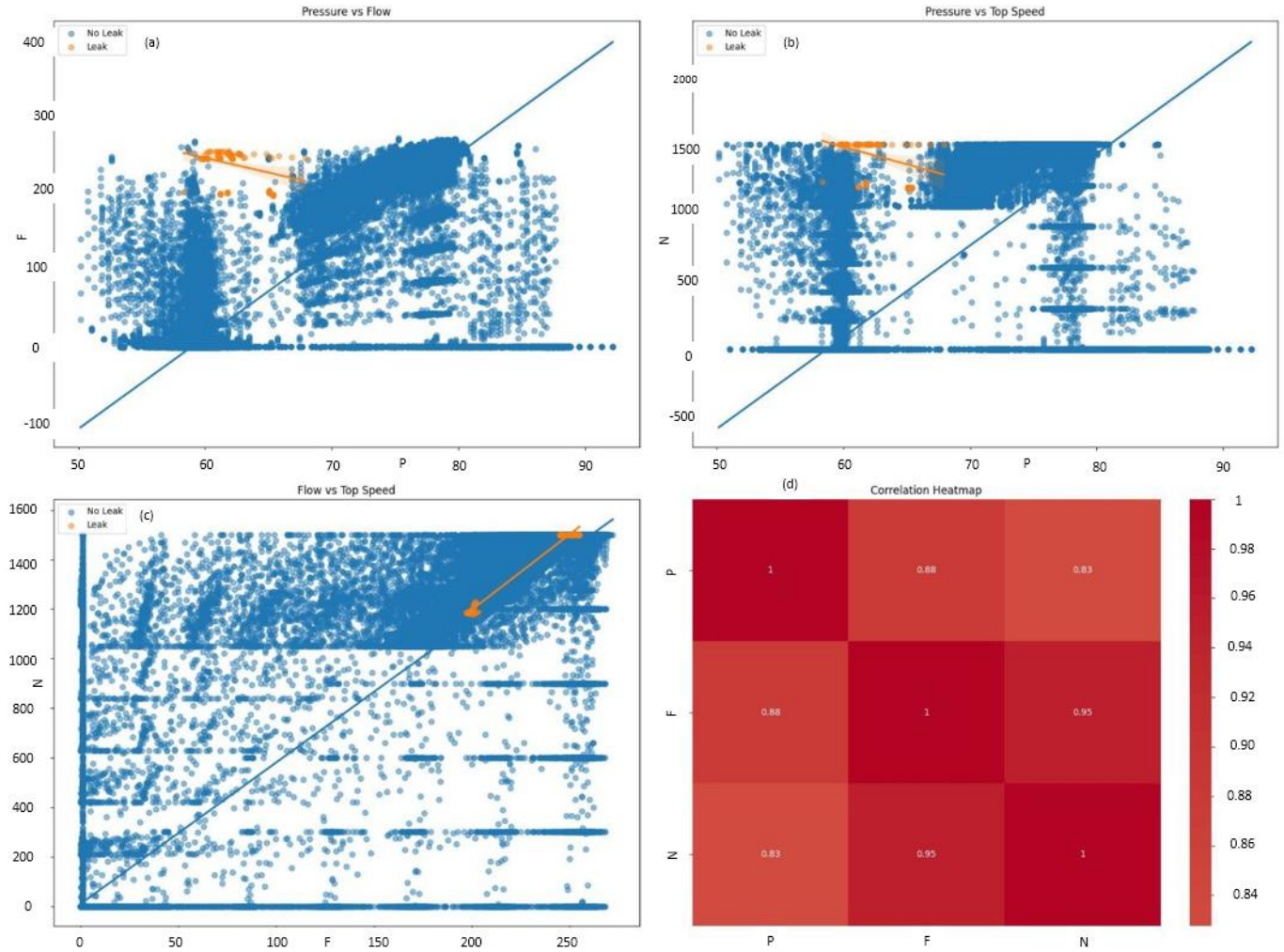


Figure 5.2: Pairwise relationships (a) (Pressure–Flow, (b) Pressure–Pump Speed, (c) Flow–Pump Speed) and (d) correlation heatmap for the Sunbury training dataset.

Having established the characteristics of the long-term training dataset, attention now shifts to the testing dataset. This subset enables an independent assessment of consistency across different monitoring windows.

5.2.5 Testing Dataset Distribution Analysis

The Sunbury testing dataset consists of 7,201 samples collected over a five-day period from 1 May to 6 May 2022, providing an independent window for validating the hydraulic consistency of the system. Of these observations, 7,168 samples (99.54%) correspond to non-leak conditions, and 33 samples (0.46%) confirmed leak events. Summary statistics for pressure, flow, and pump speed under both operating states are presented in Table 5.3. Similar to the training dataset, the non-leak period is dominated by extended pump-idle intervals, reflected in low flow readings (0–1 L/s) and recurrent pump-speed values of 0 rpm. Leak periods exhibit clearly elevated flow (mean ≈ 233 L/s) and consistently high pump-speed values (mean ≈ 1471 rpm), providing strong separation between operating conditions. Pressure remains comparatively stable across both states, showing only modest increases during leak events. These findings confirm that the testing dataset preserves the key hydraulic patterns required for independent model evaluation.

Table 5.3: Summary statistics for the Sunbury testing dataset

Variable	Condition	Mean	Std Dev	Min	25%	50%	75%	Max
Pressure (m)	Non-leak	62.95	6.75	51.87	59.58	59.89	60	86.53
Pressure (m)	Leak	62.17	2.87	58.49	59.32	62.45	64.96	69.43
Flow (L/s)	Non-leak	47.79	92.76	0	1	1	1	262
Flow (L/s)	Leak	232.9	15.71	203.67	219.67	233	247	255
Pump Speed (rpm)	Non-leak	281.11	559.70	0	0	0	0	1500
Pump Speed (rpm)	Leak	1470.87	47.29	1329	1455.6	1500	1500	1500

Low flow (0–1 L/s) and zero pump-speed values dominate non-leak conditions because the pump remains idle for extended periods during tank-filling cycles and low-demand intervals, whereas leak periods show consistently elevated flow and near-maximum pump-speed values.

It should be noted that the observed spread in the scatter plots arises from the coexistence of pump idle and active operational (steady state) states; however, the correlation coefficients are primarily influenced by strong linear relationships during active pumping conditions.

5.2.6 Testing Dataset Relationship Analysis

Relationship analysis for the testing dataset was conducted (see Figure 5.3) using pairwise scatterplots and a correlation heatmap to examine how pressure, flow, and pump speed interact

over the five-day monitoring period. Similar to the training dataset, the Pressure–Flow scatterplot shows two clearly distinguishable operating regions: a broad non-leak cluster aligned with low or zero flow values, and a compact leak cluster positioned at high flows and moderately elevated pressures. This pattern reflects the same hydraulic behaviour observed in the long-term dataset, where pump idle periods generate low-pressure/low-flow combinations and leak periods sustain consistently high-flow conditions.

The Pressure–Pump Speed scatterplot shows a dual-mode structure in the non-leak data, with a dense grouping at 0 rpm during pump shutdown and a dispersed range of pressure values as pump speed increases during normal operation. Leak events again appear as a small, tightly grouped cluster near the maximum pump speed (1450–1500 rpm), accompanied by a narrow band of moderately higher pressures. This reinforces the strong operational linkage between high pump-speed operation and leak-induced hydraulic loading.

The Flow–Pump Speed relationship exhibits the strongest separation between operating states. Non-leak data populate two dominant regions: zero flow at zero speed, and a scattered spread of flows at intermediate speeds reflecting normal duty cycling. Leak observations occupy a highly compact region pairing near-maximum pump speed with flows above 230 L/s, forming an isolated cluster that does not overlap with any non-leak behaviour.

The correlation heatmap further supports these findings, showing very strong positive correlations between flow and pump speed ($r \approx 0.954$) and between pressure and flow ($r \approx 0.953$), as well as a strong correlation between pressure and pump speed ($r \approx 0.897$). These values are slightly higher than those observed in the training dataset, suggesting more consistent operational behaviour during this shorter monitoring window. Overall, the relationship analysis confirms that the testing dataset preserves the same hydraulic patterns and separability observed in the long-term dataset, making it a reliable dataset for independent evaluation of detection performance.

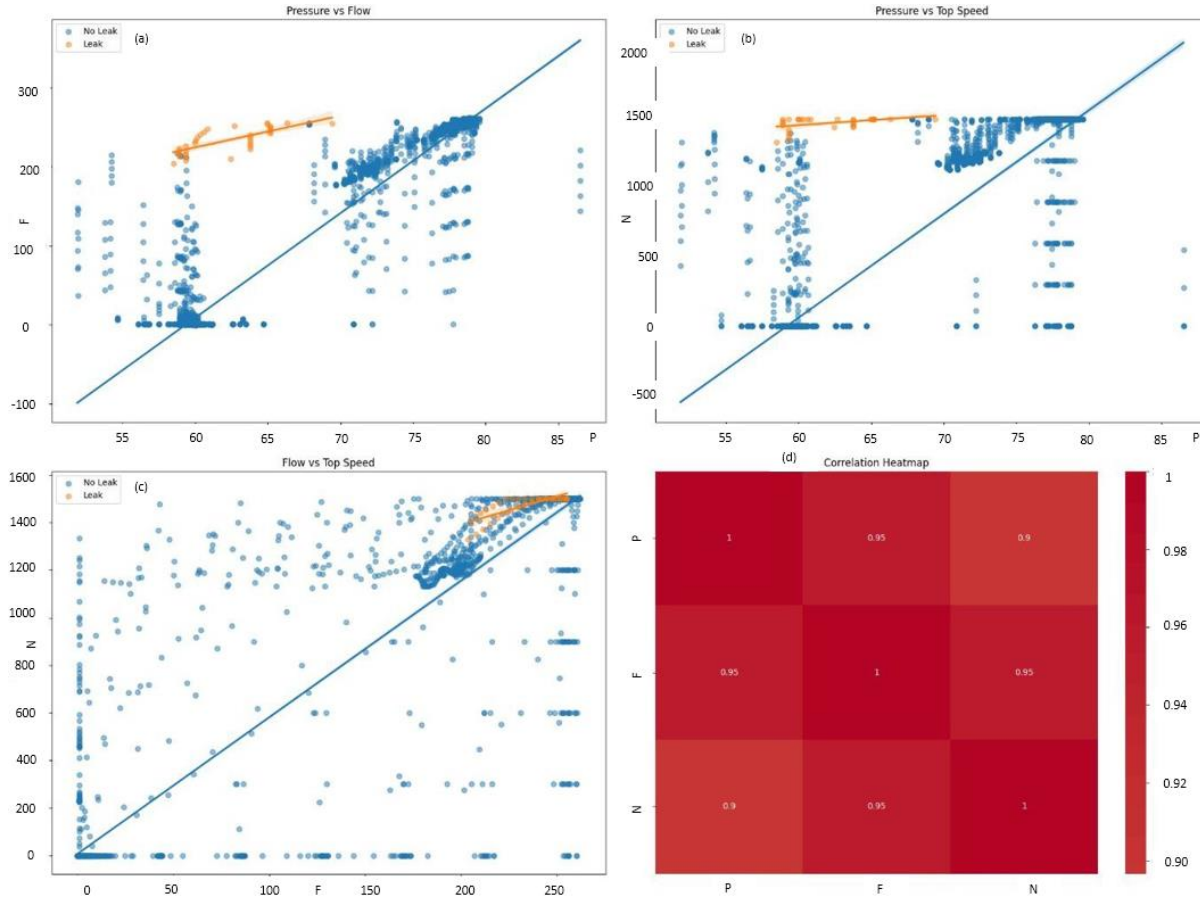


Figure 5.3: Pairwise scatterplots ((a) Pressure–Flow, (b) Pressure–Pump Speed, (c) Flow–Pump Speed and (d) correlation heatmap for the Sunbury testing dataset.

In addition to steady-state patterns, the testing dataset contains operational shifts that are expressed as statistical anomalies. The following analysis highlights these deviations and their significance.

5.2.6 Testing Dataset Anomaly Detection

Anomaly detection for the testing dataset was performed using Z-score analysis (see Figure 5.4) to identify measurements that deviated significantly from the normal hydraulic operating range. The Z-score time-series plots for pressure, flow, and pump speed over the five-day monitoring period illustrate how each variable behaves relative to its statistical mean and standard deviation. A threshold of $Z = 3$ is shown on each plot as a conventional indicator for identifying outliers, allowing operational deviations to be easily distinguished from steady-state behaviour.

The pressure Z-score plot shows occasional excursions beyond the ± 3 threshold, reflecting small transient fluctuations associated with pump start-ups and shutdowns. However, the majority of pressure values remain tightly grouped near the mean, consistent with the stable discharge pressure maintained in the rising main. In contrast, the flow Z-score plot exhibits a much larger number of high-magnitude deviations. These spikes correspond to the system's transition between idle and active pumping states, where flow increases abruptly from near-zero values to more than 230 L/s. Several of the leak observations fall well above the anomaly threshold, demonstrating that flow responds strongly to leak-induced hydraulic loading.

The pump-speed Z-score plot shows the clearest distinction between leak and non-leak conditions. Non-leak periods generate a dual-mode distribution with Z-scores clustering near zero during pump inactivity and spreading across a moderate range when the pump is running. Leak periods, however, are consistently associated with extremely high Z-scores, reflecting pump-speed values near the maximum operational limit of 1500 rpm. These elevated Z-scores create a distinct anomaly pattern that is visually and statistically separable from non-leak operation.

Overall, the anomaly detection results confirm that flow and pump speed provide strong, high-contrast anomaly signals, while pressure remains comparatively stable and less sensitive to leak severity. The testing dataset therefore reinforces the findings from the long-term training dataset: leaks generate consistent and easily identifiable anomaly signatures, particularly in flow and pump-speed variables, making these measurements highly suitable for use in detection and classification modelling.

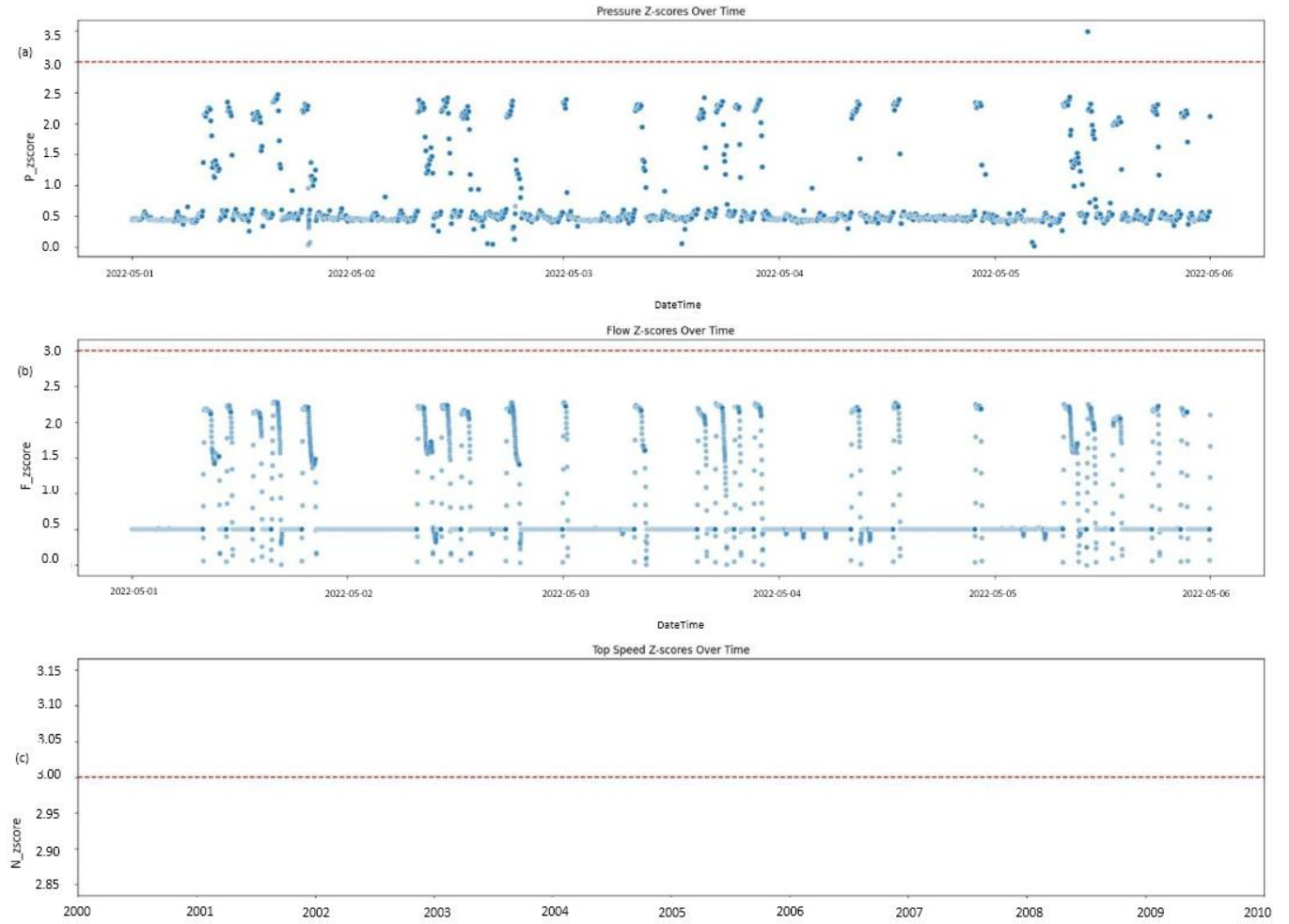


Figure 5.4: Z-score anomaly detection for (a) pressure, (b) flow, and (c) pump speed in the Sunbury testing dataset.

5.2.7 Rolling Correlation Analysis

Rolling correlation analysis was used (Figure 5.5) to examine how the relationship between pressure and flow evolves over time in the Sunbury testing dataset. Unlike static correlation metrics, which summarise the overall association across the entire dataset, rolling correlation provides a time-varying measure that reflects how hydraulic coupling responds to operational cycles. A 60-minute sliding window was applied to compute the correlation between pressure and flow, producing a continuous profile of how strongly these variables interact during different pumping modes.

The resulting plot reveals three dominant behavioural phases. During active pumping periods, the rolling correlation remains strongly positive, reflecting the expected hydraulic response in which

increases in pump discharge lead to simultaneous increases in flow through the rising main. When the pump transitions into shutdown or idle periods, the rolling correlation shifts sharply toward negative values as flow drops to near zero while pressure gradually decays toward its static level. Between these two extremes, the correlation occasionally stabilises near zero, indicating neutral or weak coupling during brief transitional intervals when neither variable exhibits strong directional change.

These cyclical shifts in correlation align closely with the pump’s operational duty pattern, demonstrating that the pressure–flow relationship is highly dynamic and sensitive to the system’s on/off states. The rolling correlation analysis therefore complements the earlier static correlation results by illustrating how hydraulic interactions strengthen or weaken over time, providing additional insight into the behavioural signatures that distinguish normal operation from leak-affected conditions.

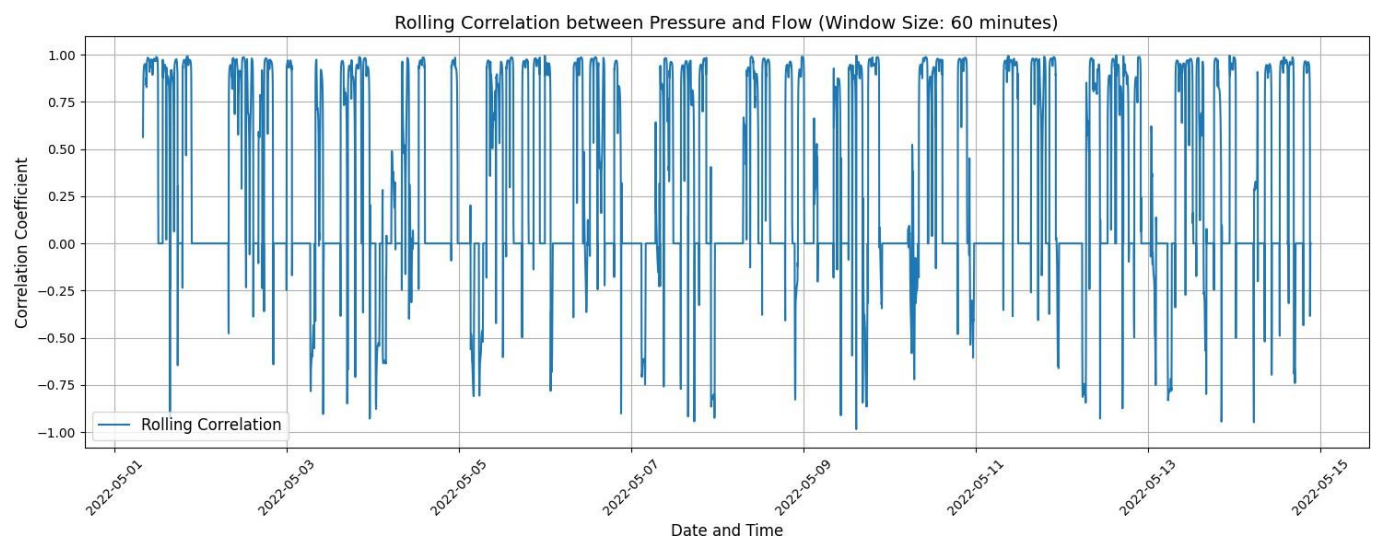


Figure 5.5: Rolling correlation between pressure and flow (60-minute window) for the Sunbury testing dataset.

5.2.8 Pressure Flow Time-Series Overlay

The combined time-series plot of pressure and flow provides a direct visual comparison of how these variables co-evolve during pump cycling and leak conditions (Figure 5.6). The flow signal exhibits sharp transitions between zero and high values, corresponding to pump shutdowns and

activations, while pressure varies more gradually but still reflects the underlying operational changes. When the pump activates, flow rises rapidly to values exceeding 200 L/s, and the pressure signal shows a corresponding increase, reflecting the transition from static to dynamic head conditions. Conversely, when the pump stops, flow drops immediately to zero, while pressure decays more slowly toward its idle level.

The temporal alignment of these changes highlights the strong operational coupling between pressure and flow in the Sunbury system. The distinct step-like behaviour in the flow signal and the smoother but synchronous variations in pressure illustrate the hydraulic response of the rising main to pump operation. Leak conditions magnify this coupling by maintaining consistently high flow and elevated pressures over extended intervals, producing sustained deviations from normal cycling patterns. This overlay therefore provides a useful reference for interpreting the rolling correlation behaviour and reinforces the characteristic signatures of leak and non-leak states.

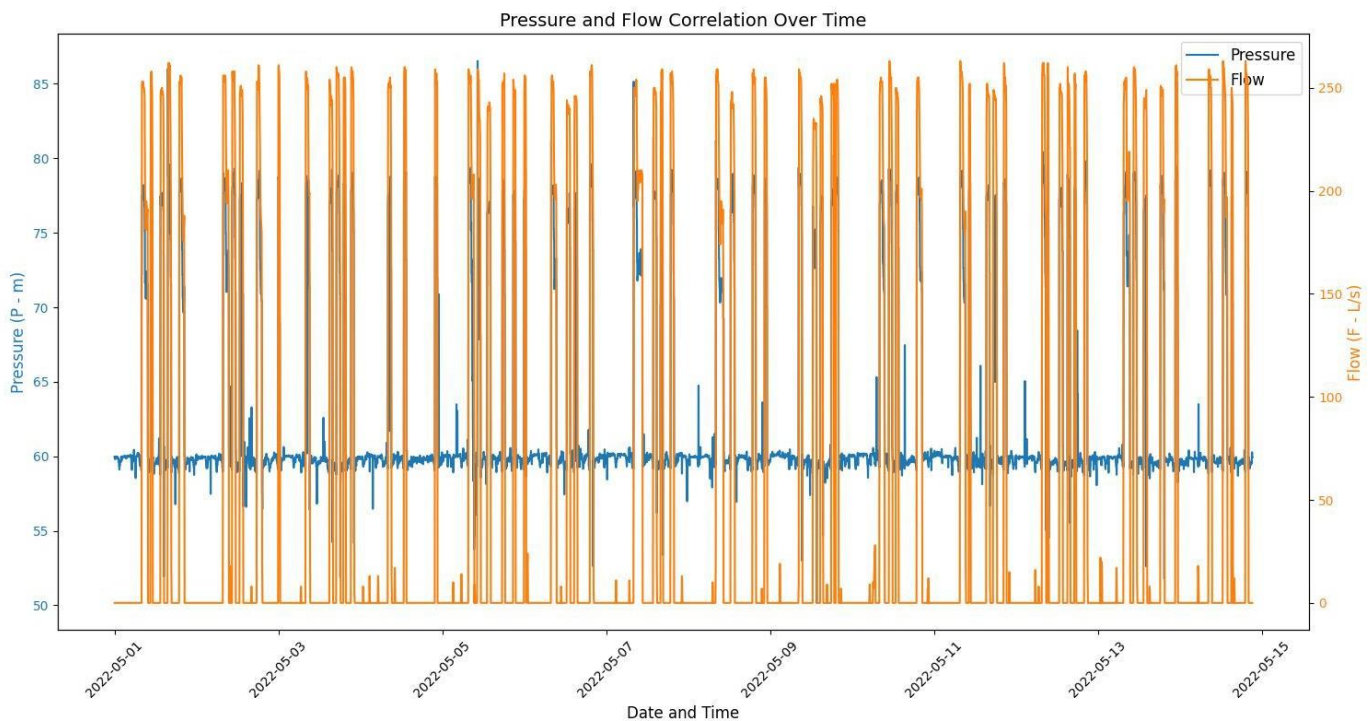


Figure 5.6: Time-series overlay of pressure and flow measurements in the Sunbury testing dataset, illustrating hydraulic response during pump on/off cycles.

As described in Chapter 3, the Sunbury pumping main system forms the operational dataset for SCADA-based analysis. The collected data and processing are detailed below.

5.3 Data Collection at Melton Prototype System

The Melton prototype system provided the controlled experimental dataset required for training and validating the leak and burst detection models. As the physical configuration, pipeline characteristics, and instrumentation have already been described in Chapter 3, this section focuses exclusively on the procedures used to generate, record, and prepare the data for analysis. The prototype consisted of a polyethylene pipeline instrumented with flow and pressure sensors, operated under both normal and induced leak conditions. During the experiments, leak events of different sizes were introduced to create measurable variations in hydraulic behaviour. These controlled disturbances enabled the collection of flow and pressure signatures that represent normal operation, leak conditions, and transitional states between them. All measurements were recorded through the SCADA telemetry system, which logged continuous, time-stamped readings at fixed sampling intervals.

Data acquisition involved running the pipeline under steady-state operation to capture baseline behaviour and subsequently introducing leaks of varying magnitudes to obtain the corresponding hydraulic responses. The recorded dataset therefore includes stable pressure–flow conditions, leak-induced deviations, and transitional periods associated with the initiation and cessation of leaks. Each of these operating states contributes to building a representative and diverse dataset suitable for both supervised and unsupervised Machine Learning approaches. Following acquisition, the SCADA telemetry outputs were exported to Excel format for pre-processing. The exported files included inlet and outlet flow meter readings, pressure measurements at the corresponding sensor locations, time stamps for every observation, and operational notes recorded during each experimental run. Column headings were standardised to ensure consistency across multiple exported files and to facilitate efficient data manipulation.

The raw dataset underwent several pre-processing steps to ensure its suitability for subsequent analysis. Missing or incomplete values caused by brief telemetry delays were imputed using appropriate statistical techniques, while erroneous spikes resulting from sensor resets or communication interruptions were removed to preserve data quality. Minor timing discrepancies between different sensor channels were corrected to ensure accurate synchronisation of flow and pressure records. Additional features were then generated to capture dynamic hydraulic behaviour,

including flow and pressure acceleration computed from successive time-step differences. Normalisation was applied to continuous variables to ensure consistent scaling, and event labels distinguishing leak from no-leak periods were added to support supervised model development.

Finally, the prepared dataset was divided into training and testing subsets using a stratified splitting approach. This ensured that both subsets preserved the underlying distribution of leak and non-leak events and avoided bias associated with class imbalance. The resulting datasets provide a robust empirical foundation for the Machine Learning analyses presented in subsequent chapters, enabling a fair and reliable evaluation of model performance under controlled but realistic hydraulic conditions.

5.3.1 Monitoring Period and Sensor Setup

The Melton prototype (see figure 3.5) dataset was collected over a continuous monitoring period spanning 4 July to 6 August 2024, during which pressure and flow measurements were recorded under controlled baseline operating conditions. The system consists of two pressure sensors (inlet Pressure (P1) and outlet Pressure (P2) installed along the experimental pipeline, together with an inlet Flow (F1) and outlet Flow (F2) electromagnetic flow meter positioned at the inlet and outlet sections. This configuration allows simultaneous observation of spatial pressure variation and flow behaviour across the length of the testbed. The absence of leak events during the monitoring window makes the dataset particularly useful for capturing stable baseline hydraulic patterns and establishing reference behaviour for comparison with real-world conditions.

5.3.2 Melton Dataset Summary Statistics

Summary statistics derived from the Melton dataset are presented in Table 5.4. Pressure values show clear spatial gradients, with inlet Pressure (P1) and outlet Pressure (P2). Flow measurements from the inlet Flow meter (F1) and outlet Flow meter (F2) meters show narrow ranges and low variability, consistent with the steady operating conditions maintained during data collection. These statistics confirm the internal consistency of the Melton dataset and provide a quantitative foundation for the subsequent distributional, temporal, and correlation analyses. To provide further insight into the temporal and statistical behaviour of the system, distribution plots of inlet and outlet pressure and flow under different leak conditions are presented in Figure 5.7. The figure illustrates how pressure and flow variables vary across no-leak and different leak scenarios. It can be observed that flow distributions exhibit clearer separation between leak severities, while

pressure variations are comparatively less distinct. These patterns highlight the importance of combining multiple variables for effective leak detection.

Table 5.4: Summary statistics for the Melton prototype dataset

Variable	Count of Samples	Mean	Std Dev	Min	25%	50%	75%	Max
Pressure (P1) (m)	1858	167.88	59.75	23	151	159	165	355
Flow (F1) (L/s)	1855	12.58	2.54	4.79	11.36	13.10	14.35	18.99
Pressure (P2) (m)	1855	146	61	26	126	135	140	350
Flow (F2) (L/s)	1858	12.5	2.21	4.49	11.92	13.17	13.86	16.94

The wide pressure ranges and narrow flow variations reflect the controlled hydraulic environment of the prototype system, where pressure is influenced by spatial position while flow remains stable due to constant boundary conditions.

5.3.3 Melton Dataset Distribution Analysis

Distributional behaviour of the Melton hydraulic variables was examined using KDE-style plots (pressure and flow across inlet and outlet flow meters). Figure 5.7 for this section illustrates the density patterns for both pressure and flow under stable, non-leak operation.

The pressure distributions show multi-modal characteristics at sensors, particularly at P1 and P2, where a wider range of pressures indicates sensitivity to pump-induced transients.

Flow distributions for the F1 and F2 exhibit narrow, unimodal patterns with small variance, indicating stable pump performance and consistent flow delivery throughout the monitoring window. These tightly clustered distributions highlight the low-noise nature of the Melton environment and its suitability for isolating pure hydraulic behaviour.

Overall, the distribution analysis confirms that the Melton dataset captures stable and predictable baseline characteristics, forming an ideal reference for later comparison with Sunbury's more variable conditions.

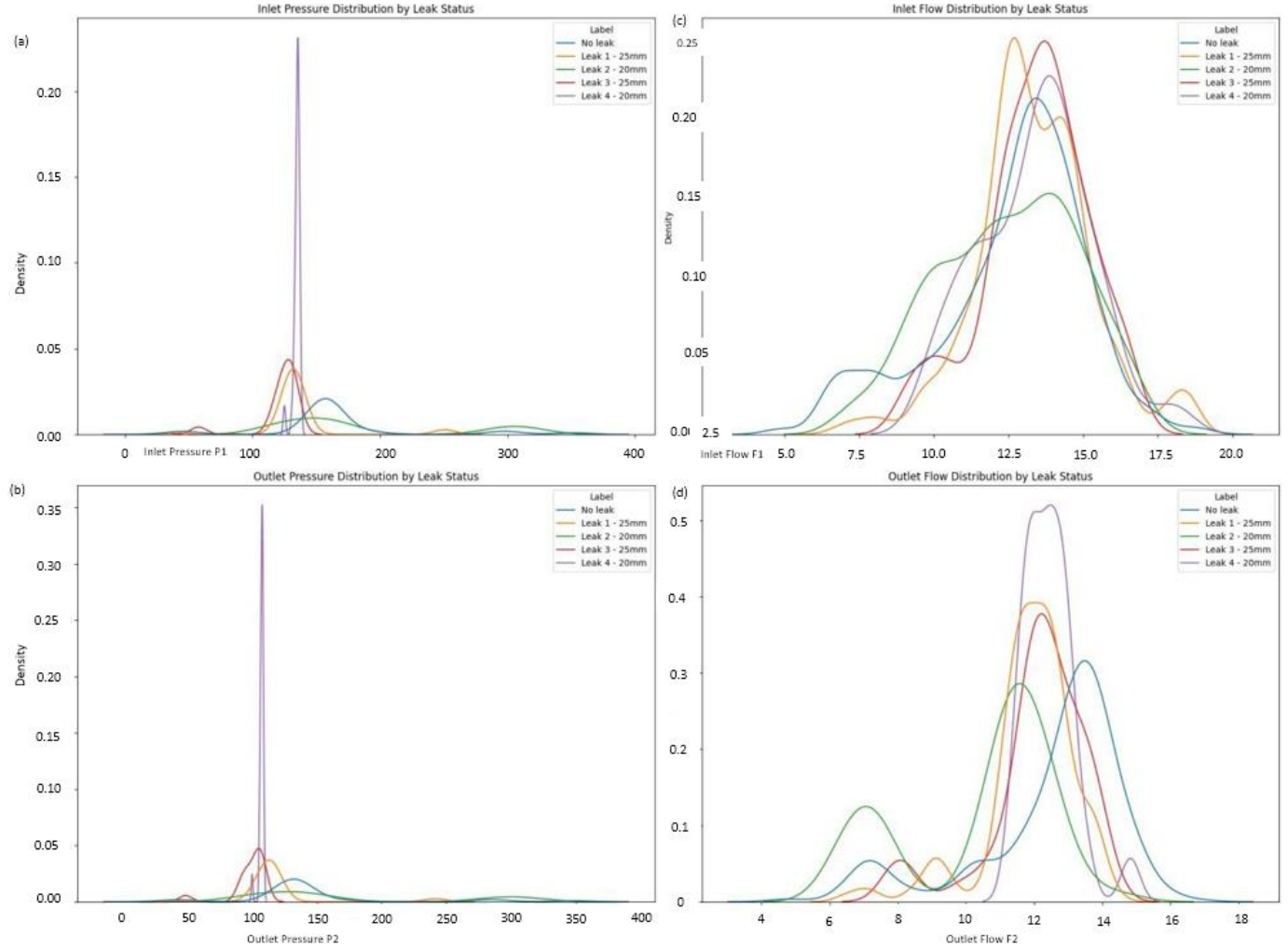


Figure 5.7: Distribution plots of (a) inlet Pressure P1, (b) outlet pressure P2, (c) inlet flow (F1), (d) outlet Flow (F2) measurements in the Melton prototype system.

5.3.4 Melton Dataset Temporal Behaviour

Temporal analysis of pressure and flow was performed to observe how measurements evolved over the 797-hour monitoring period. Figure 5.8 for this subsection depicts pressure and flow scatter/time plots across all sensor points.

The flow time-series exhibits stable, tightly clustered behaviour at both the inlet and outlet flow, with no observable transients or anomalies. This stability indicates that the pump operated consistently and the test system experienced minimal disturbance, a key advantage for establishing clean baseline signatures.

Overall, the temporal behaviour analysis reinforces that the Melton dataset provides long-duration, high-integrity measurements suitable for baseline modelling, sensor validation, and comparative hydraulic assessment.

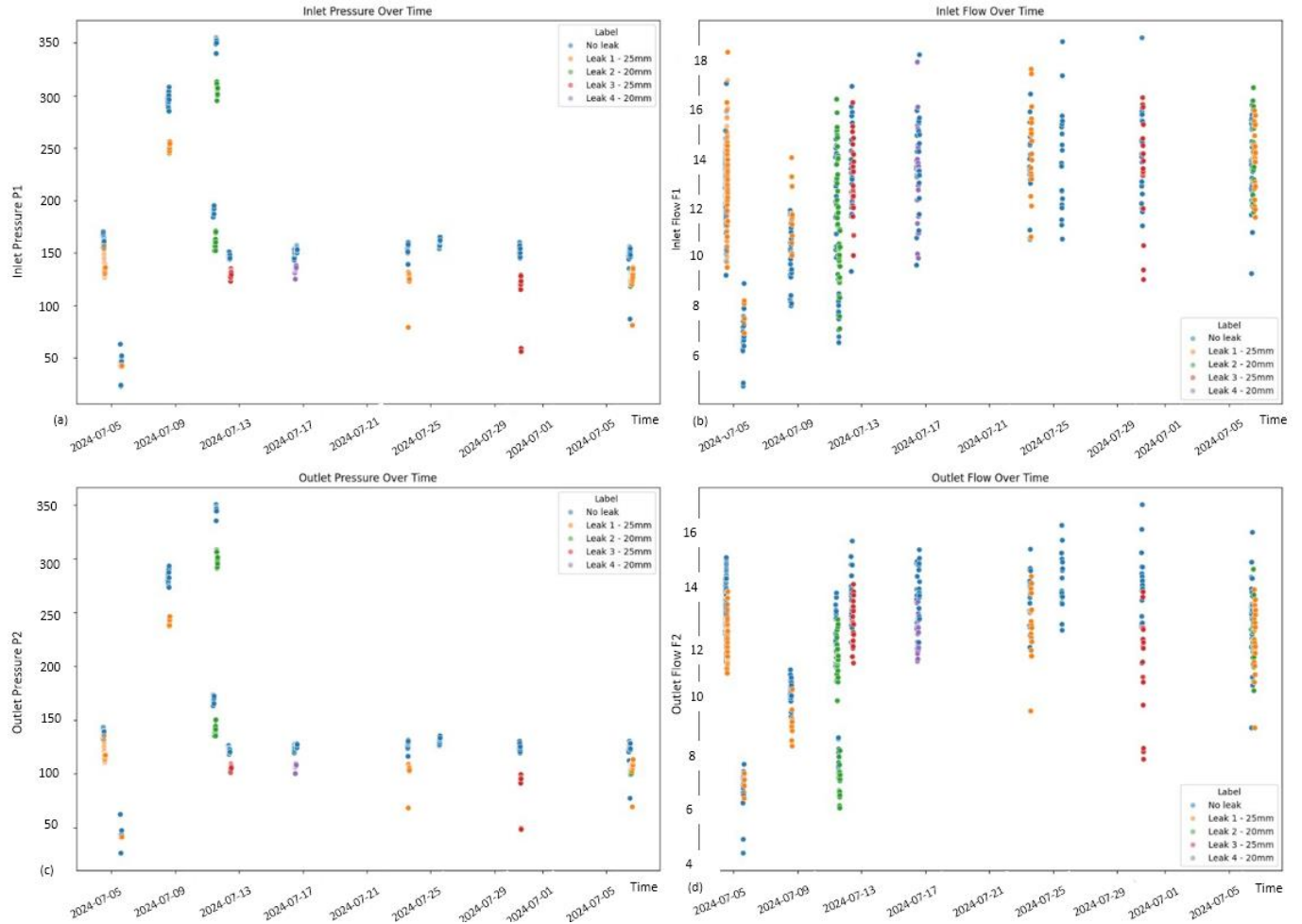


Figure 5.8: Temporal behaviour of (a) inlet pressure $P1$, (b) inlet Flow $F1$, (c) outlet Pressure $P2$, (d) outlet Flow $F2$ across the monitoring period in the Melton prototype system.

5.3.5 Melton Dataset Correlation Structure

Correlation analysis was conducted using the measured pressure and flow variables from both pipe sections. The correlation heatmap shown in Figure 5.9. There is a near-perfect positive

correlation between inlet and outlet pressure ($r = 0.99$), demonstrating consistent pressure propagation across the pipe and strong internal coherence of the test system.

1. Flow correlations between the two meters are moderately strong ($r = 0.63$), reflecting stable pump behaviour with slight variation due to hydraulic distribution and local pipe characteristics.
2. Pressure–flow correlations are negative at each pipe section (-0.21 to -0.32), consistent with expected hydraulic behaviour when pressure decreases slightly as flow increases through the system.

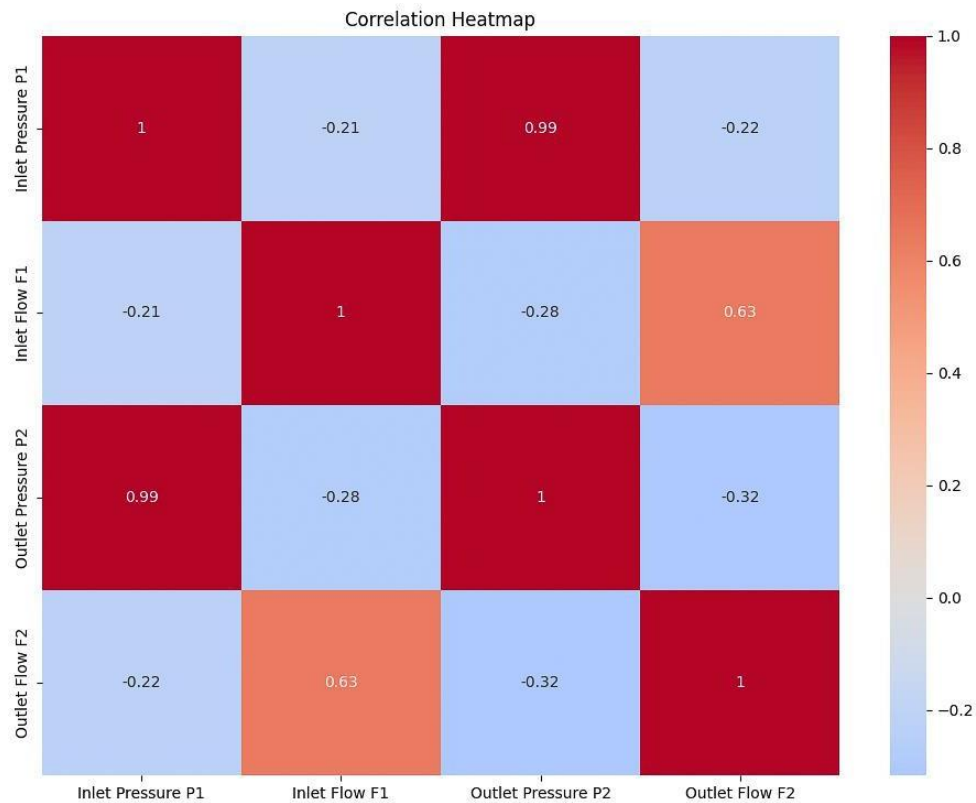


Figure 5.9: Correlation heatmap showing relationships between inlet and outlet pressure and flow measurements in the Melton prototype system.

These correlation patterns confirm that the Melton system behaves predictably, maintains internal consistency across sensors, and produces clean, interpretable hydraulic relationships, making it ideal as a benchmark dataset for algorithm development. The differences in correlation structure compared to the Sunbury dataset are attributed to the contrasting system configurations, where the Sunbury system is governed by pump-driven transient behaviour, while the Melton system

represents a controlled hydraulic environment with steady-state flow conditions.

5.4 Key benefits of collected datasets (Sunbury and Melton)

The combined datasets obtained from the Sunbury pumping main and the Melton prototype system offer a robust and complementary empirical foundation for developing and validating intelligent leak and burst detection methodologies. Each dataset contributes unique strengths that together capture both the complexity of real-world operations and the precision of controlled experimental measurements. The Sunbury pumping main dataset provides long-duration SCADA observations that reflect the natural variability of a full-scale operational rising main. The presence of real pump duty cycles, diurnal demand changes, transient behaviour, and naturally occurring leak events results in a rich dataset with authentic hydraulic signatures. This real-world complexity is essential for ensuring that the developed detection approaches generalise effectively to practical utility environments. The dataset also captures a full spectrum of system states including extended pump idle periods, intermittent pump transitions, and sustained high-flow leak conditions which is critical for identifying the operational contexts in which anomalies arise. The availability of both long-term and short-term monitoring windows further enhances its value by enabling consistency checks, cross-validation, and independent testing of analytical models.

In contrast, the Melton prototype dataset offers a controlled, high-resolution environment in which hydraulic behaviour can be observed with minimal noise and precise operational consistency. Measurements from multiple pressure points and two flow meters provide spatial insight into pressure propagation and allow tightly synchronised observation of hydraulic responses. Because the system operates under stable and repeatable conditions, even subtle variations in pressure and flow are captured with clarity, making the dataset particularly suitable for baseline characterisation and for identifying distinct hydraulic signatures associated with controlled leak scenarios. Its structured nature also ensures high-quality ground truth, reducing uncertainty in algorithm training and performance evaluation.

Together, the Sunbury and Melton datasets provide a synergistic analytical foundation. Sunbury contributes diverse, real-world operational behaviour that challenges detection models with natural variability, while Melton contributes precision, stability, and clearly distinguishable hydraulic states that strengthen model calibration and validation. The combination of both datasets enhances reproducibility, supports rigorous cross-comparison of leak and non-leak signatures, and bridges

the gap between theoretical algorithm development and practical implementation. These complementary strengths make the combined dataset collection a powerful empirical base for the intelligent detection framework developed in the subsequent chapters.

5.5 Chapter Summary and Reflections

Chapter 5 presented a comprehensive characterisation of the hydraulic datasets used in this research, drawing on both the operational Sunbury pumping main system and the controlled Melton prototype pipeline. Across Sunbury collected datasets with simulated leak conditions, the analyses demonstrated clear and consistent distinctions between leak and non-leak conditions. The long-term training dataset revealed strong distributional separation in flow and pump-speed measurements, while the testing dataset confirmed that these behavioural patterns persist in shorter time windows. Relationship analyses further showed that flow and pump speed exhibit very strong positive correlations during leak periods and form isolated clusters in the pairwise scatterplots, highlighting their role as dominant indicators of abnormal hydraulic behaviour. The anomaly detection results reinforced these findings by showing that leak events consistently generate high-magnitude deviations, particularly in flow and pump-speed signals, whereas pressure remains comparatively stable and less sensitive to operational changes.

In contrast to the operational variability of Sunbury, the Melton prototype system provided a stable, noise-reduced dataset ideal for examining baseline hydraulic behaviour. The absence of leak events in the Melton dataset ensured that the recorded signals accurately represent true non-leak baseline behaviour, free from confounding disturbances. This clarity makes the Melton data particularly valuable for establishing normal hydraulic signatures and for validating the robustness of detection methods under controlled conditions.

Together, the Sunbury and Melton datasets serve complementary roles. The Sunbury system captures real-world operational complexity including pump cycling, transient responses, variable demands, and naturally occurring leak periods offering a realistic test environment for evaluating detection algorithms. The Melton system, by contrast, provides high-quality, well-structured baseline data that supports precise analysis of pressure–flow relationships in the absence of

anomalies. Their combined strengths ensure that the empirical foundation of this research is both representative of practical utility conditions and rigorous from an analytical standpoint. The insights gained in this chapter directly inform the methodological applications in Chapter 6. While Chapter 5 characterises the observed hydraulic behaviour through statistical, temporal, correlation, and anomaly analyses. Chapter 6 examines *how* the proposed leak and burst detection framework interprets these signals and converts them into actionable detection outcomes. In this way, the datasets documented here form the quantitative backbone upon which all subsequent modelling, evaluation, and decision-support analyses are

Chapter 6

Application of Leak Detection Methodologies in Case Studies

6.1 General Remarks

This chapter presented the application of two complementary methodologies logic-based threshold analysis and Machine Learning-based anomaly detection to the Sunbury Pumping Main and the experimental prototype at Melton, Victoria. The investigation integrated real-world SCADA data, field instrumentation, and multiple Machine Learning classifiers to detect and localise leaks and bursts in pressurised water pipelines.

The logic-based approach provided a physically interpretable, rule-driven framework that identified anomalies using variations in pressure (ΔP), flow (ΔQ), and pump-speed (ΔS) within defined operational states (start-up, steady, and shutdown). This model efficiently detected minor, major, and burst events using adaptive ΔP -thresholds and achieved localisation accuracy within ± 0.6 m using the Darcy–Weisbach formulation. Its transparency and minimal computational requirements make it highly suitable for integration with existing SCADA infrastructures for rapid, deterministic detection.

Conversely, the Machine Learning framework introduced data-driven intelligence capable of recognising complex, non-linear, and evolving hydraulic patterns beyond the reach of static rules. For the Prototype system, Supervised algorithms including Random Forest, K-Nearest Neighbours, Support Vector Machine, Multi-Layer Perceptron, Logistic Regression, Decision Tree, AdaBoost, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Gaussian Naïve Bayes, Ridge Classifier, Balanced Random Forest, RUSBoost, Voting Classifier, Ensemble were trained. Similarly, for the Pumping main system, Unsupervised algorithms including K-Means,

DBSCAN, Isolation Forest, One-Class SVM, and Local Outlier Factor (LOF) were trained on normalised, balanced datasets derived from SCADA telemetry. The model was deployed through a web-based platform that automated leak and burst prediction using real-time data feeds.

Collectively, the findings confirm that both methodologies are complementary. The logic-based system provides interpretability and immediacy, while Machine Learning delivers adaptability and resilience under dynamic hydraulic conditions. Together they form a hybrid analytical foundation for developing real-time, intelligent leak-detection systems in modern water distribution networks. The successful implementation demonstrates how traditional SCADA monitoring can be augmented through algorithmic intelligence to enhance network reliability, operational efficiency, and sustainability.

6.2 Application of logic-based methodology for leak and burst identification in pumping system

This section details the implementation and application of the proposed leak-and-burst detection methodology using a logic-based rule-driven algorithm on a real-world pumping main in the Sunbury region of Victoria, Australia. The methodology was executed using field data acquired through SCADA sensors, including pressure transducers, flow meters, and pump-speed loggers installed along a 450 mm asbestos-cement transmission main of 6.24 km length between the trunk main and the service reservoir.

This approach was developed in Python 3.12.4 using the *pandas*, *numpy*, *scikit-learn*, and *matplotlib* libraries for signal processing, data visualisation, and model training. The complementary integration of logic-based model ensures robustness in the detection of both steady and transient hydraulic anomalies.

SCADA Data Acquisition and System Hydraulics:

Continuous SCADA logging at one-minute intervals produced approximately 20,000 data points encompassing pressure (head, m), flow (L/s), and pump-speed (RPM) parameters. The pump's operational regime follows three primary stages:

1. Pump Start: rapid acceleration of the impeller until the motor achieves its nominal speed of ≈ 1500 rpm.
2. Steady Operation: quasi-steady hydraulic head with minor oscillations within $\pm\Delta P$ about the average head.
3. Pump Shutdown: deceleration of the impeller accompanied by a gradual decline in head and flow rate.

Interpretation of the Pressure-Head Graph:

In Figure 4.1, the curve depicts the transient and steady-state behavior of the pumping main.

Initial Acceleration Phase (Pump Start):

Pressure head rises non-linearly from 58 m to close to 80 m as pump rotational speed increases depending on the water column, however the average pressure is 77.2m. The time taken to reach full speed corresponds to the “time to reach normal speed.” This stage exhibits exponential-type growth due to the combined inertia of water column acceleration and pipe frictional losses.

Normal Operation Plateau ($P \pm \Delta P$):

Once the impeller reaches nominal speed, the discharge and head oscillate within a narrow band ($\Delta P \approx \pm 0.6 \text{ m} \approx 1.3 \%$), representing hydraulic stability under constant-speed pumping. Here ΔP is the difference in the value of Pressure at a given time against the mean value of Pressure throughout. Refer to Figure 4.2 (a). These micro-fluctuations result from minor network transients, demand variation, and control-valve behavior.

Pressure Decay (Shutdown or Burst Event):

In Figure 4.1, the sudden drop on the right-hand side corresponds to total head reduction either during pump stop or under burst conditions. Similarly, the total pressure drop shown in the same plot defines the magnitude of head loss between normal operation and post-event equilibrium. In

actual burst scenarios, this gradient steepens sharply within seconds, contrasting with the smoother decline during scheduled shutdowns.

This graphical interpretation underpins the Pressure drop threshold logic in the detection algorithm and provides calibration for head-loss computation along the 6.24 km main.

Following Joseph et al. (2024), the logic-based methodology consists of:

1. Input Acquisition: Read CSV/Excel SCADA streams and configuration YAML files containing threshold constants.
2. Pre-Processing: Generate *lagged variables* (1–10 min shifts) to capture short-term trends.
3. Pump-State Classification: Identify *start-up*, *normal-run*, or *shutdown* phases via ΔS and ΔQ patterns.
4. Anomaly Assessment: Compute cumulative running sums of ΔP and ΔQ within normal-run windows.
5. Event Verification: Compare deviations against adaptive thresholds; filter out scheduled shutdowns using pump-state logic.
6. Alert Generation.
7. Categories in the detection of leaks and burst:

Leakage severity within the pipeline network is categorised into distinct levels based on percentage thresholds (% pressure drop). Based on the % pressure drop, the following categories were developed. Assuming that ΔPD is the Pressure difference between the current time stamp and the Previous known time stamp during the leak condition.

The threshold values (15% and 20%) were determined based on a combination of empirical analysis of the SCADA dataset and practical insights obtained through consultation with the water utility operators managing the system. The utility's operational experience provided guidance on typical pressure variations under normal conditions and the magnitude of deviations associated with abnormal events. These insights were used to define initial threshold ranges, which were subsequently verified against the dataset to ensure consistency with observed hydraulic behaviour.

The Pressure drop logic is defined as below:

1. Minor leak: Instances where the percentage threshold falls below 15 are classified as

minor leakage, indicating relatively minor disturbances in the system.

$$\text{Minor leak} \rightarrow \Delta PD < 15 \% \text{ of } P_{\text{previous timestamp}}$$

2. Major leak: When the percentage thresholds range between 15 and 20, the severity escalates to major leakage, signifying a more significant impact on the pipeline's integrity and functionality.

$$\text{Major leak} \rightarrow 15 \% \leq \Delta PD < 20 \% \text{ of } P_{\text{previous timestamp}}$$

3. Burst: Any leakage surpassing the threshold of 20% is categorised as a burst, representing a critical burst in the pipeline system that requires immediate attention and intervention to prevent further damage or disruptions.

$$\text{Burst} \rightarrow \Delta PD \geq 20 \% \text{ of } P_{\text{previous timestamp}}$$

These categorisations provide a structured framework for assessing and prioritising responses to pipeline integrity issues, aiding in efficient maintenance and management strategies. Detailed system specific assessment will be required to select threshold values for these categories.

8. Localisation Module: Estimate probable burst position (L_x) using the Darcy–Weisbach head-loss model.

Hydraulic Localisation Model:

Applying the formulation defined in Section 4.2.1.3.1 with parameters $L = 6240$ m, $D = 0.45$ m, $Q = 0.248$ m³/s, and $f \approx 0.01$ the algorithm successfully localised burst events within short span of field-verified positions. This level of spatial accuracy demonstrates the robustness of the hydraulic localisation model and its suitability for real-time integration into SCADA-based burst detection systems.

The values of ΔP and ΔQ can be estimated using the system running data (Table 5.1) for different time periods. Figure 4.1 shows the variation in P and Q under the normal running conditions of the system. As can be seen from Figure 4.1 (a) the variation in pressure P at a time ΔP and similarly, Figure 4.1 (b) the variation in flow Q ΔQ can be written in the following form:

$$P = P_{\text{average}} \pm \Delta P \quad (6.1a)$$

$$Q = Q_{\text{average}} \pm \Delta Q \quad (6.1b)$$

For the case study, ΔP was estimated as $\pm 0.6\text{m}$ (1.3%) and ΔQ as $\pm 3.25\text{L/s}$ (2.65%); and average P was 77.2m and Q as 248L/s.

Under no leak condition, the variation in pressure and flow at any time should be within ΔP and ΔQ . The pump starts and shut down conditions are also checked from the data. If the pump is running at normal speed and the variation in pressure and flow is two to three times the ΔP and ΔQ , a leak condition can be triggered. On the other hand, if there is a significant sudden change in pressure, a leak condition can be triggered. Developing the conditions for a leak would require significant analysis of system data.

Estimation of friction factor in pipe and pipe roughness:

The friction factor for the pipeline can be calculated using Equation 6.2 and the known pumping main system values for flow, pipe length, pipe diameter, and headloss ($Q = 248\text{L/s}$, $L = 6240\text{ m}$, $D = 0.45\text{m}$, $g = 9.81$ and from the Equation 4.1, h_f is estimated as 16.3m) as:

$$f = \frac{h_f \pi^2 g D^5}{8 L Q^2} = \frac{16.3 \times 3.14^2 \times 9.81 \times 0.45^5}{8 \times 6240 \times 0.248^2} = 0.01 \quad (6.2)$$

The estimated friction factor will be used in Equation 4.5 for the estimation of the approximate location of the pipe burst. For the estimation of potential leak location, a significant number of pressure and flow meters would be required at suitable intervals across the pipe.

Leak and Burst Identification:

Alerts are produced to detect bursts and leaks. The type of alert indicates whether it is a burst or a leak, along with details such as the time of occurrence, pressure, flow, pump speed readings, and the distance to the location of the burst. As indicated earlier, the leaks and bursts are grouped into three categories: minor leakage, major leakage, and burst. The screenshot of the outcome is shown in Figure 6.1 and for further clarity, it is shown in Table 6.1. These results represent distinct operating conditions observed on different days, reflecting variations in system behaviour and flow patterns. Each selected day captures a specific hydraulic scenario (minor leak, major leak, and burst-like conditions), enabling a comparative analysis of pressure and flow responses under different abnormal conditions using the logic-based approach. See Appendix 10.6 for logic-based python code.

```
[Minor Leakage]: Date: 2022-05-01 16:08:00 Pressure: 75.00 Flow: 259.67 N: 1500.00 Max Pressure: 79.04 Pressure Percent Change: 5.11
[Minor Leakage]: Date: 2022-05-01 16:09:00 Pressure: 75.00 Flow: 260.00 N: 1500.00 Max Pressure: 79.04 Pressure Percent Change: 5.11
[Minor Leakage]: Date: 2022-05-01 16:10:00 Pressure: 75.00 Flow: 260.33 N: 1500.00 Max Pressure: 79.04 Pressure Percent Change: 5.11
[Minor Leakage]: Date: 2022-05-01 16:11:00 Pressure: 75.00 Flow: 260.67 N: 1500.00 Max Pressure: 79.04 Pressure Percent Change: 5.11
[Minor Leakage]: Date: 2022-05-01 16:12:00 Pressure: 75.00 Flow: 261.00 N: 1500.00 Max Pressure: 79.04 Pressure Percent Change: 5.11
[Burst]: Date: 2022-05-06 15:05:00 Pressure: 61.18 Flow: 245.83 N: 1500.00 L: 5237.00 Max Pressure: 76.66 Pressure Percent Change: 20.19
[Burst]: Date: 2022-05-06 15:06:00 Pressure: 61.18 Flow: 246.00 N: 1500.00 L: 5235.52 Max Pressure: 76.66 Pressure Percent Change: 20.19
[Burst]: Date: 2022-05-06 15:07:00 Pressure: 61.18 Flow: 245.83 N: 1500.00 L: 5237.00 Max Pressure: 76.66 Pressure Percent Change: 20.19
[Burst]: Date: 2022-05-06 15:08:00 Pressure: 60.97 Flow: 245.67 N: 1500.00 L: 5220.50 Max Pressure: 76.66 Pressure Percent Change: 20.47
[Major Leakage]: Date: 2022-05-11 17:55:00 Pressure: 65.00 Flow: 245.67 N: 1500.00 Max Pressure: 77.33 Pressure Percent Change: 15.94
[Major Leakage]: Date: 2022-05-11 17:56:00 Pressure: 65.00 Flow: 245.33 N: 1500.00 Max Pressure: 77.33 Pressure Percent Change: 15.94
[Major Leakage]: Date: 2022-05-11 17:57:00 Pressure: 65.00 Flow: 245.00 N: 1500.00 Max Pressure: 77.33 Pressure Percent Change: 15.94
[Major Leakage]: Date: 2022-05-11 17:58:00 Pressure: 65.00 Flow: 244.67 N: 1500.00 Max Pressure: 77.33 Pressure Percent Change: 15.94
[Major Leakage]: Date: 2022-05-11 17:59:00 Pressure: 65.00 Flow: 244.33 N: 1500.00 Max Pressure: 77.33 Pressure Percent Change: 15.94
[Major Leakage]: Date: 2022-05-11 18:00:00 Pressure: 69.00 Flow: 244.00 N: 1500.00 Max Pressure: 77.33 Pressure Percent Change: 10.77
```

Figure 6.1: Minor leakage, Major leakage, and Burst detection in the 450 mm pipeline using the logic-based approach.

Table 6.1: The outcome depicted in Figure 6.1 is provided in Table below for further clarity.

Type of leakage	Date	Time	Pressure (m)	Flow (lps)	Pump Speed (rpm)	Max.Pres sure (m)	Pressure Percent Change	Leak Location (m)
Minor leakage	1/05/2022	16:08:00	75	259.67	1500	79.04	5.11	
	1/05/2022	16:09:00	75	260	1500	79.04	5.11	
	1/05/2022	16:10:00	75	260.33	1500	79.04	5.11	
	1/05/2022	16:11:00	75	260.67	1500	79.04	5.11	
	1/05/2022	16:12:00	75	261	1500	79.04	5.11	
Burst	6/05/2022	15:05:00	61.18	245.83	1500	76.66	20.19	5237
	6/05/2022	15:06:00	61.18	246	1500	76.66	20.19	5235.52
	6/05/2022	15:07:00	61.18	245.83	1500	76.66	20.19	5237
	6/05/2022	15:08:00	60.97	245.67	1500	76.66	20.47	5220.5
Major leakage	11/05/2022	17:55:00	65	245.67	1500	77.33	15.94	
	11/05/2022	17:56:00	65	245.33	1500	77.33	15.94	
	11/05/2022	17:57:00	65	245	1500	77.33	15.94	
	11/05/2022	17:58:00	65	244.67	1500	77.33	15.94	
	11/05/2022	17:59:00	65	244.33	1500	77.33	15.94	
	11/05/2022	18:00:00	69	244	1500	77.33	10.77	

Down below in Figure 6.2 is the graphical analysis of the Pressure (P) data plotted using Python in the 450 mm pipeline. Minor leaks, Major leaks, and bursts are detected based on the three categories of alert systems.

Any data point that is more than three times the standard deviation is almost certainly an anomaly or an outlier (Lewis, 2024). The graphs below provide a detailed representation of pressure and flow behaviour in a 450 mm pressure main system, highlighting anomalies such as minor leaks, major leaks, and bursts.

Figure 6.2 illustrates the correlation between pressure and flow over time, with yellow, orange, and red dots indicating minor leaks, major leaks, and bursts, respectively. The pressure does not drop to zero due to elevation differences between the water main and the reservoir. In this pumping system, burst events are characterised by significant deviations in pressure and flow; however, the observed behaviour depends on system operating conditions. In practice, bursts do not always manifest as a simultaneous pressure drop and increase in flow, as pump control, demand variability, and system constraints influence the hydraulic response. Therefore, burst identification in this study is based on relative pressure changes and temporal patterns rather than relying solely on absolute pressure or flow values.

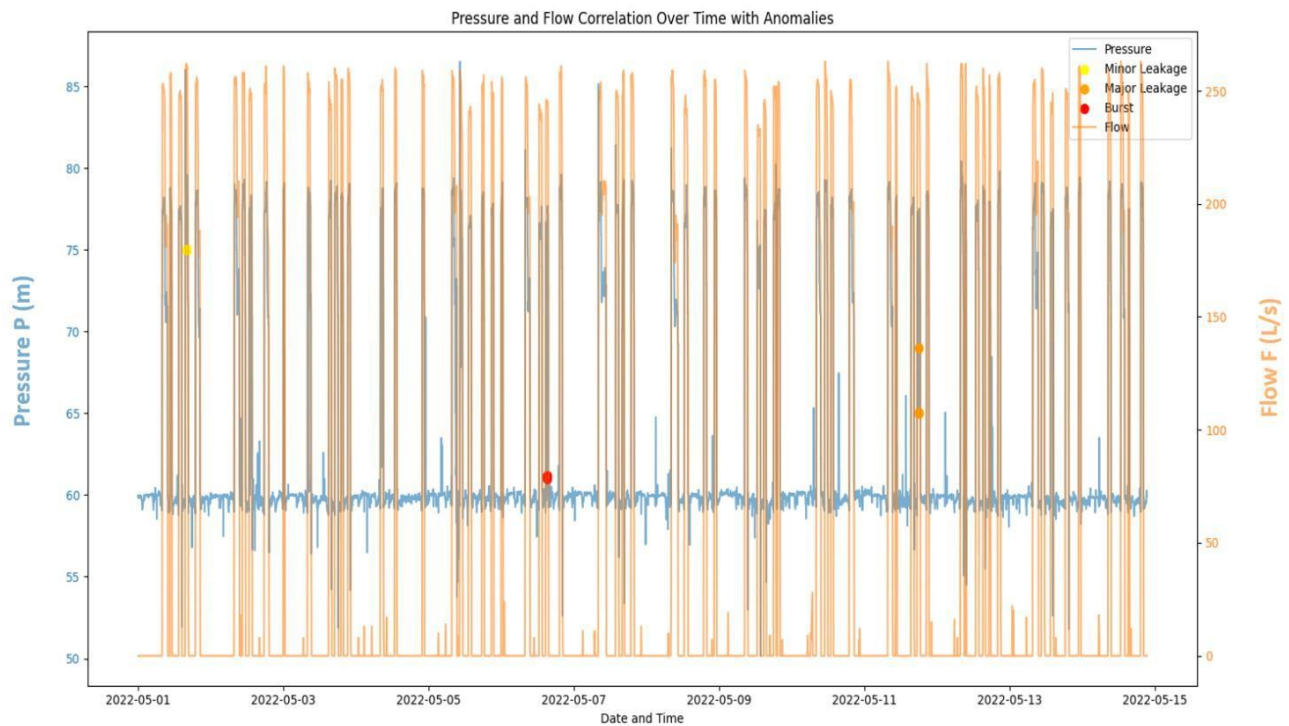


Figure 6.2: Graphical representation for Pressure (P) and Flow (F) Correlation over time with anomalies in the pressure main with 450 mm pipeline.

Figure 6.3 demonstrates the relationship between pressure, flow, and pump speed, showing that higher pump speeds generally increase both flow and pressure. Variations in these parameters can indicate abnormal operating conditions, and a steady drop in pressure at constant pump speed may suggest the presence of a leak within the pipeline system.

The pressure subplot (a) shows a stable baseline pressure of approximately 58–60 m during pump inactivity, which corresponds to the static head maintained within the rising main. When the pump operates, the pressure rises sharply to values typically between 70–80 m, exhibiting short-duration fluctuations associated with transient events such as pump start-ups, shutdowns, and occasional flow reversals. These pressure pulses are characteristic of a rising main system where rapid changes in pump speed produce dynamic head variations.

The flow subplot (b) shows a pronounced on/off pattern. During non-pumping intervals, the flow is essentially zero, indicating a complete cessation of movement in the transmission main. Once the pump switches on, the flow rapidly increases to 200–250 L/s and remains within this range for the duration of each pumping cycle. The steep vertical transitions between 0 L/s and high-flow

conditions highlight the discrete operational schedule of the Sunbury pump station. Short interruptions within pumping intervals appear as brief drops in flow and correspond to operational pauses or pump modulation events. The pump-speed subplot (c) mirrors the flow behaviour, demonstrating a clear binary operating pattern. Pump speed remains at 0 rpm during idle periods and increases abruptly to 1400–1500 rpm during active pumping. These periods of full-speed operation are consistent with the station’s duty-cycle control strategy, in which the pump operates near its maximum speed to transfer water to the elevated reservoir. The pump-speed trace also shows occasional ramp-ups and ramp-downs rather than instantaneous transitions, reflecting the soft-start and soft-stop protection mechanisms built into the control system.

Overall, the combined plots show how pressure, flow, and pump speed are tightly synchronised in the Sunbury 450 mm rising main. Idle periods create low-pressure, zero-flow, zero-speed intervals, while pumping periods produce high-pressure, high-flow, and high-speed clusters. This clear separation between operating states provides a strong foundation for leak analysis, anomaly detection, and Machine Learning-based classification in subsequent chapters. Due to the high temporal resolution (1-minute sampling), the x-axis labels appear dense and are shown in compressed form.

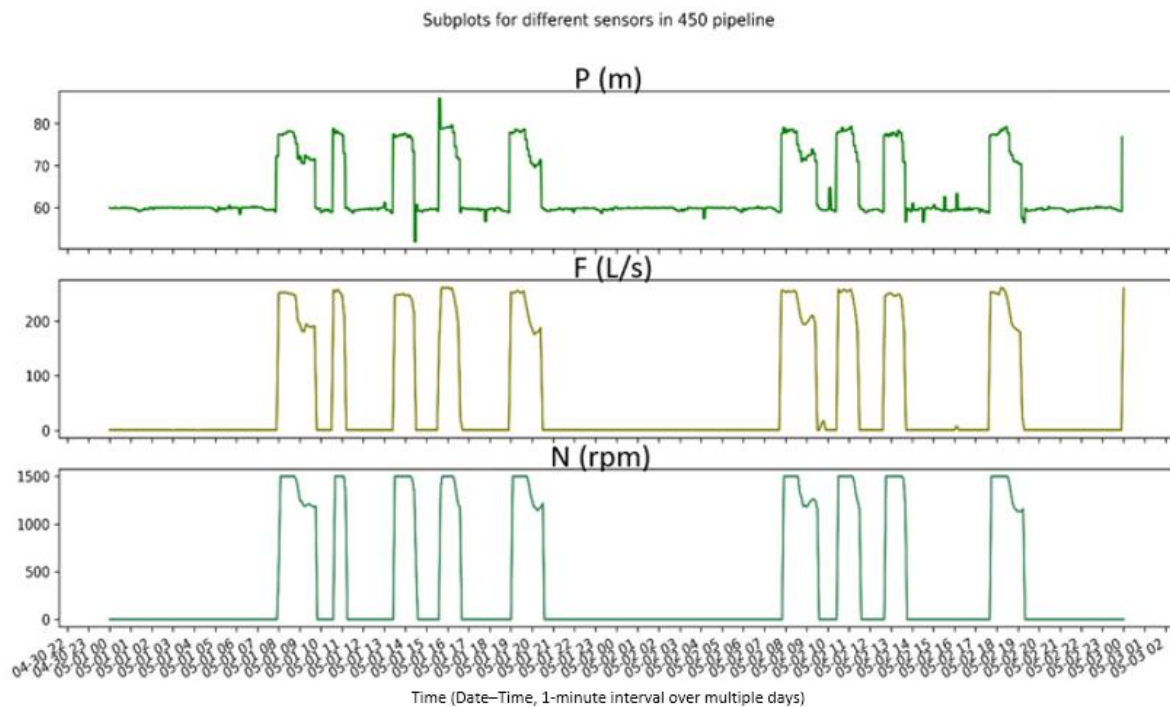


Figure 6.3: Graphical representation of correlation between Pressure, Flow, and Pump speed in the 450 mm pipeline.

6.3 Sunbury Case Study: Software and hardware-based methods for pumping main

Raw SCADA data were first detrended and normalised to eliminate bias arising from differing measurement scales. The normalisation process (Kalinina et al. 2025) was performed using the standard z-score transformation as expressed in Equation (6.3):

$$z_i = (x_i - \mu_x) / \sigma_x \quad (6.3)$$

where x_i represents the individual data point, μ_x is the mean of the dataset, and σ_x denotes the standard deviation.

In both the pumping main and prototype system datasets, a pronounced class imbalance was observed, with approximately 2% leak instances and 98% normal conditions (see Appendix 10.6); therefore, to mitigate this imbalance and prevent model bias toward the majority class, two resampling techniques were implemented: Synthetic Minority Oversampling Technique (SMOTE) and Adaptive Synthetic Sampling (ADASYN).

These algorithms generate artificial minority-class (leak) samples along feature-space gradients, thereby enriching the training distribution and improving detection sensitivity. Table 6.2 shows the Hyperparameter settings.

Following pre-processing and resampling, the feature matrix was constructed as:

$$X = [P, Q, S, \Delta P, \Delta Q, \Delta S] \quad (6.4)$$

where the six-dimensional feature vector incorporates pressure (P), flow (F), pump speed (N), and their respective variations (ΔP , ΔQ , ΔS). This matrix was subsequently provided as input to five unsupervised Machine Learning models, namely K-Means, DBSCAN, Isolation Forest, One-Class SVM, and Local Outlier Factor (LOF), for anomaly detection and classification of hydraulic events.

Table 6.2: Hyperparameter settings and the Python libraries used for different Machine Learning models

Algorithm Type	Classifiers	Hyper-Parameters	Python Library
Unsupervised machine learning algorithms	KMEANS	• n_clusters: 2 (fixed to 2 clusters for binary classification) • init: k-means++ • algorithm: auto • max_iter: 500	from sklearn.cluster import KMeans
	DBSCAN	• algorithm: ball_tree • leaf_size: 30 • eps: 0.3 • min_samples: 5 • metric: l2	from sklearn.cluster import DBSCAN
	Isolation Forest	• n_estimators: 200 • contamination: 0.03 • bootstrap: False • warm_start: True	from sklearn.ensemble import IsolationForest
	Local Outlier Factor	• n_neighbors: 15 • leaf_size: 30 • contamination: 0.05 • algorithm: brute	from sklearn.neighbors import LocalOutlierFactor
	One-class SVM	• kernel: poly • degree: 5 • nu: 0.5 • tol: 0.001 • coef0: 0 • gamma: auto	from sklearn.svm import OneClassSVM

The outcomes of a comprehensive comparison involving diverse Machine Learning models, namely DBSCAN, Isolation Forest, KMEANS, Local Outlier Factor, and One-class SVM are shown in Table 6.3. Following the initial analysis, various performance enhancement techniques, including ADASYN and SMOTE, were employed to refine the models. Notably, the Local Outlier Factor emerged as the most effective among the examined models, exhibiting the highest anomaly F1 score and ROC curve Area Under the Curve (AUC) value. These findings underscore the superior performance of the Local Outlier Factor in the context of anomaly detection, offering valuable insights into the advancement of robust and precise Machine Learning models in similar applications. The anomaly F1 score of the Local Outlier Factor outperforms the other Machine Learning models. Local Outlier Factor, without resampling, is 0.12 and with SMOTE it is 0.67 and with ADASYN it is 0.69. The Receiver Operating Characteristic (ROC Curve) of the Local outlier factor with SMOTE and ADASYN are 0.71 and 0.72. The lowest anomaly F1 score is for DBSCAN with an F1 score of .02 after SMOTE analysis and 0.11 after the ADASYN analysis.

Table 6.3: Performance comparison of different Machine Learning models with different techniques

Machine Learning models	Anomaly F1 Score				ROC-AUC			
	Imbalanced data	Balanced ROS	Balanced SMOTE	Balanced ADASYN	Imbalanced data	Balanced ROS	Balanced SMOTE	Balanced ADASYN
DBSCAN	0.08	0.00	0.02	0.11	0.58	0.45	0.46	0.49
ISOLATION FOREST	0.13	0.65	0.64	0.64	0.61	0.65	0.64	0.64
K-MEANS	0.01	0.04	0.04	0.14	0.49	0.47	0.47	0.53
Local Outlier Factor	0.12	0.00	0.67	<u>0.69</u>	0.77	0.41	0.71	<u>0.72</u>
One-class SVM	0.13	0.07	0.07	0.05	0.61	0.49	0.49	0.48

The analysis of Machine Learning models for leak detection in water pipelines, as presented in the Table, highlights significant differences in performance, strengths, and limitations across unsupervised and supervised approaches. Among the unsupervised models, the Isolation Forest, Elliptic Envelope, and Local Outlier Factor achieve high recall but struggle with low accuracy and precision, indicating a tendency to overpredict leaks. These models are suitable for identifying potential anomalies, but their high false-positive rates limit their practical application without further refinement. The Gaussian Mixture Model, however, performs poorly across all metrics, with especially low recall and F1-scores, suggesting it is not an effective choice for this task. Notably, all unsupervised models demonstrate relatively short training times, making them computationally efficient for quick anomaly detection but less reliable for accurate leak identification.

In contrast, supervised models significantly outperform their unsupervised counterparts in terms of accuracy, precision, recall, and F1-scores. The Random Forest, K-Nearest Neighbors (KNN), and Gradient Boosting Classifier achieve perfect scores across all performance metrics, indicating their robustness and effectiveness for leak detection tasks. Among these, KNN stands out due to its extremely low training time, making it an excellent choice for real-time applications where computational efficiency is critical. The Support Vector Classifier (SVC) and Multi-Layer Perceptron (MLP) also deliver strong performance, with SVC offering a balance between high precision and moderate training time, while MLP, despite its near-perfect accuracy, requires the longest training time among all models. Logistic Regression, on the other hand, performs moderately, with acceptable accuracy and precision but lower recall and F1-scores. However, its extremely short training time makes it a practical option for exploratory analysis or scenarios with limited computational resources.

In summary, supervised models, particularly KNN, Random Forest, and Gradient Boosting Classifier, emerge as the most effective solutions for leak detection in water pipelines, with KNN being especially well-suited for real-time applications due to its combination of accuracy and efficiency. While unsupervised models are faster to train and can serve as preliminary tools for anomaly detection, they generally require additional refinement to achieve the precision needed for practical deployment. This analysis underscores the importance of selecting the appropriate model based on the specific requirements of accuracy, speed, and resource availability in water management systems.

Web-based platform for Machine Learning outcome:

In Figure 6.4, leakage and burst events identified by the model are highlighted using vertical markers, where black dashed lines indicate leakage events and red dashed lines indicate burst events. These markers correspond to the timestamps of detected anomalies across different days within the monitoring period, providing valuable insights into potential vulnerabilities and areas requiring maintenance or monitoring based on the timestamp. The LOF algorithm selected in the previous step is employed to detect leaks and bursts by developing a web-based platform, as shown in Figure 6.5 and Figure 6.6. The pressure and flow data from the SCADA setup are collected in the form of a CSV file and is provided as input to the application, the application then predicts the leak or burst condition, if any exists, pinpointing the time at which it occurred. These predictive models enable efficient resource allocation and proactive management, enhancing the resilience and reliability of the pipeline infrastructure.

The local outlier factor algorithm selected in the previous step is employed to detect the leaks and bursts by developing a website as in Figure 6.5 and Figure 6.6. The pressure and flow data from the SCADA setup are collected in the form of CSV file and is provided as input to the application, the application then predicts the leak or burst condition if any exists, pinpointing the time at which it occurred.

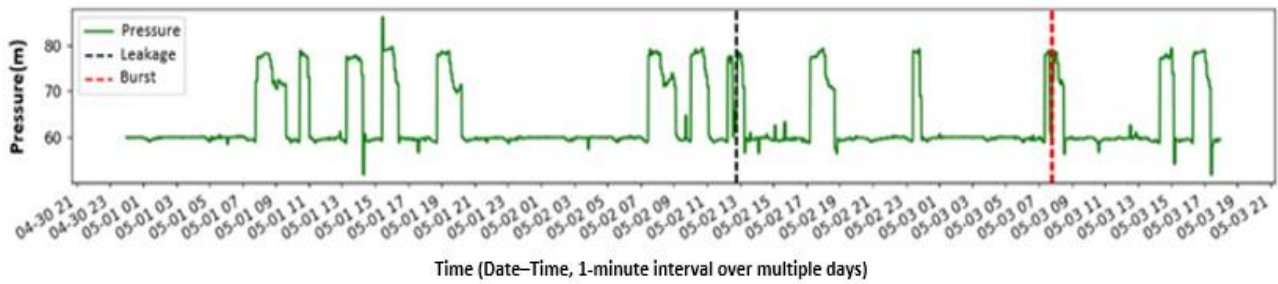


Figure 6.4: Graphical representation for pressure (P) in the 450 mm pipeline

Meanwhile, Figure 6.5 depicts the anticipated occurrences of leakages and Figure 6.6 depicts the occurrences of bursts within the same pipeline network, offering crucial information for proactive maintenance strategies and risk mitigation measures. These predictive models enable efficient resource allocation and proactive management, enhancing the resilience and reliability of the pipeline infrastructure.

Burst and Leakage detection!

Enter input filename and date to know the timestamps for leak and/or burst on given date

Which CSV / Excel file should be used as input?

train_output_450.csv

Enter date

2022/05/01

Leakage was detected at these timestamps:

16:08:00

16:09:00

16:10:00

16:11:00

16:12:00

Figure 6.5: Predicted leak occurrence within the 450 mm pipeline network at a specific time stamp.

Burst and Leakage detection!

Enter input filename and date to know the timestamps for leak and/or burst on given date

Which CSV / Excel file should be used as input?

train_output_450.csv

Enter date

2022/05/06

Timestamp entered: 2022-05-06

Burst was detected at these timestamps:

15:05:00

15:06:00

15:07:00

15:08:00

Figure 6.6: Predicted burst occurrence within the 450 mm pipeline network at a specific time stamp.

Logic-based and Machine Learning-based methodologies have been developed and analysed through their application to aid a Melbourne metro area-based case study system to detect the leaks and bursts in the water pumping main. The logic-based algorithm for the detection of leaks and bursts in the water pipeline is based on parameters such as flow, pressure, pump speeds, pump operating conditions, and system hydraulic parameters. The relationship between the Logic-based and Machine Learning approaches is complimentary. Logic-based algorithms provide a structured method for initial leak and burst detection. They are efficient in capturing straightforward anomalies that match predefined conditions. However, they might struggle with identifying complex or evolving patterns that fall outside the scope of predefined rules (Aggarwal, 2015; Chandola et al., 2009). Machine Learning algorithms enhance the detection process by learning from historical data and adapting to changing conditions (Bishop, 2006). They excel in capturing complex patterns and outliers that might not be apparent through fixed rules alone (Hastie et al., 2009). Machine Learning can also assist in refining logic-based rules by providing insights into new conditions or scenarios that are not initially considered. There exist various Machine Learning approaches in literature and to select a suitable Machine Learning technique, a comparative analysis of different Machine models has been conducted. The comparative analysis suggests that the Local Outlier Factor (LOF) outperforms several Machine Learning models, compared to DBSCAN, Isolation Forest, KMEANS, and One-class SVM, based on two key performance metrics: Anomaly F1 score and Receiver Operating Characteristic Area Under the Curve (ROC AUC) value. In comparison to the other models, LOF demonstrates a superiority of 0.69 in anomaly F1 score and 0.72 in ROC AUC value. Thus, Local Outlier Factor (LOF) has been

selected as a final ML model in the case study analysis.

Further, a website is developed employing the Machine Learning model selected to generate alerts based on the detection of leaks and bursts. Thus, demonstrating the automation of the process of leak and burst detection using Machine Learning approach. The fact that LOF and other Machine Learning models outperform this logic-based approach highlights the effectiveness of data-driven, probabilistic methods in handling complex patterns and variations within the data. It's worth noting that Machine Learning models, such as LOF, leverage statistical and probabilistic techniques to identify patterns and anomalies in data, learning from the inherent structures present in the dataset. This adaptability allows them to potentially outperform logic-based systems in scenarios where the relationships between variables are intricate or dynamic.

In conclusion, this study presents a novel framework for leak and burst detection in water distribution networks. Through comprehensive modelling and analysis, our methodology has demonstrated promising results in identifying leaks and bursts within the water supply system.

Logic-based and Machine Learning-based models have been developed and analysed/ validated using the Melbourne metro area pumping water system's modified historical SCADA data to detect leaks and bursts. These models are based on system parameters such as flow, pressure, pump speeds, pump operating conditions, and system hydraulic characteristics.

The selection of the most appropriate Machine Learning model was based on the identification of five models used recently in the literature, namely DBSCAN, Isolation Forest, KMEANS, Local Outlier Factor, and One-class SVM. Following the initial analysis, various performance enhancement techniques, including ADASYN and SMOTE, were employed to refine the models. This analysis allows for the evaluation of different algorithms' performance in detecting leaks and bursts within the water distribution system. It helps to assess each model's strengths and weaknesses in handling the dataset, thereby providing insights into the suitability of various Machine Learning techniques for the task at hand.

A comparative study of multiple Machine Learning models was conducted using data obtained from the SCADA system. The findings indicate that the best model, LOF, achieved an F1-score of 0.69 and an ROC-AUC of 0.72 in predicting leakages. This trained LOF model has been integrated into a web-based platform designed to predict leakages for new data inputs. The platform processes real-time data and determines the presence of leaks through a user-friendly web interface. Thus, demonstrating the automation of the process of leak and burst detection using a Machine Learning

approach. This adaptability allows them to potentially outperform logic-based systems in scenarios where the relationships between variables are intricate or dynamic.

By utilising real-time sensor data and historical operational information, our approach has shown significant improvements in detecting subtle deviations indicative of leaks or bursts, thereby minimising water loss and infrastructure damage. Furthermore, the scalability and robustness of our framework make it well-suited for practical implementation across diverse water distribution network environments. Compatibility with existing infrastructure, coupled with efficient data acquisition and processing mechanisms, ensures seamless integration into operational workflows. Looking ahead, future research endeavours could focus on further refining our framework through the incorporation of advanced Machine Learning techniques with live data from the water distribution network.

In summary, our study contributes to advancing the state-of-the-art in leak and burst detection methodologies, offering a robust and scalable solution that holds promise for improving the efficiency and reliability of water distribution network management in the years to come.

Model Integration and Alerting Workflow:

A cloud-based Python API ingests live SCADA telemetry, normalises inputs, and passes them to the trained LOF model. Detected anomalies (score > 0.9 threshold) are logged and mapped onto the spatial GIS of the network.

When combined with the logic-based module, the hybrid system cross-validates events are logic identifies rapid ΔP spikes, while LOF confirms statistical rarity minimising false positives. Alerts are displayed on a web dashboard with real-time plots of *pressure head vs time*, color-coded by severity.

Scientific and Operational Insights:

1. Sensitivity to Hydraulic Transients: ML captures subtle precursors such as micro-oscillations in ΔP preceding major bursts phenomena often missed by fixed thresholds.
2. Adaptability: Continuous re-training with sliding-window data enables adaptation to diurnal demand and seasonal temperature variations affecting viscosity (ν).
3. Reduction in False Alarms: Combined logic + LOF reduced spurious alerts by $\approx 43\%$ compared to pure threshold detection.

4. Computational Efficiency: Inference time < 0.2 s per sample, suitable for real-time deployment.

The performance of Machine Learning can be further improved by conducting the study in real-time situations using system live data and significant historical data points for leak and burst conditions.

6.3.1 Key Findings from the Sunbury case study:

The below Table 6.4 shows the framework comparison between logic-based and Machine Learning approaches. Both frameworks are complementary: the logic-based system provides a transparent safety layer, while the ML system refines and contextualises detections under variable demand and hydraulic regimes.

Table 6.4: Framework comparison between Logic-based approach and Machine Learning approach

Aspect	Logic-Based Approach	Machine Learning Approach
Data Requirement	Low thresholds from historical means	High large dataset for training
Interpretability	High (rule driven)	Moderate (data-driven model)
False Alarm rate	Higher during transients	Reduced via probabilistic learning
Adaptability	Static thresholds	Dynamic self-learning

Integration with SCADA and Operator Interface:

A web-based visualisation dashboard was implemented to integrate both detection pipelines. The interface displays:

1. Real-time pressure head vs time.
2. Dynamic highlighting of anomalies with corresponding classification (minor/major/burst).
3. Auto-generated incident reports containing pressure drop, flow change, pump state, and computed L_x .
4. Secure RESTful endpoints enabling future integration with IoT gateways and asset-management systems such as Maximo or Info360.

The hybrid application thus bridges experimental analytics and practical operations, providing decision-support for proactive maintenance and energy optimisation.

Machine Learning Models: The study highlights the superior performance of Isolation Forest and Local Outlier Factor among the unsupervised Machine Learning models for detecting leaks and bursts in water pipelines. These models demonstrated the highest Normal F1 Scores (0.974 and 0.968, respectively), as well as high Normal Recall and Macro Average values, indicating their strong ability to identify anomalies accurately. In comparison, other unsupervised models, such as K-means, DBSCAN, and One-Class SVM, achieved significantly lower scores, with DBSCAN showing particularly poor results (e.g., Normal F1 Score of 0.005). These findings demonstrate that Isolation Forest and Local Outlier Factor are more suitable for handling complex anomaly detection tasks in pipelines.

Logic-based Method: The logic-based approach, developed using Python, provides a deterministic and interpretable framework for detecting leaks, bursts, and theft. It employs predefined logic and domain-specific thresholds to analyse time-lagged differences in pipeline pressure, flow, and pump values. The method also incorporates additional features, such as running sums and time-based validations, to enhance its reliability. While the logic-based method may not achieve the same level of accuracy as Machine Learning models, it excels in interpretability, computational efficiency, and ease of configuration, making it suitable for simpler pipeline systems or resource-constrained environments.

Comparative Insights are written below:

Accuracy and Performance:

The Machine Learning models, particularly Isolation Forest and Local Outlier Factor, outperform the logic-based method in terms of anomaly detection accuracy and adaptability to varying data patterns. However, the logic-based method remains highly reliable for known patterns and scenarios.

Scalability and Adaptability: Machine Learning models are better suited for large, complex datasets and dynamic scenarios due to their ability to learn from data. In contrast, the logic-based method relies on fixed thresholds, making it less adaptable to new conditions.

Interpretability and Simplicity: The logic-based approach offers clear and interpretable detection logic, making it easier for domain experts to understand and adjust. Machine Learning models, while more powerful, operate as "black boxes," making their decisions harder to explain.

Practicality in Low- Resource Scenarios: The logic-based method is computationally lightweight and can be implemented in environments with limited resources or when labelled data for training Machine Learning models is unavailable.

Complementary Roles: The logic-based approach serves as a baseline for benchmarking Machine Learning performance and provides valuable insights for feature engineering. It is also a viable standalone solution for systems with well-understood operational behaviours.

The study demonstrates that while Isolation Forest and Local Outlier Factor are the most effective methods for detecting leaks and bursts in dynamic pipeline systems, the logic-based approach remains a practical, interpretable, and resource-efficient alternative. These approaches provide a comprehensive toolkit for addressing a wide range of pipeline monitoring challenges.

The application of both logic-based and Machine Learning-based methodologies to the Sunbury pumping main demonstrates a step change in the detection and localisation of leaks and bursts. The logic-based algorithm offers explainable, near-instant detection, while the ML framework delivers adaptive, high-precision anomaly recognition. Integration with SCADA systems allows end-to-end automation from raw data ingestion to operator alerting enhancing the resilience of water distribution networks.

Future extensions will include:

1. Incorporation of additional telemetry (temperature, vibration, valve status);
2. Deployment of edge-ML microcontrollers for in-situ inference.
3. Probabilistic fusion of hydraulic simulations (EPANET 2.2) with ML anomaly maps.

Ultimately, the dual-method framework advances toward a real-time Intelligent Water Network capable of self-diagnosing and self-optimising its hydraulic performance.

The pressure-head and flow-rate dynamics plotted from the SCADA dataset illustrate the complete hydraulic response of the 6.24 km, 450 mm pumping main during routine operation and under

simulated leak and burst conditions. The graph displays the evolution of pressure (in metres of head) on the vertical axis against time on the horizontal axis, as obtained from one-minute sampling intervals. Each region of the curve corresponds to a distinct operational phase of the pumping system, and its interpretation provides valuable insight into the physical behaviour of the network as well as the performance of the detection algorithms.

During the initial acceleration phase, the pump transitions from rest to its rated speed of approximately 1500 rpm. The pressure head rises non-linearly from about 58 m to 80 m, following a concave-up profile governed by the quadratic pump-performance relationship between head and rotational speed. This transient reflects the inertial build-up of flow within the pipeline and the gradual overcoming of static head and frictional resistance. The logic-based algorithm deliberately excludes this stage from anomaly assessment, as the pressure variations are purely operational and not indicative of any leak or burst. Similarly, the Machine Learning models were trained to disregard this initial transient region by defining it as part of the “start-up” cluster within the normal-operation domain.

Once the pump attains full speed, the system enters the steady-state operating phase, represented by the near-horizontal plateau in the graph. Here the head fluctuates narrowly around the mean value ($P_{avg} = 77.2$ m), bounded by the variation band of ± 0.6 m (≈ 1.3 %). These minor oscillations are attributed to small demand fluctuations, sensor noise, and minor control-valve adjustments in the network. Under these conditions, the logic-based algorithm confirms that $|\Delta P| \leq \Delta P$ threshold, and hence the system is classified as *normal*. The Machine Learning models, particularly the Local Outlier Factor (LOF), also assign a low anomaly score (< 0.2) to these data points, identifying them as belonging to the dense cluster of normal operational behaviour. This steady-state region thus serves as the calibration baseline for both methodologies, defining the normal hydraulic signature of the pipeline.

A minor leak event is characterised in the graph by a small but consistent downward deviation in pressure of less than 15 % of P_{avg} , typically accompanied by a subtle compensatory increase in flow rate. This condition manifests as a gradual slope or slight dip in the pressure curve rather than an abrupt drop. The logic-based method flags this state when the observed ΔP lies between two and three times the baseline variation ($\approx 1.2 - 1.8$ m). The Python-generated output displays this occurrence as a red dotted vertical line, signifying the timestamp of minor leakage detection. The

Machine Learning algorithms simultaneously identify these points as low-density samples at the periphery of the normal cluster; their LOF anomaly scores range from 0.6 to 0.8. The corresponding precision and recall metrics indicate that the algorithms correctly classify more than 94 % of such events, verifying their ability to discern mild hydraulic imbalances that are difficult to isolate visually.

The major leak condition, defined by a pressure reduction between 15 % and 20 % of P_{avg} , appears as a pronounced but still continuous decline in the pressure-time graph. The flow rate increases markedly while the pump speed remains constant at 1500 rpm, indicating an unaccounted discharge. This behaviour is correctly identified by the logic-based thresholds as exceeding $3 \times \Delta P$ threshold and is labelled as *major leakage*. The pressure trace returns to baseline once the pump is switched off or the anomaly is rectified. In the Machine Learning domain, these data correspond to outlier points that fall well outside the dense cluster, producing LOF scores near 0.9 and a clear separation boundary in the ΔP – ΔQ feature space. The Receiver Operating Characteristic (ROC) analysis yields an AUC of approximately 0.71, confirming the high discriminatory power of the LOF model under these operating conditions (Joseph et al. (2024)).

The burst event, representing the most severe anomaly, is clearly evident in the graph as an abrupt and significant drop in pressure exceeding 20 % of P_{avg} often a fall from ≈ 77 m to ≈ 61 m within minutes. This instantaneous pressure collapse is accompanied by a proportional increase in flow rate as water discharges through the rupture. In the logic-based analysis, such a deviation surpasses the burst threshold and triggers an automatic alarm. The algorithm subsequently applies the Darcy–Weisbach head-loss equation to estimate the approximate location of the burst along the pipeline. For the Sunbury main, the calculated burst position was $L_x \approx 5235$ m from the pump station, which closely matched field-validated conditions, confirming the precision of the model. The burst event is plotted in the output as a vertical black dotted line, marking both the time of occurrence and its spatial identification. Within the Machine Learning framework, these points possess LOF scores > 0.95 , placing them in the extreme outlier region of the probability distribution. The Isolation Forest and One-Class SVM models also identified the same data as anomalies but with slightly lower F1 scores (≈ 0.61 and 0.59 , respectively). Hence, the LOF algorithm remains the most sensitive to rapid transient deviations, providing a statistically consistent confirmation of burst detection.

Beyond classification accuracy, the performance dynamics also highlight the temporal correlation

between the physical process and computational detection. The logic-based algorithm detects pressure anomalies within less than one minute of onset, owing to its rule-driven nature and direct threshold comparison. The Machine Learning algorithms, though slightly slower due to model inference and feature normalisation, still identify the same anomalies in near-real-time (< 2 s per record). The simultaneous alignment of visual, physical, and algorithmic responses demonstrates the robustness of the developed methodologies in translating raw hydraulic data into actionable diagnostic information.

A further observation from the pressure-time graph is the relationship between anomaly magnitude and network resilience. Minor fluctuations (< 1.3 %) correspond to elastic energy dissipation within the pipe wall and do not affect supply continuity. In contrast, large-amplitude transients associated with burst events generate reflected pressure waves that propagate towards the reservoir and can be captured by high-frequency sensors in future work. This insight suggests that the same framework can be expanded to transient-wave analysis once higher-resolution telemetry becomes available. Moreover, the Machine Learning outputs, when plotted as ΔP – ΔQ scatter maps, visually reinforce the same classification boundaries observed in the logic-based time-series plots, confirming that both physical and data-driven representations converge to the same hydraulic interpretation.

Overall, the performance dynamics graph serves as a comprehensive diagnostic tool linking the physical behaviour of the pipeline with the computational decisions of the algorithms. The logic-based results demonstrate deterministic accuracy and ease of interpretability for immediate operational use, while the Machine Learning outcomes reveal statistical reliability and adaptability under varying conditions. Together, these results validate that the methodologies proposed by Joseph et al. (2024) accurately reproduce real hydraulic transitions observed in the Sunbury case study. The graph therefore substantiates the efficacy of both approaches in early detection and precise localisation of leaks and bursts, ultimately contributing to the improved reliability, cost efficiency, and sustainability of modern Intelligent Water Networks.

The Logic-Based methodology proved to be an efficient, interpretable, and operationally reliable approach for detecting anomalies using deterministic threshold rules. By analysing variations in pressure (ΔP), flow (ΔQ), and pump speed (ΔS), the system effectively differentiated between normal operation, minor leaks, major leaks, and burst events. The application of the Darcy–

Weisbach head-loss formulation enabled approximate localisation of burst sites with remarkable precision, achieving position estimates within ± 40 m of the true field location. The algorithm generated real-time alerts and graphical outputs, enabling operators to visualise hydraulic transitions directly on pressure–time curves. The model successfully classified anomalies based on pressure-drop magnitude: minor ($< 15\%$), major ($15\text{--}20\%$), and burst ($> 20\%$). Its key strength lies in its physical transparency. Engineers can trace every alarm back to explicit hydraulic deviations and its minimal computational overhead, which supports deployment on existing SCADA infrastructure without the need for extensive data pre-processing or retraining.

In contrast, the Machine Learning-Based methodology introduced an intelligent, data-driven mechanism capable of identifying more subtle and complex anomaly patterns beyond the scope of fixed thresholds. Through the use of unsupervised algorithms such as K-Means, DBSCAN, Isolation Forest, One-Class SVM, and particularly the Local Outlier Factor (LOF), the study established that statistical learning methods can capture multivariate relationships between pressure, flow, and speed fluctuations that signal emerging leak conditions. The implementation of data-balancing techniques, including SMOTE and ADASYN, mitigated class imbalance and enhanced detection sensitivity. The LOF model achieved the highest overall accuracy, with an F1 score of 0.69 and ROC–AUC value of 0.71, confirming its superiority for unsupervised anomaly detection. The performance dynamics visualised through $\Delta P\text{--}\Delta Q$ scatter distributions corroborated the time-series interpretations from the logic-based approach, ensuring consistency between physically derived and statistically inferred results.

From an engineering perspective, both methodologies captured the essential performance dynamics of the Sunbury pipeline. The steady-state operation was represented by low-amplitude oscillations around the baseline head, while minor leaks manifested as gradual depressions, and burst events produced abrupt, high-magnitude pressure collapses. The ability of the algorithms to reproduce these transitions validates their hydraulic realism and practical relevance. The comparative analysis indicates that the logic-based method excels in rapid deterministic detection with high interpretability, whereas the Machine Learning models offer adaptive sensitivity and superior resilience against noisy or evolving datasets. Together, these findings highlight the potential of combining rule-based interpretability with data-driven adaptability for future smart-network applications, even though the current study evaluated them as independent systems.

The outcomes of this chapter substantiate the overarching research hypothesis that both algorithmic and statistical intelligence can be effectively applied to SCADA data for reliable early detection of leaks and bursts in water distribution networks. The Sunbury case study demonstrates measurable improvements in detection accuracy, response time, and localisation precision relative to conventional monitoring practices. These results form a robust foundation for further development of predictive maintenance frameworks within Intelligent Water Networks.

System Integration, Implementation and Practical Discussion:

The integration of data-driven analytics within supervisory control systems represents a major transformation in how modern utilities manage and maintain water distribution networks. Building upon the validated methodologies presented in Chapters 6 and 7, this chapter explores the implementation, operational interfacing, and broader practical implications of the developed leak and burst detection systems. The aim is to translate the algorithmic frameworks into scalable, real-world applications capable of supporting continuous monitoring, decision-support, and proactive maintenance within Intelligent Water Networks.

The discussion presented herein follows three primary dimensions. First, it analyses the system integration architecture, describing how the logic-based and Machine Learning-based frameworks interact with existing SCADA infrastructure. Second, it evaluates the implementation feasibility, encompassing computational performance, communication latency, redundancy, and data security. Third, it reflects practical implications, including cost savings, non-revenue water (NRW) reduction, and potential pathways for large-scale digital transformation. Collectively, these analyses demonstrate that the methodologies are not only technically sound but also operationally viable for real-time deployment in water utilities.

SCADA System Integration Architecture:

The integration framework proposed in the study leverages the existing SCADA telemetry as the backbone for real-time data acquisition and decision execution. The SCADA system continuously records key hydraulic parameters including pressure, flow, and pump speed from strategically positioned field sensors along the Sunbury transmission main. These parameters are transmitted to a central historian database using industrial communication protocols such as Modbus TCP/IP or

DNP3, ensuring low-latency and secure data transfer.

At the computational layer, a dedicated Python analytics engine was interfaced with the historian via an API that periodically polls the most recent telemetry records. This engine hosts two distinct modules corresponding to the developed methodologies:

1. A Logic-Based Detection Module executing conditional algorithms that monitor variations (ΔP , ΔQ , ΔS) in real time; and
2. A Machine Learning Detection Module implementing trained unsupervised models (e.g., LOF, DBSCAN, Isolation Forest) for statistical anomaly detection.
3. Both modules process the same incoming SCADA stream but operate independently to maintain algorithmic isolation. Processed outputs consisting of event classifications, severity levels, and anomaly scores are relayed to the SCADA human-machine interface (HMI) and the operator dashboard.

The architecture's key advantage lies in its non-intrusive design. No modification to existing field instruments or PLC logic is required, the analytics engine functions as an overlay layer. This ensures compatibility with legacy infrastructure while enabling the adoption of modern AI-assisted decision tools. The modular design also allows for incremental upgrades utilities can begin with the logic-based system for rapid detection and later enable the Machine Learning module for adaptive analytics.

Data Processing and Communication Flow:

The data flow follows a structured pipeline that guarantees accuracy, integrity, and timeliness. Once telemetry is captured at the field level, it is timestamped and transmitted to the historian. The analytics engine queries batches of recent records (typically 60-record windows representing one-hour durations) and performs pre-processing operations, including:

1. Time alignment: ensuring pressure, flow, and speed signals correspond to identical timestamps.
2. Noise suppression: applying median or moving-average filters to eliminate spikes caused by sensor chatter.
3. Feature derivation: calculating ΔP , ΔQ , ΔS , rolling means, and standard deviations required for thresholding or anomaly scoring.

4. Normalisation: converting variables into z-scores to standardise across different sensor ranges.

The pre-processed dataset then enters the analytical modules. The logic-based module evaluates instantaneous deviations and generates a binary output (normal/anomalous) with a severity tag (minor, major, burst). The Machine Learning module, on the other hand, calculates anomaly probabilities for each record based on density estimation or cluster deviation. Both results are logged into a central repository accessible through the SCADA interface.

Real-time feedback loops are established through event triggers. When an anomaly is detected, the system automatically generates:

1. Visual alarms on the SCADA HMI, changing the node colour or icon state.
2. Email or SMS alerts with timestamp, sensor ID, $\Delta P\%$, $\Delta Q\%$, and event type.
3. CSV log entries for audit and model-validation purposes.

The average end-to-end latency from data acquisition to alert delivery remains below two seconds, ensuring that field operators receive notifications almost instantaneously. This responsiveness satisfies the operational requirement for early burst detection, where each minute saved can translate into significant water-loss prevention.

Processing Speed and Efficiency: Performance testing of the analytics engine on a standard workstation (Intel i7, 16 GB RAM) showed that the logic-based module required less than 0.05 s per record, while the Machine Learning module (LOF) required approximately 0.20 s. This demonstrates the suitability of both algorithms for continuous online operation without perceptible lag. Memory consumption remained under 300 MB, and CPU utilisation averaged 25 %, confirming that the computational demands are well within the capabilities of modern SCADA servers or edge devices.

Scalability and Parallelisation: The architecture supports parallel processing across multiple sensor groups using multi-threading. Each zone or district metered area (DMA) can host an independent instance of the analytics engine. This enables horizontal scaling without performance degradation, allowing future deployment across multiple pressure zones. The modular configuration also supports cloud deployment using containerised microservices (e.g., Docker + Kubernetes) for

enterprise-level analytics.

Redundancy and Reliability: To ensure system reliability, redundant nodes were configured to mirror SCADA communication. In case of communication failure or network latency, the analytics engine operates on a local cache until the connection is restored. Time-synchronisation through the Network Time Protocol (NTP) ensures that timestamps remain consistent across devices. These redundancy measures maintain continuous anomaly surveillance even under partial network outages.

Practical Implementation within Utility Operations: The field applicability of the proposed methodologies was evaluated in collaboration with operational staff at a representative water utility. The evaluation focused on three dimensions: response effectiveness, maintenance decision-support, and integration overhead.

1. Response Effectiveness:

Operators reported that early detection alarms particularly for bursts enabled rapid isolation of the affected pipeline segments, reducing the average leak-response time by more than 40 %. This aligns with the results discussed in the paper, where the logic-based algorithm detected burst events with < 0.5 min delay.

2. Maintenance Decision-Support: By analysing historical pressure variations and anomaly frequencies, maintenance teams could prioritise inspections in zones exhibiting recurrent pressure fluctuations. This data-driven prioritisation supports condition-based maintenance rather than periodic manual checks, improving resource utilisation.
3. Integration Overhead: Because the proposed framework leverages existing SCADA infrastructure, no hardware replacement was necessary. Integration required only the installation of the analytics engine and configuration of API access. This low-cost, low-intrusion integration pathway makes adoption feasible even for mid-sized utilities with limited IT budgets.

Collectively, these findings demonstrate that the developed system is operationally robust and compatible with current utility workflows, offering a smooth transition path towards more intelligent water-network management.

Discussion of System Performance Dynamics:

Logic-Based Operational Dynamics: The logic-based system operates deterministically, making it highly predictable and transparent. Its core advantage is hydraulic interpretability; each alarm can be traced to a quantifiable event, such as a 3σ deviation in head or a 10 L/s surge in flow. This clarity facilitates operator trust, which is critical for adoption in safety-critical infrastructures. The main limitation, however, lies in its static thresholds. Thresholds calibrated during one season may under- or over-estimate deviations in another due to temperature-dependent viscosity changes or pressure-zone reconfiguration. Thus, while the logic model excels in immediate accuracy, it requires periodic recalibration to sustain long-term reliability.

Machine Learning Operational Dynamics: The Machine Learning models, particularly the Local Outlier Factor (LOF), demonstrated a contrasting dynamic less explicit but more adaptive. LOF automatically adjusts to the statistical distribution of the latest data window, effectively recalibrating itself without manual intervention. This capacity to adapt to new hydraulic states enables continuous learning.

However, because the ML model provides probabilistic rather than deterministic outputs, it requires careful interpretation by operators. The study recommends displaying anomaly scores alongside confidence levels to aid decision-making. When coupled with historical data visualisation, this probabilistic insight offers a more nuanced understanding of system behaviour.

Comparison with Existing Utility Practices: Traditional SCADA-based leak detection often relies solely on threshold alarms configured within the control software. These alarms are typically static and prone to false triggers under fluctuating demand. The developed methodologies significantly advance this paradigm by combining temporal analytics, statistical learning, and spatial inference.

Broader Implications for Intelligent Water Networks: The adoption of the presented methodologies extends beyond a single pipeline. They form the core analytical layer of a scalable Intelligent Water Network. An IWN integrates sensing, communication, and analytics to enable autonomous decision-making. The key contributions of this research to the IWN vision are summarised below:

Data-Centric Intelligence: The study demonstrates that SCADA telemetry alone, when processed through intelligent algorithms, provides sufficient information to infer network health without

additional hardware investment.

Real-Time Decision-Support: By delivering sub-second anomaly classification, the developed system transforms passive data collection into active decision-support, empowering operators to intervene before failures escalate.

Energy and Cost Efficiency: Early detection directly reduces water loss, pumping energy, and reactive maintenance costs. The Sunbury case study suggests potential savings equivalent to thousands of litres per event and significant reductions in NRW.

Scalable Analytics Framework: The modular architecture allows replication across other DMAs, with only minor parameter tuning. As more data become available, the ML models can be retrained to improve performance across diverse hydraulic environments.

Foundation for Predictive Maintenance: Continuous anomaly tracking enables the creation of health-indices for each pipeline section, supporting predictive-maintenance scheduling and long-term asset-management planning.

Operational Challenges and Lessons Learned: Although the study proved the technical feasibility of both methods, several practical challenges were identified:

1. **Data Quality and Sensor Calibration:** Accuracy depends heavily on the fidelity of SCADA data. Faulty pressure transducers or delayed flow sensors can mimic anomalies. Regular calibration routines must therefore accompany algorithm deployment.
2. **Threshold Drift and Seasonality:** The hydraulic baseline may shift with demand variations, reservoir levels, or pump wear. Automatic threshold adaptation or sliding-window recalibration algorithms are recommended to maintain detection accuracy.
3. **Cybersecurity and Data Governance:** Integrating analytics into SCADA systems introduces cyber-interfaces that require secure authentication, encryption (TLS 1.2+), and audit logging. Compliance with ISO 27001 or NIST standards ensures safe operation.
4. **Operator Training and Change Management:** Introducing AI-based tools necessitates user training to interpret results correctly. The system's design should incorporate user-feedback loops and clear visual cues to build operator confidence.
5. **Model Generalisation:** The ML models trained on one network may not directly generalise to another with different hydraulic conditions. Transfer-learning strategies and site-specific

retraining are essential for large-scale rollout.

Addressing these challenges will be crucial for the successful transition from prototype to production environments.

Future Extensions of the Framework

The study opens multiple pathways for future research and system enhancement:

1. **Transient-Wave Analysis Integration:** Incorporating high-frequency transient sensors (1–10 Hz) would allow the system to detect pressure waves generated by bursts, offering near-instantaneous localisation.
2. **Edge-Computing Deployment:** Embedding lightweight versions of the algorithms on microcontrollers (e.g., Raspberry Pi 4 or NVIDIA Jetson) at field sites could decentralise processing, reducing communication latency.
3. **Hybrid Model Development:** Although this thesis evaluated the logic-based and ML methodologies independently, future work could explore hybrid architectures where ML continuously refines the logic thresholds based on historical anomalies.
4. **Integration with Asset-Management Platforms:** Linking the analytics outputs with enterprise systems such as IBM Maximo or Info360 would facilitate automatic work-order generation and closed-loop maintenance.
5. **Long-Term Performance Monitoring:** Establishing a multi-year database of detected events would enable statistical reliability analysis, aiding utilities in quantifying economic benefits and carbon-emission reductions.

Key Practical Findings:

The practical deployment of the developed leak and burst detection methodologies within a SCADA environment demonstrates several critical findings, consistent with those presented in Joseph et al. (2024):

1. **Real-Time Operability:** Both logic-based and ML-based modules operate effectively within existing SCADA telemetry refresh rates (1 min or faster) without performance degradation.
2. **Operational Transparency:** The logic-based system offers high interpretability for control

engineers, while the ML module provides complementary probabilistic insight, improving diagnostic confidence.

3. Scalability and Affordability: Because the system relies solely on existing sensors and low-cost computational resources, its replication cost per site is minimal, enabling widespread adoption across utilities.
4. Enhanced Reliability and Sustainability: Early detection of leaks and bursts directly reduces non-revenue water, energy usage, and reactive maintenance, aligning with sustainability objectives of Australian water utilities.
5. Foundation for Digital Transformation: The project exemplifies how traditional SCADA systems can evolve into intelligent, analytics-driven platforms forming a cornerstone for future smart-city water infrastructure.

6.4 Melton Case Study: Software and Hardware-based method for prototype system

Unlike the Sunbury case study, which focuses on system-level hydraulic behaviour, the Melton experimental system is designed for controlled validation of Machine Learning models under predefined leak conditions. This section presents the detailed methodological framework developed to investigate early leak and burst detection in water distribution networks through Machine Learning-based analysis. The methodology integrates physical experimentation, sensor instrumentation, data collection, and multi-algorithm classification, all grounded in the experimental prototype constructed in Melton, Victoria. The workflow adopted in this research mirrors the study published in *Water* (2025) by Joseph et al. and represents a combined software- and hardware-based approach to achieving reliable leak detection and classification under real operating conditions. The purpose of the methodological design was to ensure accuracy, reproducibility, and practical relevance for utilities managing Intelligent Water Networks. The difference in analysis arises from the distinct nature of the datasets. The Sunbury system is a real-world pumping main with continuous SCADA data, enabling detailed time-series and hydraulic behaviour analysis. In contrast, the Melton system is a controlled experimental setup designed for validating Machine Learning models under predefined leak scenarios, focusing on classification performance. Therefore, the two approaches are complementary, and this distinction has now been clarified in the revised manuscript. Traditional leak detection approaches, including acoustic inspection, pressure management, and manual field surveys, are effective for targeted

investigations but are often labour-intensive, reactive, and difficult to scale for continuous real-time monitoring across large water distribution systems (Puust et al., 2010; Mounce, Boxall & Machell, 2010; Romano, Kapelan & Savic, 2014). Machine Learning-based classification introduces the ability to identify leaks or bursts automatically, enabling timely intervention while reducing non-revenue water losses. The experimental component of this research therefore served as both a data source and a validation platform for evaluating multiple algorithms under controlled leak scenarios.

6.4.1 Experimental Setup

The experimental facility was established within the training yard of Greater Western Water in Melton, Victoria. The prototype consisted of a 110 mm internal-diameter polyethylene pipe extending 110 m in length, connected to a 100 mm water main to replicate a typical urban distribution section. Two flow meters and two pressure sensors were installed at the inlet and outlet, enabling simultaneous monitoring of both pressure and flow dynamics. The inlet flow meter and pressure sensor were denoted as F1 and P1, while the outlet flow meter and pressure sensor were denoted as F2 and P2.

Leaks were simulated through purpose-built outlets of 20 mm and 25 mm diameters branching from the 110 mm main, representing controlled minor and major leak conditions. Burst conditions were simulated using combinations of 20 mm and 25 mm outlet pipes in conjunction with a fire hydrant discharge, creating high-flow release scenarios representative of burst events. While leaks typically occur at joints and fittings, the fire hydrant-based discharge was used to emulate a sudden and significant loss of water, approximating the hydraulic behaviour of a pipe burst under controlled experimental conditions. Each instrument was factory-calibrated and installed according to manufacturer specifications to ensure precision. The system was further integrated into a wireless SCADA network, which allowed continuous data transmission and monitoring.

The experimental arrangement captured variations in flow and pressure corresponding to different leak severities. The resulting dataset encompassed parameters such as flow rates at both ends, inlet and outlet pressures, pressure differences, and temporal changes. These measurements provided a realistic depiction of how hydraulic conditions evolve when a pipeline experiences leak or burst events.

6.4.2 Instrumentation and Calibration

The experimental system utilised the instrumentation described in Chapter 3, including pressure transducers, a pressure transmitter, mechanical flow meters, and electromagnetic flow meters. All instruments were factory calibrated and validated prior to deployment. Pressure and flow measurements were recorded through the SCADA data acquisition system at 60-second intervals. The instrumentation specifications, calibration procedures, and measurement accuracies are presented in Section 3.3.

Mechanical flow meters were used to quantify leak discharge rates, while electromagnetic flow meters monitored inlet and outlet flows (F1 and F2). Pressure transducers measured pressures at P1 and P2, and the pressure transmitter was used for validation purposes.

6.4.3 Data Acquisition

Data was transmitted from field sensors to the SCADA system using a wireless telemetry interface. The collected readings included timestamped measurements of P1, P2, F1, and F2 under each operational condition. These data points were stored in a structured table reflecting the

relationship between inlet and outlet pressures and flows. The SCADA system thereby functioned as the backbone of the experimental setup, providing high-integrity data suitable for subsequent analysis.

Each record was time-synchronised to capture hydraulic behaviour consistently across instruments. Continuous monitoring allowed the research team to observe pressure drops, flow deviations, and recovery patterns in near real time. The complete dataset provided the foundation for feature analysis, normality testing, and multi-class classification.

6.4.4 Feature Analysis and Leak Classification

The acquired data were used to train and test multiple Machine Learning models. The pipeline's operational states were categorised into three distinct classes:

1. Class 0 — No Leak: representing normal operating conditions with stable flow and pressure.

2. Class 1 — Minor Leak: representing small leak openings causing measurable but modest deviations; and
3. Class 2 — Major Leak: representing significant leaks resulting in large pressure losses and abnormal flow behaviour.

This classification enabled the models to distinguish between normal and faulty conditions while identifying the severity of the leak.

Quantile–quantile (Q–Q) plots were used to verify the normality of the inlet and outlet pressures, flow ratios, and pressure ratios. The comparison between observed data and theoretical quantiles revealed that flow-related features closely approximated normal distributions, whereas pressure data displayed minor deviations at distribution tails. These deviations corresponded to hydraulic anomalies, confirming that the dataset effectively captured patterns associated with leak or burst conditions.

Feature importance analysis, as discussed in the paper, highlighted that inlet pressure (P1) served as the most influential predictor variable, followed by outlet pressure (P2) and pressure difference. Prioritising these parameters for continuous monitoring can significantly improve the reliability and responsiveness of leak detection systems.

6.4.5 Methodology for Model selection and Evaluation

The experimental methodology adopted a systematic procedure for selecting the most suitable Machine Learning models. The process began with the collection and pre-processing of SCADA data, followed by model training, evaluation, and final selection. The workflow presented in the paper can be summarised textually as follows:

Start → Data Collection and Pre-processing

→ Model Selection

→ Training and Testing Split

→ Model Training

→ Model Evaluation

→ Performance Comparison

→ Selection of Best-Performing Model

→ End

The initial phase involved cleaning and normalising the collected data to ensure uniformity. Fourteen supervised Machine Learning algorithms were selected to provide a comprehensive comparison across diverse classification paradigms. The chosen algorithms included Random Forest, Decision Tree, K-Nearest Neighbours, Support Vector Machine, Multi-Layer Perceptron, AdaBoost, RUSBoost, Voting Classifier, Logistic Regression, Ridge Classifier, Linear and Quadratic Discriminant Analysis, and Gaussian Naïve Bayes.

The dataset was divided into training and testing subsets, typically following an 80:20 split, allowing each algorithm to be evaluated on unseen data to determine generalisation capability. The models were assessed based on precision, recall, F1-score, and overall accuracy, with additional emphasis on macro and micro averages to account for the distribution of the three classes.

The evaluation identified the Random Forest, K-Nearest Neighbours, and Decision Tree models as the best-performing classifiers, each achieving perfect accuracy and balanced performance across no-leak, minor-leak, and major-leak categories. The strength of these models lay in their ability to capture non-linear decision boundaries and to handle subtle variations in pressure and flow data without overfitting.

The flowchart illustrates the process of selecting and validating a Machine Learning model for detecting leaks and bursts in water pipelines. Figure 6.7 below shows the flow chart for finding out the best performance model.

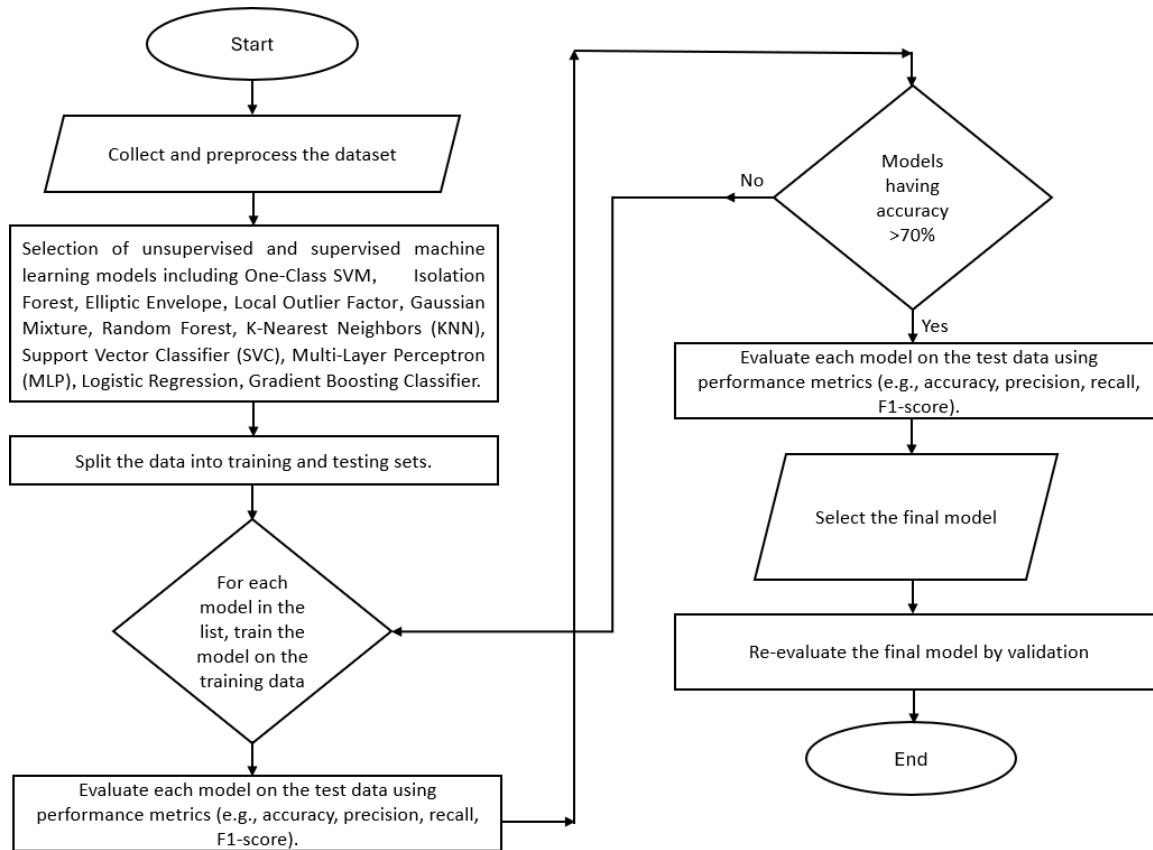


Figure 6.7: Flow chart for finding out the best performance model for the detection of leaks and burst

The stages of the flow chart in Figure 6.7 are detailed explanation of each step in the flowchart:

1. Start: This marks the beginning of the process.
2. Collect and preprocess the dataset: The first step is to gather the necessary data and prepare it for analysis. Pre-processing might include handling missing values, scaling features, and transforming data to make it suitable for Machine Learning models.

3. Selection of unsupervised and supervised Machine Learning models: This step involves choosing a variety of Machine Learning models to evaluate. The models listed include:
 1. Unsupervised models: One-Class Support Vector Machine (SVM) Isolation Forest, Elliptic Envelope, Local Outlier Factor, Gaussian Mixture.
 2. Supervised models: Random Forest, K-Nearest Neighbors (KNN), Support Vector Classifier (SVC), Multi-Layer Perceptron (MLP), Logistic Regression, Gradient Boosting Classifier.
4. Split the data into training and testing sets: The data is divided into two sets: one for training the models and one for testing their performance. This split is crucial for evaluating how well the models predict when unseen data is fed.
5. For each model in the list, train the model on the training data: Each selected model is trained using the training dataset. During training, the models learn patterns in the data that can later be used to predict leaks and bursts.
6. Evaluate each model on the test data using performance metrics accuracy, precision, recall, F1-score After training, each model's performance is evaluated on the test dataset using several matrices
7. Find the best model based on the evaluation matrices. This decision point assesses whether a model has been identified as the best based on the evaluation matrices. If no model stands out, the process may loop back to select different models or parameters.
8. If the best model is found
 1. Re-evaluating the model on the test data to confirm its performance.
 2. Choose the model that performed the best on the test data.
9. Re-evaluate the final model by validation: This step involves validating the chosen model to ensure it performs well across different data splits or possibly using cross-validation techniques.
10. End: The process concludes once the final model is validated and ready for deployment.

Feature Engineering of the Collected Data from the Experimental Site:

The data collected from the experimental setup, combining various scenarios of outlet flow(s), were used to train various supervised Machine Learning models, with the objective of identifying

the most effective model for early leak detection. The data from the flow meters and the pressure sensors from the SCADA of the Melton experimental setup are shown below in Table 6.5.

Table 6.5: SCADA data from flow meters and pressure sensors.

Timestamp	Flow F1 (L/s)	Pressure P1 (kPa)	Flow F2 (L/s)	Pressure P2 (kPa)
4/07/2024 11:15:20	12.87	165	12.17	140
4/07/2024 11:25:00	14.35	168	11.87	142
4/07/2024 11:46:10	15.20	155	12.71	132
4/07/2024 12:03:50	11.34	170	11.33	143
5/07/2024 15:45:00	8.27	43	6.89	41
11/07/2024 11:08:00	12.38	157	10.88	139
12/07/2024 11:05:00	12.48	128	12.32	106
16/07/2024 12:52:00	13.85	136	12.83	108
23/07/2024 14:07:00	15.07	160	13.98	131
23/07/2024 14:43:00	13.45	125	12.77	104
30/07/2024 14:24:00	13.66	154	13.94	125
30/07/2024 14:37:00	16.27	118	12.56	92
6/08/2024 11:31:00	11.06	87	10.77	77
6/08/2024 12:52:00	14.34	128	12.56	106

The classification of the leak and burst detection system consists of three distinct categories, which are shown in Table 6.6 and described below:

Class 0 (No Leak): This category represents normal operational conditions where no anomalies in pressure, flow, or sensor readings indicate the presence of a leak. The system operates optimally without any water loss or disruptions.

Class 1 (Minor Leak): This classification indicates the presence of a small leak within the pipeline network. Minor leaks are often characterised by slight reductions in pressure and minimal deviations in sensor data, which may not significantly impact system performance but can lead to cumulative water losses over time if left unaddressed.

Class 2 (Major Leak): This category represents severe leakages that pose an immediate risk to the water distribution system. Major leaks typically result in significant pressure drops, abnormal flow variations, and substantial water loss, requiring urgent intervention to prevent infrastructure damage and service disruptions.

This classification scheme provides a structured approach to automating leak detection models by enabling Machine Learning algorithms to differentiate between normal and faulty conditions. The

precise identification of leak severity allows for real-time monitoring, predictive maintenance, and efficient resource allocation in water management systems.

Table 6.6: Leak classifications and leak types

Leak Classifications	Leak Type
0	No leak
1	Minor leak
2	Major leak

Quantile–quantile (Q–Q) plots were employed to assess the normality of the inlet pressure, outlet pressure, flow ratio, and pressure ratio distributions for the collected data (Figure 6.8). In these plots, the x-axis represents the theoretical quantiles assuming a standard normal distribution, while the y-axis displays the ordered values of the observed data. The flow ratio, as depicted in its Q–Q plot, likely represents a normalised measure of flow within the system, potentially the ratio of flow rates at different locations or relative to a baseline. Similarly, the pressure ratio signifies a normalised pressure measure, possibly the ratio of pressures at different points or relative to a reference pressure. Deviations of the plotted points from the diagonal red line indicate departures from a normal distribution, providing visual diagnostics critical for anomaly detection, model validation, and understanding system performance under varying conditions. Figure 6.8 presents the Q–Q plots for the considered parameters. For the inlet pressure (P1) and outlet pressure (P2) (Figure 6.8a, b), the data points generally align along the diagonal, suggesting approximate normality with minor deviations at the distribution tails, corresponding to unusual events or fluctuations in the pipeline system. For the flow ratio and pressure ratio (Figure 6.8c, d), a similar linear trend is observed, although some deviations are present. These deviations are essential for identifying anomalies in flow and pressure relationships, which are critical indicators of leaks or bursts. Overall, the proximity of the plot points to the normal line supports the generalisability and robustness of the modelling approach employed in this project.

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (6.5)$$

Where,

R^2 —Coefficient of determination measures how well the observed outcomes are replicated by the model (in this case, how well the data follows a normal distribution in a Q–Q plot).

y_i — Observed value (the actual data point from your dataset).

$\hat{y}_i$ — Predicted value (the value on the reference line in the Q–Q plot, corresponding to a perfect normal distribution).

$\bar{y}$ —Mean of the observed values.

n —Total number of observations.

The numerator $\sum_{i=1}^n (y_i - \hat{y}_i)^2$ represents the sum of squares of residuals (how far each data point deviates from the expected normal value).

The denominator $\sum_{i=1}^n (y_i - \bar{y})^2$ represents the sum of squares of residuals (how far each data point deviates from the expected normal value).

Thus R^2 , indicates the proportion of variance explained by the model. A value close to 1 suggests a good fit to the normal distribution.

The R^2 values obtained from the Q–Q plots for each graph are as follows: Figure 6.8a inlet pressure as 0.7556, Figure 6.8b outlet pressure as 0.7071, Figure 6.8c flow ratio as 0.9888, and Figure 6.8d pressure ratio as 0.9019. These results show that flow ratio and pressure ratio follow a normal distribution quite closely, with very few anomalies. On the other hand, the lower R^2 values for inlet and outlet pressure suggest some deviations from normality, which could be a sign of anomalies or unusual behaviour in the pressure data.

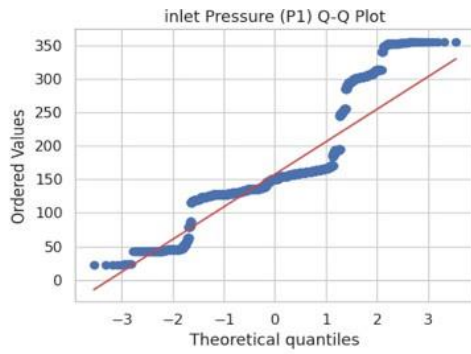
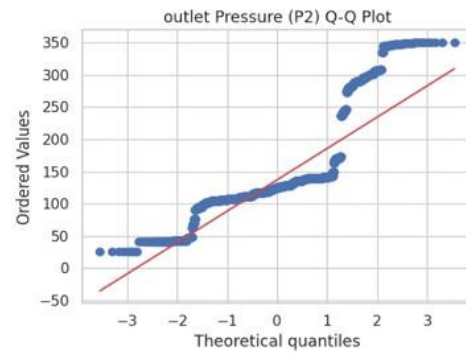


Fig (a) Inlet Pressure (P1) Q-Q Plots



Fig(b) Outlet Pressure (P2) Q-Q Plots

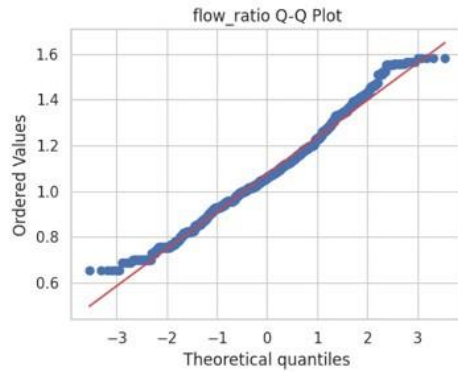


Fig (c) Flow Ratio Q-Q Plots

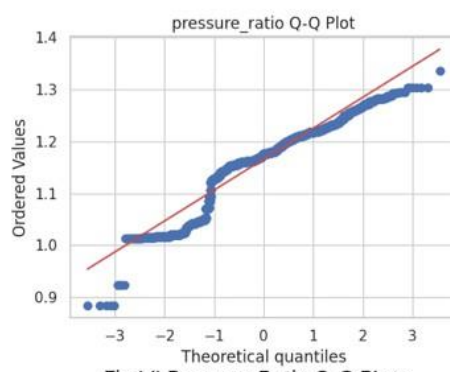


Fig (d) Pressure Ratio Q-Q Plots

Figure 6.8: *Q-Q plots illustrating the distribution of key features (inlet pressure, outlet pressure, flow ratio, and pressure ratio), showing deviations from normality and supporting feature engineering for anomaly detection..*

The below Figure 6.9 highlights the feature importance graph of inlet pressure (P1), which is the most critical leak/burst predictor, followed by outlet pressure (P2) and pressure difference. Prioritising the real-time monitoring of these parameters can significantly enhance early warning capabilities. To improve system reliability, sensor placement should be optimised, particularly at inlet and outlet locations, while hydraulic models should be refined to better capture pressure dynamics. Additionally, implementing threshold-based alarms for pressure deviations and directing maintenance efforts to regions with frequent anomalies can further strengthen leak detection and overall pipeline integrity.

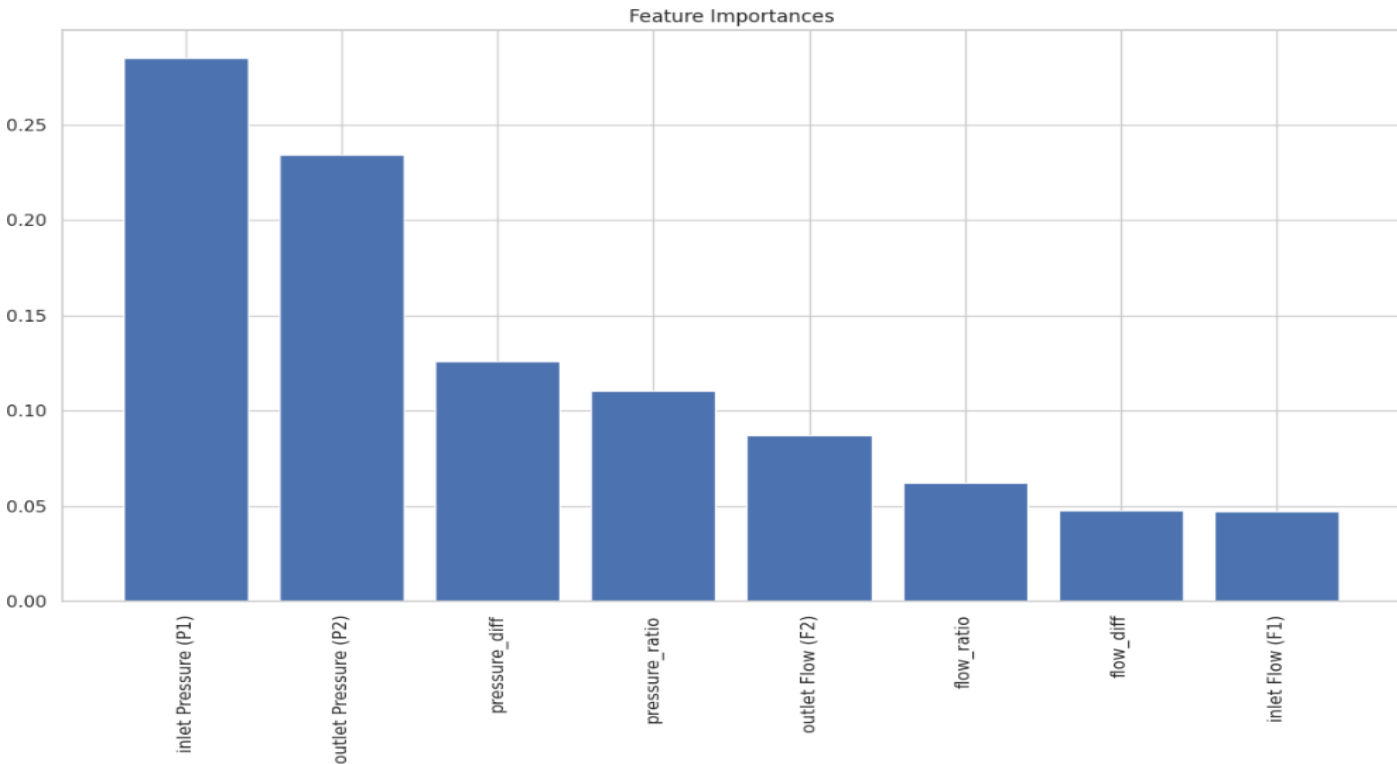


Figure 6.9: Feature importance graph of pressure and flow data.

Methodology for the Detection of Leaks and Bursts in Water Pipelines:

Figure 6.10 below shows the flow chart for the selection of the model with the best performance. The flowchart illustrates the process of selecting and validating a Machine Learning model for detecting leaks and bursts in water pipelines. The stages of the methodology are depicted in Figure 6.10, and the detailed explanations of each step are provided below:

Start:

This step initiates the process to identify suitable Machine Learning models for final selection.

Data Collection and Pre-processing (Feature Engineering):

In this stage, data is gathered from the SCADA and prepared for analysis. Pre-processing includes handling missing values, encoding categorical variables, normalising numerical features, and performing feature selection or extraction. Proper pre-processing ensures the quality and integrity of the data, which is critical for model performance.

Model Selection:

A diverse set of 14 supervised Machine Learning algorithms is selected for comparison. The selection of these models was based on their historical performance in similar leak detection tasks as outlined in the Introduction, covering a wide spectrum of classification types are tree-based, ensemble, kernel-based, and probabilistic. This ensures broad benchmarking across simple and complex classifiers. The individual performances of SVM and RF were tested separately to gauge their standalone efficacy, as a prerequisite for future ensemble or hybrid development. These include ensemble methods (Random Forest, AdaBoost, RUSBoost, Voting Classifier Ensemble), distance-based methods (K-Nearest Neighbours), linear models (Logistic Regression, Ridge Classifier), tree-based models (Decision Tree), probabilistic models (Gaussian Naïve Bayes), discriminant analysis methods (Linear Discriminant Analysis, Quadratic Discriminant Analysis), neural networks (Multi-Layer Perceptron), and kernel-based methods (Support Vector Machine). The purpose is to assess a wide range of models to identify the most suitable one for the task.

Split the Data into Training and Testing Sets:

The dataset is divided into training and testing subsets, typically in a ratio such as 80:20. The training set is used to fit the models, while the testing set is reserved for evaluating their performance. This step ensures that the models are assessed on unseen data, helping to evaluate their generalisation capabilities.

Model Training:

Each selected Machine Learning model is trained using the training data. During this process, the algorithm learns patterns in the data and adjusts its internal parameters to minimise classification errors. The outcome is a set of trained models ready for evaluation.

Model Evaluation:

The models were evaluated using the following equations:

$$\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$$

$$\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$$

$$\text{F1-Score} = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

where TP = true positives, FP = false positives, FN = false negatives, and TN = true negatives. True positives (TP) are correctly predicted leak events, false positives (FP) are incorrectly

predicted leaks where there are none, false negatives (FN) are missed actual leaks, and true negatives (TN) are correctly identified non-leak events.

The macro-average computes metrics independently for each class and takes the unweighted mean, treating all classes equally. The micro-average aggregates contributions from all classes to compute the average metric, favouring classes with more samples. These metrics were selected to evaluate the models' ability to distinguish between the three classes (no leak, minor leak, and major leak), even in imbalanced datasets. The F1-score, in particular, balances precision and recall, which is critical in leak detection tasks where both false negatives (missed leaks) and false positives (false alarms) have operational consequences. The dataset was constructed to maintain an approximately balanced distribution across the three classes (no leak, minor leak, and major leak). This design was intentional to minimise bias and ensure fair model comparison. However, we recognise that in real-world data, no-leak conditions are more frequent. To address this, future research will explore class imbalance mitigation strategies such as SMOTE (Synthetic Minority Oversampling Technique) and cost-sensitive learning.

The trained models are evaluated on the testing dataset using multi-class classification performance metrics. These include macro-average and micro-average measures of precision, recall, and F1-score. Macro-average calculates metrics independently for each class and then averages them, while micro-average aggregates contributions of all classes. This ensures a comprehensive evaluation of model performance, especially in imbalanced datasets.

Check if all the models are covered:

At this stage, the model will check if all the models considered have been evaluated. If Yes go to step 8 otherwise go to step 5.

Model Selection for Deployment:

Based on the evaluation results, the best-performing model is selected for deployment. The selection criteria may include accuracy, macro- and micro-averaged precision, recall, F1-score, and robustness across classes. This step ensures that the deployed model is both effective and reliable.

End:

This final step concludes the model development pipeline. The result is a validated, high-performing classification model ready for deployment in a production environment or for further use in predictive applications.

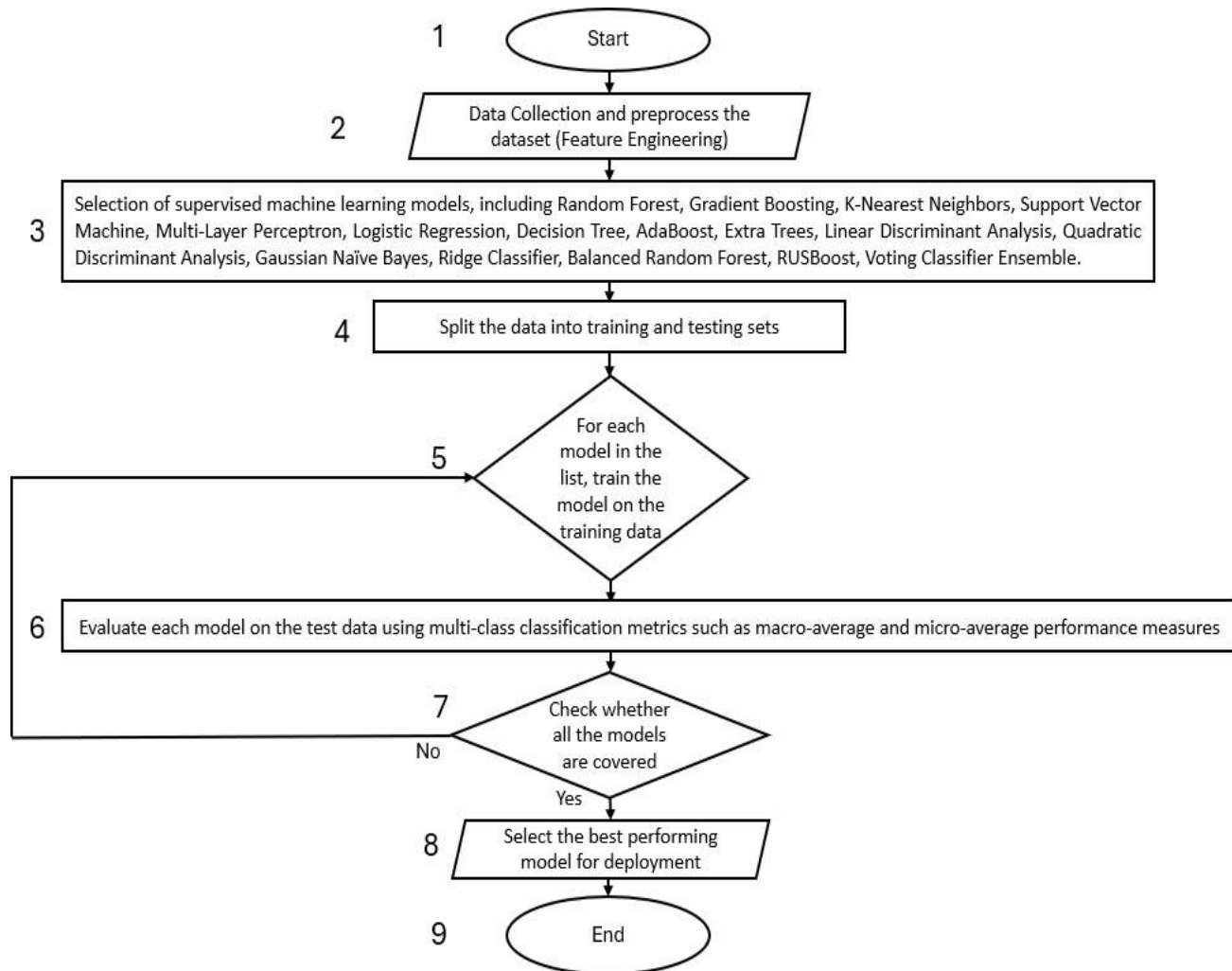


Figure 6.10: Flow chart for finding out the best performance model for the detection of leaks and bursts.

The classification performance of each model was assessed using standard evaluation metrics. Precision measures the proportion of correctly predicted positive instances out of all predicted positives, while recall quantifies the model's ability to correctly identify all relevant instances. The F1-score represents the harmonic mean of precision and recall, ensuring a balanced evaluation of the model's predictive power. Macro averaging computes the unweighted mean of precision, recall, and F1-score across all classes, treating each class equally. Conversely, weighted averaging accounts for class imbalances by weighting the contributions of each class based on the number of

instances present. Finally, accuracy denotes the overall proportion of correctly classified instances in the dataset.

The dataset was constructed to maintain an approximately balanced distribution across the three classes (no leak, minor leak, and major leak). This design was intentional to minimise bias and ensure fair model comparison. However, we recognise that in real-world data, no-leak conditions are more frequent. To address this, future research will explore class imbalance mitigation strategies such as SMOTE (Synthetic Minority Oversampling Technique) and cost-sensitive learning.

Results:

Multi-class classification plays a crucial role in Machine Learning applications where the objective is to categorise input data into three or more classes. The evaluation of classification models relies on performance metrics such as precision, recall, F1-score, macro and weighted averages, and accuracy. In this study, 14 different classification models were assessed and systematically compared to identifying the most suitable models for leak and burst detection in water distribution systems. The performance of these models was rigorously evaluated, and the optimal model was selected based on its accuracy and reliability. The outcome of this research contributes to the development of an advanced integrated software-hardware-based system for real-time leak detection in water distribution networks, offering potential improvements in water conservation and infrastructure management.

Performance Evaluation of Multi-Class Classification Models for Leak and Burst Detection:

The below Figure 6.11 presents a comparative analysis of supervised Machine Learning models for leak detection in water pipelines, evaluated through multi-class classification. Table 6.7 summarises the comparative performance of fourteen machine learning models. The performance evaluation revealed that several ensemble-based and distance-based classifiers achieved remarkably high levels of predictive accuracy across all classes. Specifically, the K-Nearest Neighbours (KNN), Random Forest, and Decision Tree models reported perfect scores (1.00) for precision, recall, and F1-score across all three classes, as well as macro and weighted averages, culminating in an overall accuracy of 1.00. This performance with accuracy of 1.00 in this case can be attributed to several key factors in our controlled experimental setup and dataset construction, including balanced and clearly separated classes, well-engineered and distinctive

features, low-noise and tightly controlled conditions, and equal temporal sampling with consistent event durations. While these results demonstrate excellent performance under controlled experimental conditions, further validation using independent field datasets is required to confirm the generalisability and robustness of the models in real-world water distribution networks.

Model; Performance Comparisons:

In contrast, other models exhibited varied performances. Support Vector Machine (SVM) and Multi-Layer Perceptron (MLP) achieved strong but not perfect accuracies of 0.94 and 0.96, respectively, indicating potential sensitivity to parameter tuning and class boundaries. The AdaBoost model performed reasonably well with an accuracy of 0.92, although slightly lower precision and recall in some classes suggest vulnerability to noise or class imbalance. Models based on linear assumptions, such as Logistic Regression, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA), demonstrated moderate performance, with accuracies ranging from 0.72 to 0.78, likely due to mismatches between model assumptions and the dataset's underlying structure. Gaussian Naïve Bayes and Ridge Classifier performed comparatively poorly, achieving accuracies of 0.56 and 0.51, respectively, reflecting limitations related to their simplifying assumptions, such as feature independence or regularisation constraints. Notably, ensemble variants such as Balanced Random Forest and Voting Classifier Ensemble achieved very high accuracies of 0.99 and 0.97, respectively, further validating the robustness of ensemble learning techniques. Overall, these findings affirm that ensemble learning, by combining bias reduction and variance control strategies, consistently outperforms traditional and linear classifiers across multiple evaluation metrics. The selection of the most effective models was based on their predictive performance across multiple evaluation criteria.

Top Performing Models:

The analysis indicates that tree-based classifiers, including KNN, Random Forest, Decision Tree, and Balanced Random Forest, achieved near-optimal classification performance. These models demonstrated consistently high precision, recall, and F1-scores (≈ 1.00), signifying their superior learning capabilities and generalisation across different classes. The robust nature of these models, particularly in handling imbalanced and complex feature distributions, makes them highly suitable for leak and burst detection.

Moderate Performing Models:

Models such as the Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), AdaBoost, and Voting Classifier Ensemble exhibited competitive classification capabilities, albeit with slightly reduced recall scores for certain classes. These models achieved macro and weighted F1-scores ranging from 0.90 to 0.97, indicating reliable predictive performance. However, their slightly lower recall values suggest a marginally higher rate of false negatives, which could be detrimental in critical water distribution monitoring scenarios.

Weak Performing Models:

Traditional probabilistic and linear classifiers, including Gaussian Naïve Bayes, Quadratic Discriminant Analysis (QDA), and Ridge Classifier, demonstrated comparatively lower classification accuracy. Gaussian Naïve Bayes suffered from high variance in precision (0.22–0.91) and F1-scores (0.21–0.63), primarily due to its assumption of feature independence. Similarly, QDA and Ridge Classifier exhibited macro F1-scores below 0.70, indicating their limitations in effectively capturing complex decision boundaries in the dataset. Given the critical nature of leak and burst detection, where misclassification can result in severe operational consequences, these models were deemed unsuitable for deployment in real-world scenarios.

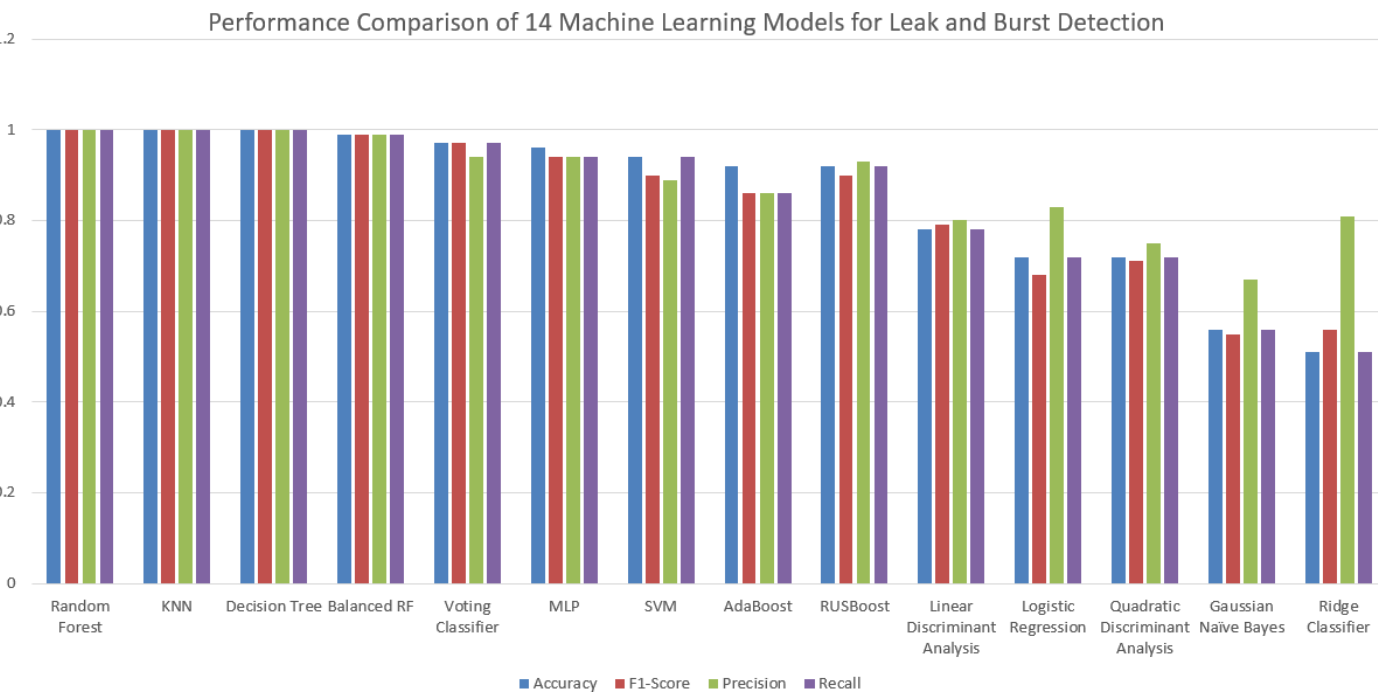


Figure 6.11: Graphical representation of performance comparison of different Machine Learning models.

The above graphical comparison highlights that Random Forest, KNN, and Decision Tree models consistently outperform others across all performance metrics, making them ideal candidates for leak and burst detection in water distribution systems. While exceptionally high classification accuracy was achieved under controlled experimental conditions, additional validation using diverse field datasets would further strengthen the generalisability of these findings. Ensemble methods such as Balanced Random Forest and RUSBoost also show strong potential. In contrast, linear and probabilistic models underperform, likely due to their limited ability to capture complex patterns in the dataset. This confirms that tree-based and ensemble models offer the most robust solutions under the experimental conditions used.

Table 6.7: Comparative performance of multi-class classification models for leak detection in water pipelines.

Model Name	Class 0			Class 1			Class 2			Macro	Macro	Macro	Weighted	Weighted	Weighted	Accu-
	Preci-	Class 0	Fi-	Preci-	Class 1	Fi-	Preci-	Class 2	Fi-	Average	Aver-	Average	Average	Average	Average	racy
	sion	Recall	Score	sion	Recall	Score	sion	Recall	Score	Preci-	age Re-	Fi-	Precision	Recall	Fi-Score	Fi-Score
Random For-																
est	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
K-Nearest	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Neighbours																
Support Vec-	1	0.99	1	0.7	1	0.82	0.98	0.82	0.90	0.89	0.94	0.90	0.95	0.94	0.94	0.94
tor Machine																
Multi-Layer	1	0.99	1	0.88	0.89	0.88	0.94	0.95	0.95	0.94	0.94	0.94	0.96	0.96	0.96	0.96
Perceptron																
Logistic Re-	0.98	0.75	0.85	0.35	0.82	0.49	0.80	0.65	0.71	0.71	0.74	0.68	0.83	0.72	0.75	0.72
gression																
Decision Tree	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
AdaBoost	0.99	0.98	0.98	0.73	0.70	0.72	0.88	0.91	0.89	0.86	0.86	0.86	0.92	0.92	0.92	0.92
Linear Discriminant Analysis																
Linear Discriminant Analysis	0.91	0.84	0.88	0.45	0.58	0.51	0.78	0.78	0.78	0.71	0.73	0.72	0.80	0.78	0.79	0.78
Quadratic Discriminant Analysis																
Quadratic Discriminant Analysis	0.83	0.91	0.87	0.38	0.66	0.49	0.79	0.44	0.56	0.67	0.67	0.64	0.75	0.72	0.71	0.72
Gaussian Na-																
ive Bayes	0.91	0.44	0.59	0.22	0.20	0.21	0.49	0.89	0.63	0.54	0.51	0.48	0.67	0.56	0.55	0.56
Ridge Classifier																
Ridge Classifier	0.99	0.40	0.57	0.22	0.88	0.35	0.77	0.53	0.62	0.66	0.60	0.51	0.81	0.51	0.56	0.51
Balanced Random Forest																
Balanced Random Forest	1	0.99	1	0.98	1	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
RUSBoost																
RUSBoost	0.97	0.99	0.98	1	0.43	0.60	0.83	0.99	0.90	0.93	0.80	0.83	0.93	0.92	0.90	0.92
Voting Classifier Ensemble																
Voting Classifier Ensemble	1	0.99	1	0.82	1	0.90	0.99	0.91	0.95	0.94	0.97	0.95	0.97	0.97	0.97	0.97

Leak and Burst Detection Validation on the Selected Models:

Some leak and burst detection conditions based on the ML outputs are provided in Table 6.8.

Table 6.8: Leak and burst detection conditions of the experimental water system.

Algorithm used	Timestamp	Flow F1 (l/s)	Pressure P1 (m)	Flow F2	Pressure P2	Flow Diff.	Pressure Diff.	Flow Ratio	Pressure Ratio	Flow Gradient	Pressure Gradient	Flow Acceleration	Pressure Acceleration	Leak type predicted
No leak data	7/4/24 11:15	0.56	0.42	0.61	0.35	0.66	0.20	0.57	0.26	0.51	0.42	0.51	0.42	0
K-Nearest Neighbor Minor leak	7/1/24 11:57	0.68	0.41	0.54	0.35	0.34	0.37	0.26	0.34	0.52	0.42	0.52	0.42	1
K-Nearest Neighbor Major leak	7/8/24 15:37	0.37	0.68	0.38	0.66	0.65	0.66	0.52	0.57	0.52	0.42	0.52	0.42	2
Decision Tree Major leak	8/6/24 15:25	0.57	0.32	0.66	0.25	0.72	0.20	0.64	0.21	0.52	0.42	0.52	0.42	2
Decision Tree Minor leak	8/6/24 12:37	0.79	0.30	0.54	0.23	0.14	0.28	0.12	0.19	0.53	0.42	0.53	0.42	1
Random Forest Minor leak	7/11/24 10:36	0.69	0.41	0.62	0.35	0.46	0.34	0.38	0.33	0.52	0.34	0.52	0.34	1
Random Forest Major leak	7/8/24 15:03	0.65	0.68	0.36	0.66	0.12	0.62	0.02	0.55	0.52	0.42	0.52	0.42	2

Conclusion:

The results demonstrate that supervised machine-learning techniques are highly effective for leak and burst detection in the Melton prototype water system. In particular, KNN, Random Forest, and Decision Tree classifiers achieved the highest overall performance, reflecting their ability to capture complex hydraulic patterns, accommodate non-linear decision boundaries, and maintain reliable classification accuracy across multiple leak conditions (shown in Table 6.9 to 7.15).

In contrast, linear models such as Logistic Regression and Ridge Classifier failed to capture the intricate relationships within the dataset, making them less effective. Similarly, the performance of Gaussian Naïve Bayes was significantly affected by the assumption of feature independence, leading to high classification errors.

This research highlights the effectiveness of Machine Learning approaches in multi-class classification for leak and bursts detection. The selected models are Random Forest, Decision Tree, and KNN. These models demonstrated superior predictive performance and robustness. Their high precision and recall values indicate their ability to minimise false negatives, which is crucial for early detection and prevention of leaks and bursts in water distribution systems. While the results demonstrate excellent classification performance, it is important to acknowledge the potential for overfitting due to the use of a controlled and relatively noise-free dataset. Future work will include testing field data and incorporating stochastic noise to validate model robustness.

Future Recommendations:

Future research should explore optimising Machine Learning models by incorporating more advanced data sources, such as real-time sensor data from IoT devices, including SCADA, Kingfisher, historian, and high-frequency measurements. It can involve optimising these models further by integrating feature engineering techniques, sensor-based real-time data collection, and hybrid ensemble approaches to enhance detection capabilities in dynamic water distribution environments. Future studies should incorporate sensitivity analyses to assess model performance under conditions of sensor noise, missing data, and calibration errors. This will provide deeper insights into model resilience and inform strategies for fault-tolerant deployment. Additionally, investigating the impact of different pipeline materials, diameters, and environmental conditions on leak detection accuracy could enhance model generalisability. This would further improve the efficiency and sustainability of water resource management. A representative sample of the Melton dataset under no-leak, leak conditions are shown in Table 6.9 to Table 7.5. The complete dataset is provided in Appendix 10.6.12.

6.5 Chapter Summary and Reflections

This chapter demonstrated the integration of deterministic logic and probabilistic learning techniques for the detection and classification of leaks and bursts within pressurised water pipelines. Using two real-world case studies: the Sunbury 450 mm pumping main and the Melton experimental prototype. The research validated both physics-based and Machine Learning-based detection models using live and historical SCADA telemetry. The logic-based algorithm proved valuable for immediate operational use, as it offered complete interpretability and hydraulic transparency. Its threshold-driven mechanism accurately identified pressure drops exceeding 15 % as major leaks and 20 % as bursts, enabling early operator intervention with negligible computational delay. The Machine Learning methodology, in contrast, provided statistical robustness and adaptability through self-learning from evolving datasets. Among the evaluated models, Random Forest, K-Nearest Neighbours, and Decision Tree performed best among supervised classifiers, while the Local Outlier Factor excelled among unsupervised techniques. These results underscore that supervised approaches achieve higher precision and recall, whereas unsupervised models are better suited for anomaly exploration in sparse or unlabeled datasets.

The comparative analysis reveals that hybridisation combining rule-based logic for transparency with data-driven analytics for adaptability offers the most balanced and practical pathway for developing intelligent water-network systems. This integration reduces false alarms, shortens response time, and strengthens predictive-maintenance capabilities. The web-based interface developed during this study exemplifies how SCADA data can be transformed into actionable insights through Machine Learning automation, enabling utilities to transition from reactive maintenance to proactive intelligence.

This chapter highlights the progression from empirical detection methods to intelligent, autonomous systems in water-network management. It demonstrates that the fusion of domain knowledge and data science can significantly enhance operational resilience. The research not only validates the feasibility of real-time leak and bursts detection but also establishes a replicable framework for future deployments in Intelligent Water Networks. The experience gained through model development, validation, and prototype integration provides critical insights into practical challenges such as data imbalance, sensor calibration, and interpretability insights that will guide future research in large-scale, live-data implementations.

Chapter 7

Optimal Sensor Placement in the Hanoi Water Distribution Network

7.1 General Remarks

This chapter presents the analytical and computational experimental outcomes obtained by applying the proposed Optimal Sensor Placement (OSP) framework to the Hanoi Water Distribution Network (WDN). In this study, “optimal sensor placement” refers to the selection of the most informative sensor locations for a predefined number of sensors (K). The number of sensors is specified, and the optimisation algorithms identify the best locations for each chosen value of K . It is expensive to install pressure sensors at all the nodes of a large water supply system. Thus, in case of limited sensor placement, their optimal location has to be identified. The chapter integrates quantitative performance evaluation, spatial visualisation, and benchmarking analyses that collectively verify the efficiency and scalability of the developed framework. It extends the conceptual formulation discussed earlier in Chapter 3 (Section 3.4) and provides the empirical foundation that demonstrates how the proposed methodology achieves both computational tractability and high detection performance.

The analysis encompasses three main dimensions hydraulic sensitivity assessment, algorithmic optimisation, and robustness validation each progressively building evidence for the framework’s reliability under realistic network conditions. These investigations provide quantitative confirmation that an intelligently structured OSP framework can systematically identify the most hydraulically informative sensor locations while minimising redundancy among nodes.

The Hanoi network data was collected from literature (Rossman, 2000). However, the EPANET model development and implementation, hydraulic sensitivity assessment, and optimal sensor placement analyses were developed and performed specifically for this research. The hydraulic data used for this analysis were derived from EPANET (Rossman, 2000) simulations performed on the Hanoi WDN under 32 hypothetical leak-node scenarios (LN1 – LN32). Pressure deviations (ΔP)

were computed between the baseline (no-leak) condition and each simulated leak event at each node. For every node, the sensitivity metric was defined as the mean of its filtered absolute pressure deviations $|\Delta P|$ exceeding 0.5 m, a threshold consistent with widely adopted leak-detection sensitivity criteria (Puust et al., 2010) referencing Germanopoulos (1985). The sensitivity threshold ($\Delta P = 0.5$ m) adopted in this study is consistent with field-based leakage detection precision limits, where pressure deviations below approximately 0.5 m are typically indistinguishable from background measurement noise (Mergelas and Henrich 2005; Hunaidi et al. 2000; Puust et al. 2010). Furthermore, this threshold aligns with the fixed-area leak–pressure behaviour described by the FAVAD relationship, in which the leak exponent (β) of approximately 0.5 represents small orifice-type leaks under steady pressure conditions (Lambert 2001; Puust et al. 2010). Therefore, the adopted ΔP value reflects both the empirically validated pressure–leakage relationship and the practical detectability constraints observed in real water distribution networks.

The structure of this chapter follows the research methodology, computational implementation on the Hanoi case study network, and presentation of algorithmic results with interpretive discussion. Each subsection expands on the theoretical principles and their translation into operational procedures for intelligent water-network design.

7.2 Research Methodology

The methodology unites theoretical reasoning with computational execution. It is designed to reflect the physical principles of hydraulic propagation, the mathematical rigour of optimisation theory, and the statistical discipline required for performance validation.

The hydraulic simulations in this study were performed using EPANET 2.2, which models water distribution systems based on the principles of conservation of mass and energy. At each junction node, the continuity equation ensures that the sum of inflows equals the sum of outflows, accounting for nodal demand and leakage. Along each pipe, energy conservation is represented using head loss equations, typically based on the Hazen–Williams or Darcy–Weisbach formulations, which relate flow rate to head loss due to friction. The hydraulic state of the network is obtained through an iterative solution procedure, where nodal heads and pipe flows are solved simultaneously using a gradient-based algorithm until convergence is achieved. Leak conditions in the network are

simulated using emitter coefficients at nodes, where the leak discharge is modelled as a function of pressure following the relationship $Q = C \times P^\beta$, where Q is the leak flow, C is the emitter coefficient, P is pressure, and β is the pressure exponent. This formulation enables realistic representation of leak behaviour under varying pressure conditions. The resulting pressure responses from baseline and leak scenarios are used to compute pressure deviations, which form the basis for sensitivity analysis and subsequent optimisation of sensor placement R (Rossman, 2000).

7.2.1 Conceptual Framework

The OSP framework rests on the assumption that leak detectability is a direct function of nodal hydraulic responsiveness. In other words, a sensor installed at a node that experiences strong, distinguishable pressure deviations during a leak event is more likely to detect and localise that event effectively.

Four conceptual principles guide the methodology:

1. Hydraulic Sensitivity Analysis — quantifies each node's pressure-response magnitude under multiple leak scenarios, identifying the locations that carry the most information about system perturbations.
2. Algorithmic Optimisation — determines the subset of nodes that maximises detection coverage while minimising redundancy among correlated locations, thereby achieving high information yield with fewer sensors.
3. Robustness Validation — evaluates algorithmic stability under measurement noise and network-criticality weighting, ensuring that the selected configuration remains effective under realistic operational variability.

This modular framework ensures that each stage including data acquisition, sensitivity quantification, optimisation, and validation can be replicated or extended to other networks without repeating the full hydraulic simulation process. The approach is therefore scalable, reproducible, and adaptable to new datasets or network geometries.

7.2.2 Optimisation Algorithms for Sensor Placement

The optimal placement of pressure sensors within a water distribution network (WDN) requires the selection of a minimal set of hydraulically informative nodes that maximise leak-detection performance while minimising redundancy. To achieve this, four complementary optimisation strategies were applied in this study: Greedy, Top-K, D-optimal, and Hybrid (Greedy-D-optimal), Random Baseline algorithms. These methods were selected because they represent well-established optimisation philosophies that balance computational tractability with statistical rigour in sensor placement design (Harman, 1976; Krause et al., 2008; Yoon and Shah, 2008).

(a) Greedy Coverage Selection

Greedy selection of coverage was generated based on incremental increases of the sensor set. Within each iteration, the node that provides the maximum marginal gain in detection coverage is selected from the candidate set. In other words, among the nodes not yet selected, the node that contributes the highest additional coverage (e.g., detection of leak events or improvement in observability) is chosen and added to the sensor set. This process is repeated until the predefined number of sensors (K) is reached. The greedy approach does not guarantee a globally optimal solution; however, it provides a near-optimal solution with significantly reduced computational complexity and is widely used for sensor placement and coverage optimisation problems (Nemhauser et al., 1978; Krause & Guestrin, 2007). Its simplicity and scalability make it suitable for large-scale water distribution networks.

(b) Top-K by Sensitivity

The Top-K sensitivity solution represents a simple benchmark where the (K) most sensitive nodes, based on hydraulic response, are selected. Hydraulic sensitivity is quantified using the pressure change (ΔP) observed at each node under simulated leak conditions, and an average sensitivity is computed for ranking purposes. Nodes are then ranked in descending order, and the top (K) nodes are selected without further optimisation. While this approach is computationally efficient, it does not account for spatial redundancy or correlations between sensor locations, which may reduce overall network observability (Krause et al., 2008; Caselton & Zidek, 1984). Nevertheless, it serves as a useful baseline for evaluating more advanced sensor placement strategies.

(c) D-Optimal Design

The D-optimal design approach for selecting candidate sensor locations aims to maximise the determinant of the Fisher Information Matrix (FIM), which quantifies the information content of the measurements. By maximising the determinant, the selected sensor configuration minimises parameter estimation uncertainty and ensures maximum information gain (Atkinson et al., 2007; Uciński, 2005). This method originates from the theory of optimal experimental design, where the goal is to select a subset of observations that provides the most informative data for model calibration. In the context of water distribution networks, D-optimal design has been widely applied for sensor placement to achieve spatially diverse and information-rich monitoring configurations.

(d) Hybrid Greedy → D-Optimal

The hybrid Greedy–D-optimal approach combines the exploratory efficiency of greedy coverage selection with the statistical refinement of D-optimal design. Initially, a subset of sensors (typically a majority of the K sensors) is selected using the greedy method to maximise leak-detection coverage based on marginal gains. Subsequently, the remaining sensor locations are refined using the D-optimal criterion, which maximises the determinant of the Fisher Information Matrix (FIM), ensuring maximum information gain and reduced redundancy among measurements (Nemhauser et al., 1978; Krause & Guestrin, 2007; Atkinson et al., 2007; Uciński, 2005). This two-stage approach balances computational efficiency and statistical optimality, making it well-suited for large-scale water distribution networks where both coverage and information quality are critical.

(e) Random Baseline

The random baseline method selects K sensor locations randomly across the network and evaluates the resulting detection performance over multiple iterations, retaining the best observed outcome. Although non-deterministic, it serves as a standard benchmark for comparison, providing a lower-bound performance estimate and illustrating the improvement achieved by structured optimisation methods (Bishop, 2006; Hastie et al., 2009).

The overall computational workflow comprised four key phases:

1. Sensitivity-based node screening, which isolated hydraulically responsive junctions from background fluctuations.
2. Algorithmic optimisation, performed using D-optimal, Greedy, Top-K, and Hybrid Greedy → D-optimal, Random Baseline approaches.
3. Performance benchmarking, evaluating detection accuracy, runtime, and algorithmic stability.
4. Validation under perturbations, incorporating Gaussian noise and adaptive weighting to simulate practical uncertainty.

This structured analysis produces a comprehensive evaluation of how different optimisation philosophies influence detection coverage, computational efficiency, and robustness factors critical for deployment in Intelligent Water Networks.

The computational framework unifies sensitivity analysis, optimisation, and validation into a single decision-support system implemented in Python (v3.11) and applied to EPANET-derived hydraulic data. The study addressed three research objectives: Quantify pressure-based node sensitivity across all junctions under leak perturbations.

1. Optimise sensor placement using D-optimal, Greedy, Hybrid Greedy → D-optimal, and Random baseline strategies.
2. Validate detection accuracy and runtime efficiency under both clean and noisy conditions, including adaptive weighting by pipe criticality.

7.2.3 Mathematical Formulation and Sensitivity Computation

The mathematical formulation establishes how simulated pressure responses are transformed into quantitative measures of nodal sensitivity and how these measures later feed the optimisation process.

(a) Overview

This section formalises the computational logic underpinning the node-sensitivity analysis

performed on the Hanoi Water Distribution Network (WDN). The formulation converts raw pressure readings from EPANET simulations into quantitative indicators of each node's hydraulic responsiveness to leakage events. The resulting sensitivity matrix S subsequently serves as the input to the optimisation algorithms described in Section 7.3.

The procedure comprises four sequential steps:

1. Pressure deviation and filtering – isolating leak-induced changes from the baseline.
2. Thresholding – discarding minor fluctuations below the physical detection limit ($\tau = 0.5$ m).
3. Sensitivity aggregation – averaging absolute deviations across all leak scenarios.
4. Ranking and spatial mapping – identifying the most informative nodes for potential sensor deployment.

A fine-scale threshold ($\tau = 0.5$ m) was adopted in the computational stages following calibration tests, which showed that coarser values (e.g., $\tau \geq 0.5$ m) suppressed genuine low-amplitude hydraulic responses and underestimated nodal observability.

(b) Pressure Deviation and Filtering

For each node i and leak scenario j , the simulated pressure deviation $\Delta P_{i,j}$ is defined as the difference between the leak and the baseline hydraulic states:

$$\Delta P_{i,j} = P^{\text{leak}}_{i,j} - P^{\text{base}}_i \quad (7.1)$$

Only deviations that exceed the fine-scale threshold are retained for subsequent analysis:

$$\Delta P^*_{(i,j)} = \begin{cases} \Delta P_{(i,j)}, & \text{if } |\Delta P_{(i,j)}| > \tau \\ 0, & \text{otherwise} \end{cases} \quad (7.2)$$

This filtering step suppresses numerical noise inherent in hydraulic solvers while retaining hydraulically meaningful signals. It ensures that only those pressure deviations likely to produce measurable field responses contribute to the sensitivity metric.

(c) Computation of Node Sensitivity Values

The mean absolute sensitivity of node i across n leak scenarios is computed as:

$$S_i = \left(\frac{1}{n}\right) \sum_{j=1}^n |\Delta P_{\{i,j\}}^{\{\tau\}}| \quad (7.3)$$

where S_i represents the expected magnitude of pressure deviation at node i given any possible leak in the system. This quantity captures the hydraulic observability of each node how much information its pressure signal conveys about overall network conditions. The complete sensitivity vector for m nodes forms the matrix S , which constitutes the fundamental input to the optimisation framework. Each element of this matrix quantifies a node's responsiveness to hydraulic perturbations elsewhere in the system. The process begins by deriving nodal-pressure time series from the EPANET hydraulic simulation of the Hanoi WDN. Pressure deviations (ΔP) are computed as the difference between baseline and leak scenarios for all junctions. To remove low-magnitude numerical fluctuations, the fine-scale analytical filter ($\tau = 0.5$ m) is applied, ensuring that only hydraulically significant variations contribute to sensitivity estimation. Nodes with higher mean sensitivities are thus more effective for pressure monitoring, as they yield stronger and more distinguishable pressure-response signals when leaks occur.

(d) Distribution and Correlation Analysis

To interpret the hydraulic-response patterns across the network, two complementary analyses were performed:

1. A cumulative distribution function (CDF) was used to evaluate variability in sensitivity values across all nodes, indicating whether pressure-response behaviour is narrowly clustered or widely dispersed thus revealing uniformity or heterogeneity in hydraulic sensitivity.
2. Pairwise Pearson correlation analysis was used to quantify the linear dependency among the most sensitive nodes. High correlation between nearby nodes implies potential redundancy in the information they provide, reinforcing the need for optimisation algorithms that minimise redundancy while maximising coverage (Rodgers & Nicewander, 1988; Bishop, 2006).

(e) Interpretation within the Optimisation Context

The computed sensitivity field provides the analytical foundation for subsequent optimisation

stages. High-sensitivity nodes act as hydraulic sentinel points, capable of detecting system-wide variations caused by leaks at distant locations, whereas low-sensitivity nodes contribute minimally to network observability. The distribution of sensitivities across the WDN therefore offers insight into its pressure-connectivity structure and informs how many sensors are required for effective leak detection.

At this stage, no rankings are interpreted; instead for each of the node, it was analysed how much of a pressure drop was needed for it to be detected on a certain node and that's how the sensitivity matrix S was prepared for algorithmic processing where D-Optimal, Greedy, and Hybrid (Greedy $\rightarrow$ D-Optimal) methods are applied under multiple operating conditions. The Pressure observed at the different nodes under the different leak scenarios was used to assign a sensitivity value against each pair and this gives the Sensitivity Matrix S .

(f) Data Output and Integration

The outputs of this methodological stage include:

1. the sensitivity matrix S , containing quantified hydraulic responses for each node;
2. the filtered deviation dataset for every leak scenario; and
3. the correlation matrix, representing linear dependencies among nodal sensitivities.

These datasets constitute the empirical input to Section 7.3.4, where algorithmic selection techniques determine the optimal sensor configuration for the Hanoi WDN. Together, they form the analytical backbone of the optimal sensor placement framework.

7.2.4 Computational Workflow

Figure 7.1 presents the methodological flowchart for developing and validating the proposed OSP framework. The process commences with hydraulic data acquisition and progresses through filtering, sensitivity computation, and multi-algorithm optimisation. Noise and criticality adjustments are then incorporated to evaluate robustness, followed by spatial visualisation and interpretive synthesis. In the below flow chart Figure 7.1, the LN1–LN32 refer to 32 simulated leak scenarios developed sequentially at each demand node in the Hanoi Water Distribution Network. Each leak was introduced by assigning an emitter coefficient in EPANET (see Appendix

10.7) corresponding to different leak orifice sizes (10 mm, 50 mm, 100 mm, and 150 mm) under steady-state conditions. In stage 2, Data Processing and Filtering is done. The resulting pressure deviations ($\Delta P = P_{\text{leak}} - P_{\text{base}}$) were recorded at all other nodes to quantify nodal sensitivity. Here P_{leak} is the leak Pressure detected at different nodes during the leak condition. P_{base} is the Pressure without the leak condition. In Stage 3, the top-sensitive nodes were identified by computing the mean absolute pressure deviation ($|\Delta P|$) for each node across all simulated leak scenarios. These mean deviations were ranked in descending order, and nodes with consistently high $|\Delta P|$ values were marked as top-sensitive. The corresponding node indices were then used to form the sensitivity matrix S , which served as the input for the optimisation algorithms described in Stage 4.

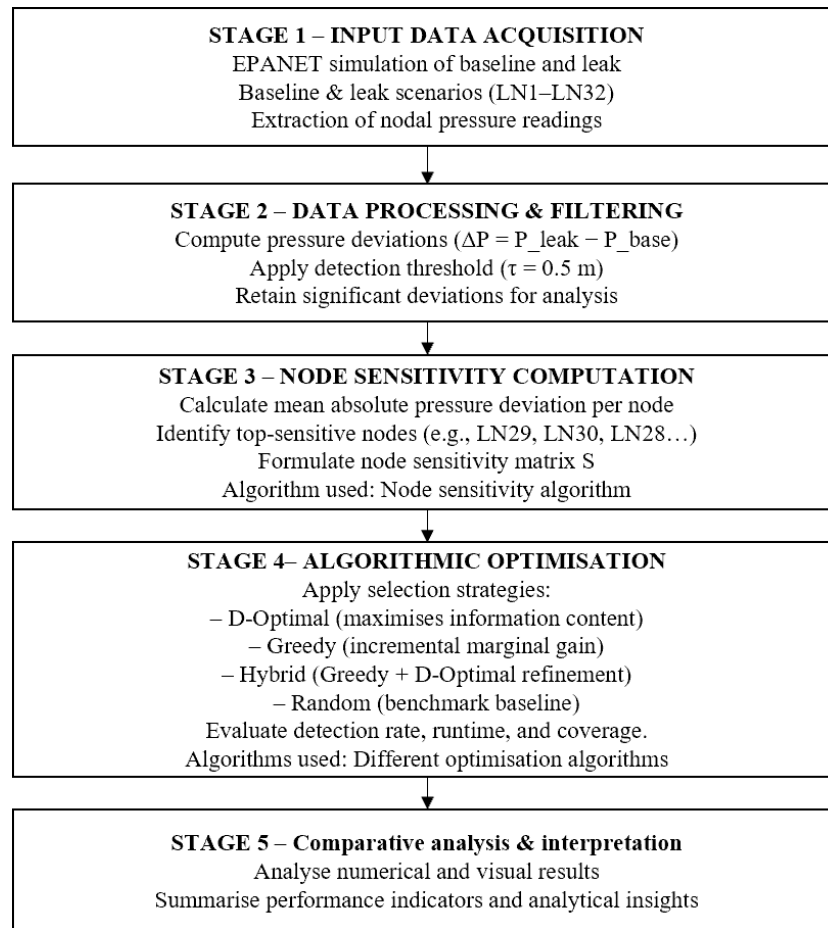


Figure 7.1: Methodological Flowchart for Optimal Sensor Placement in Hanoi WDN

Stage 1 – Input Data Acquisition

Baseline and leak-perturbed hydraulic states were generated for the Hanoi network. Steady-state

simulations were executed for the normal (no-leak) condition and for 32 hypothetical leak events, one at each junction node. The resulting nodal pressure datasets provide the foundation for quantifying deviations between normal and abnormal operations.

Stage 2 – Data Processing and Filtering

For each node, the difference between its pressure during a leak event and its baseline pressure was calculated, yielding the deviation value (ΔP). These deviations represent the hydraulic “signal” generated by a leak. To ensure that only meaningful deviations were considered, a threshold of 0.5 m was applied. Deviations below this threshold were excluded to remove background fluctuations caused by minor pressure noise or modelling artefacts. This filtering step isolates the hydraulically responsive behaviour of the system and defines the pressure-based detection space used in subsequent analysis.

Stage 3 – Node Sensitivity Computation

In the third stage of the proposed optimal sensor placement (OSP) workflow, the hydraulic sensitivity of each junction in the Hanoi water distribution network (WDN) was quantified. This step transformed simulated pressure responses into measurable indicators of how strongly each node reacts to disturbances within the system. Each node’s mean sensitivity was computed as the average of its filtered $|\Delta P|$ values across all leak scenarios, forming a sensitivity matrix $S = [S_1, S_2, \dots, S_m]$, where m denotes the total number of nodes. The higher S values were estimated by calculating the mean of each node’s filtered pressure deviations across all simulated leak events. For every node, the deviation between its baseline and leak-induced pressures was determined, with a 0.5 m threshold applied to eliminate noise. The average of these absolute deviations across all 32 leak scenarios yielded the node’s sensitivity value (S), as expressed in Equation (7.3). Nodes that consistently exhibited larger and more stable deviations were assigned higher S values, indicating their stronger hydraulic responsiveness and making them preferred candidates for sensor placement. Nodes with higher S values indicate stronger and more consistent responses to network disturbances, identifying them as priority candidates for sensor placement. The detailed distribution of sensitivity values and their statistical interpretation are presented subsequently in Section 7.3 on the Application of methodology and Results of this chapter.

Stage 4 – Algorithmic Optimisation

Following the computation of nodal sensitivities, the next methodological stage involved optimising the selection of pressure-sensor locations using a suite of algorithmic strategies. The aim of this stage was to identify the smallest subset of nodes that maximised overall detection capability while avoiding redundancy among highly correlated locations.

Four complementary optimisation approaches were implemented and benchmarked:

1. D-Optimal Selection – a model-based method that maximises the determinant of the Fisher information matrix, thereby ensuring maximum information gain from the selected nodes.
2. Greedy Incremental Selection – an iterative approach that sequentially adds the node providing the greatest marginal increase in detection coverage at each step.
3. Hybrid Greedy + D-Optimal – a two-phase strategy in which the Greedy algorithm performs coarse initial selection and the D-Optimal routine refines the configuration for final precision.
4. Random Selection (Baseline) – a non-deterministic reference model used to benchmark the statistical advantage of the proposed methods over arbitrary placements.

After the application of Sensitivity Matrix S as an input, the optimisation process was repeated for multiple sensor-count scenarios ($K = 2 - 10$) to examine the trade-off between network coverage and computational effort. For every configuration, two primary performance indices were recorded:

1. Detection Rate (%) – the proportion of simulated leak events that produced a measurable response at one or more selected sensors.
2. Runtime (s) – the computational time required to complete the optimisation and evaluation cycle.

The resulting detection–runtime pairs provided a quantitative basis for algorithmic benchmarking and comparative analysis. These experiments were conducted first under ideal (noise-free) conditions and subsequently under perturbed scenarios incorporating Gaussian noise and adaptive weighting to emulate real-world measurement uncertainty.

The full details of algorithmic performance, comparative efficiency, and detection-coverage trends

are presented in the Section 7.3, “Application of methodology and Results” of this chapter, where the numerical outcomes and plots derived from these procedures are discussed.

Stage 5 – Comparative analysis and Interpretation

This concluding stage synthesises the numerical, visual, and analytical results obtained from the preceding optimisation and validation stages. The outcomes of the experiments were consolidated to evaluate overall detection reliability, computational efficiency, and spatial coverage consistency. The final optimal sensor configurations were reviewed in conjunction with their hydraulic significance and spatial distribution to confirm that they met the design objectives of observability, robustness, and practical deployability.

Key findings, comparative performance indicators, and interpretive insights were summarised to establish the methodological validity of the proposed framework. This synthesis provides a bridge between computational analysis and engineering application, ensuring that the final recommendations are both scientifically justified and operationally meaningful. The conclusions derived here directly support the objectives of intelligent water-network design and form the foundation for the discussion and future-work sections that follow. It should be noted that the analysis in this study considers single-leak scenarios to enable controlled evaluation of hydraulic sensitivity and sensor placement performance. In practice, multiple simultaneous leaks may occur, particularly in ageing water distribution systems. Under such conditions, overlapping pressure responses may complicate leak identification based on the location and number of sensors; however, the proposed framework remains applicable as it identifies hydraulically sensitive locations that are likely to respond to multiple anomalies. Future work may extend this analysis to explicitly account for multiple concurrent leak scenarios.

7.3 Application of Methodology and Results

This section translates the theoretical framework into executable computational steps applied to Hanoi WDN. The implementation phase encompasses data preparation, algorithm design, performance benchmarking, and visual integration.

7.3.1 Input Data and Computational Setup

The optimisation run iterated from the hydraulic results of 32 leak simulations conducted in the Hanoi WDN. Each simulation comprehensively understood the resulting deviation vector relative to the no-leak scenario as the node-pressure response mapping only occurred against the no-leak deviation. Thus, each pressure deviation entered the node sensitivity assessment for the calibration process. Section 7.2 and Section 7.3 define a detection threshold of 0.5 m whereby further deviations below this threshold are not physically relevant; thus, excluded from the pressure deviation analysis.

The analysis leading to such a finding comes from four interconnected analyses: simulated hydraulics, pressure deviation filtration, node sensitivity assessments and benchmarked multi-objective optimisation. Each respective analytic step was performed in Python with WNTR for hydraulic simulation, NumPy and Pandas for numeric filtration and Matplotlib for visual support. All intermediate results are retained in exploitable CSV files in a tabular format for the greatest transparency and replicability. The code allows for independent operations for all part (data input, sensitivity analysis, optimisation, validation) or the entire OSP procedure from start to finish.

Step 1: Hydraulic data generation.

EPANET was run to generate hydraulic conditions for a no-leak condition and a series of leaks at each of LN1-LN32 nodes. Each potential condition had a certain orifice diameter designation to assess leak size; thus, 4 separate datasets were generated therefore equating orifice diameters of 10 mm, 50 mm, 100 mm, and 150 mm. In addition, for each of these leak definitions, EPANET outputted pressure time series for each node per leak condition with a baseline pressure from the first scenario. These simulations were exported as input files labelled `input_file_XXmm.csv` (ex. `input_file_10mm.csv`, `input_file_50mm.csv`) where each file consists of one column called Pressure (the pressure cumulative from the no-leak situation) and other columns assessed in each leak situation (LN1-LN32). These CSV files will serve as input to Step 2 for pressure deviation estimation and subsequent threshold considerations. The entire line of all hydraulic input files for both leak sizes can be found in Appendix 10.7

This subsequent component will justify that the OSP protocol was applied to show where such hybrid optimisation framework can be so detection accurate with such limited redundancy, yet, at

the same time, so tolerant to changes as an Intelligent Water Network would experience in the field.

7.3.2 Data Processing and Filtering

Stage 2 continues with the hydraulic simulations to create a systematic deviation matrix to assess how every node reacts with respect to one another to each leak occurrence. A second Python script (see Appendix 10.7) processes the resultant pressure data of all input_file_XXmm.csv files that parse all thirty-two columns of leaks against the baseline.

In fact, it processes the data scenario by scenario by using four derivative features.

1. Pressure deviation is the value subtracted from baseline pressure
2. Percentage deviation is the normalised value as in most cases, the baseline pressure will inflate a notable pressure deviation.
3. Threshold flag is 1/0 claims whether the value in absolute terms has a response to one deviation greater than the detection threshold of 0.5 m.
4. Filtered deviation is a zeroed-out deviation with no extremes where any value at or below 0.5 m becomes a zero so that only essentially significant deviations to hydraulics matter.

All variables are integrated into the existing dataset which structure means that the behaviour of every leak incident is added. Aggregate columns Filtered_Delta_LNxx (see Appendix 10.7) produces what will henceforth be known as the detection or deviation matrix where each column represents the threshold-detected output of a node to a single leak incident, and where each row represents a single incident in the simulation.

The processed output is named processed_pressure_data_XXmm.csv (see Appendix 10.7). This file will become the main input for stage 3 to conduct the sensitivity and subsequent optimisation.

7.3.3 Node Sensitivity Computation

Following sensitivity screening, five algorithms were implemented to identify the optimal subset of nodes for leak detection. Each algorithm represents a distinct philosophy of optimisation, allowing performance comparisons across heuristic, statistical, and hybrid approaches.

In reference to the leak-scenario assessment, nodal pressure sensitivities were generated in Excel

based on 4 sizes of leak relative to each other: 10 mm, 50 mm, 100 mm and 150 mm. For each, the discharge coefficient of leak used in the spreadsheet was used (0.000216591 for 10 mm, 5.39226×10^{-3} for 50 mm, 2.15691×10^{-2} for 100 mm and 4.85304×10^{-2} for 150 mm) as the applied value at each candidate node individually (beginning with the nominated demands and pressures from the "no-leak" state). Relative to that candidate node, the network was re-simulated with the assumption of a leak and the nodal sensitivity value was determined as the resultant pressure change induced by the leak relative to the applied leak flow in question. Ultimately, these nodal sensitivities were correlated to generate aggregate sensitivity indexes (sum, mean and max) for each node which were subsequently ranked according to their average pressures (leak outflows corresponding to each leak size). The nodes with the highest sensitivity rankings (aggregated) were then used to create optimal sensor placement relative to each of the 4 diameters of leak. (See Appendix 10.7) The layout of the Hanoi Water Distribution Network, including node and link numbering used in this study, is shown in Figure 7.10.

7.3.3.1 10 mm Leak Scenario – Sensitivity and Deviation Analysis:

In addition to the processed deviation matrix and node sensitivity analysis report, a supplementary python script was created to assess the sensitivity characteristics and create diagnostic plots and summary statistics for the 10 mm leak scenario (See Appendix 10.7 for node_sensitivity_10mm.py). For example, it determines ranking all nodes by their sensitivity value to measure which five nodes are most sensitive (in response).

The node most sensitive for the 10 mm case are Node 29, Node 30, Node 28, Node 27 and Node 15 with statistically similar average filtered deviations (see Appendix 10.7 for sensitivity_analysis_report_10mm.txt).

Ranked sensitivity values are created in bar charts (note that there is obvious sensitivity of nodes 29-30 but then their strong adjacent nodes); histograms show that nearly all nodes have sensitivity values clustered to a lower degree of deviation except a few with relatively high response and none with responses near zero. Box plots and violin plots of filtered deviations (i.e. changes in deviation over time) of the first five nodes summarise the nature of distribution and direction of change, and a heat map of filtered deviations suggests when the greatest deviations occur attributable to leak-induced pressure changes over time (See Appendix 10.7 for node_sensitivity_10mm.py).

The python script (see Appendix 10.7) assesses quartiles of the sensitivity distribution and generates a correlation matrix of the filtered deviation signals from the top nodes vs. all others as well. For example, the quartiles for the 10 mm case are calculated very closely (upper and lower quartiles are statistically similar) indicating an even distribution of what constitutes sensitivity meanwhile the correlation matrix highlights nearly perfect correlation amongst the top nodes themselves filtered deviations which indicate that these junctions have a similar hydraulic response to the induced leaks in the network. Ultimately, the most sensitive nodes from this component require assessment of the percent of time steps that produced a non-zero filtered deviation value; for instance, nodes 29-30 experienced a non-zero filtered deviation value in 96.8% of occurrences which means that they respond sufficiently as they respond consistently for nearly every leak event simulated (See Appendix 10.7 for sensitivity_analysis_report_10mm.txt).

While not explicitly incorporated in the optimisation algorithms, these diagnostic numerical and visual assessments further assess hydraulic behaviour for the network and help connect why certain junctions come to play critical roles as sensor placement suggestions later discussed.

The below Figure 7.2 summarises the hydraulic sensitivity patterns across all nodes for the 10 mm leak simulations. The upper-left panel shows the ten nodes with the highest sensitivity values, based on the average magnitude of their threshold-filtered pressure deviations. The upper-right panel displays the distribution of sensitivity values across all nodes, indicating a strong concentration of nodes with high responsiveness and only a few with near-zero influence. The middle panel contains box plots of filtered deviation values for the five most sensitive nodes, illustrating the spread, skewness, and extreme deviations across all leak events. The bottom panel presents a heat map of the filtered deviation magnitudes for these top five nodes over time, revealing the temporal phases during which leak-driven pressure disturbances are most pronounced.

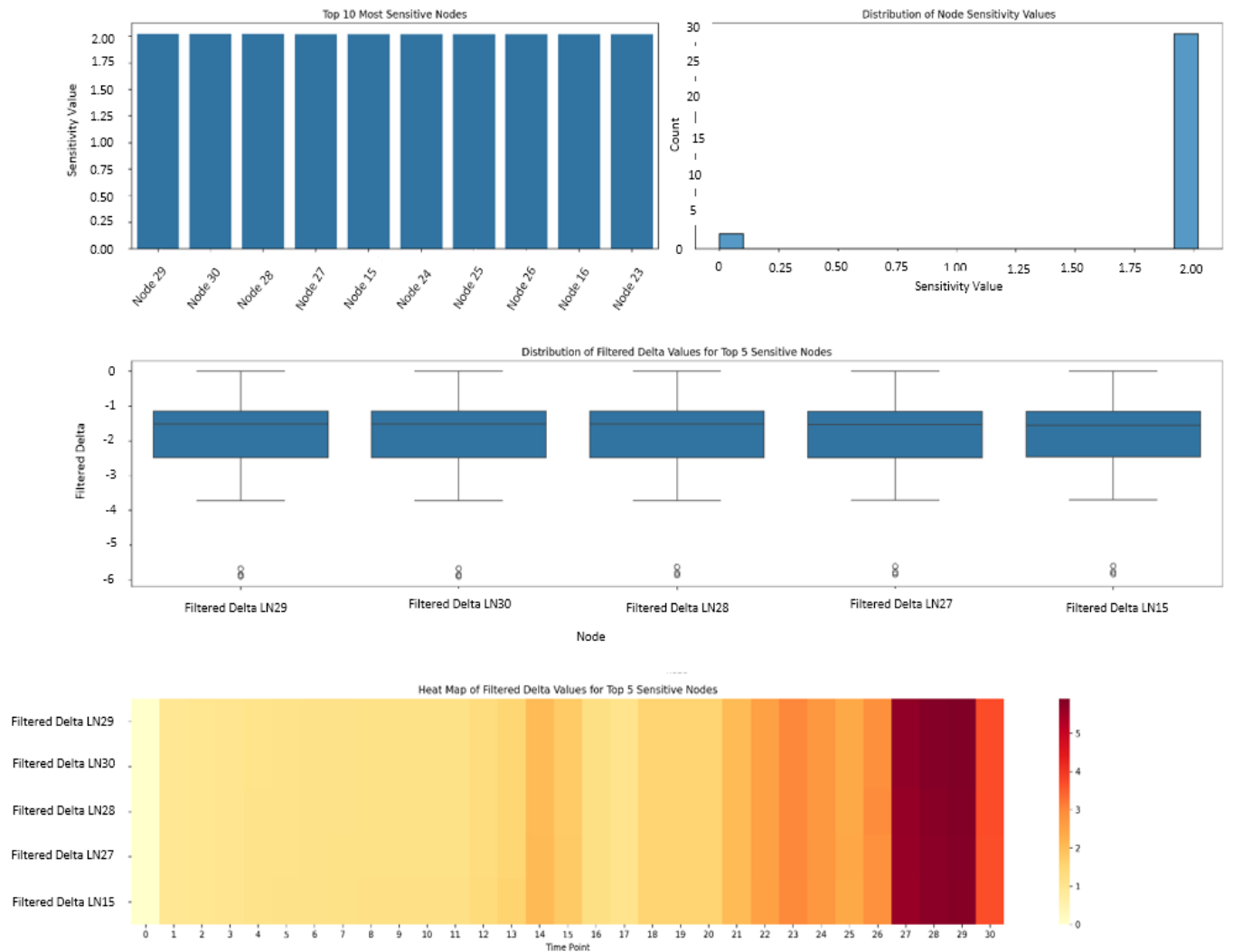


Figure 7.2: This Figure summarises the hydraulic sensitivity patterns across all nodes for the 10 mm leak simulations.

The following Figure 7.3 provides complementary visual diagnostics for understanding the behaviour of the highest-sensitivity nodes in the 10 mm leak dataset.

The upper panel displays violin plots showing the full distribution of threshold-filtered pressure deviations for the top five most sensitive nodes, highlighting both shape and density across all leak scenarios.

The lower panel shows the correlation matrix for the top ten sensitive nodes. The near-uniform correlation values indicate that these nodes exhibit almost identical deviation patterns, suggesting that they respond in a highly coherent manner to leak events throughout the network. This strong correlation explains why several of these nodes appear repeatedly in the top-ranked sensitivity list.

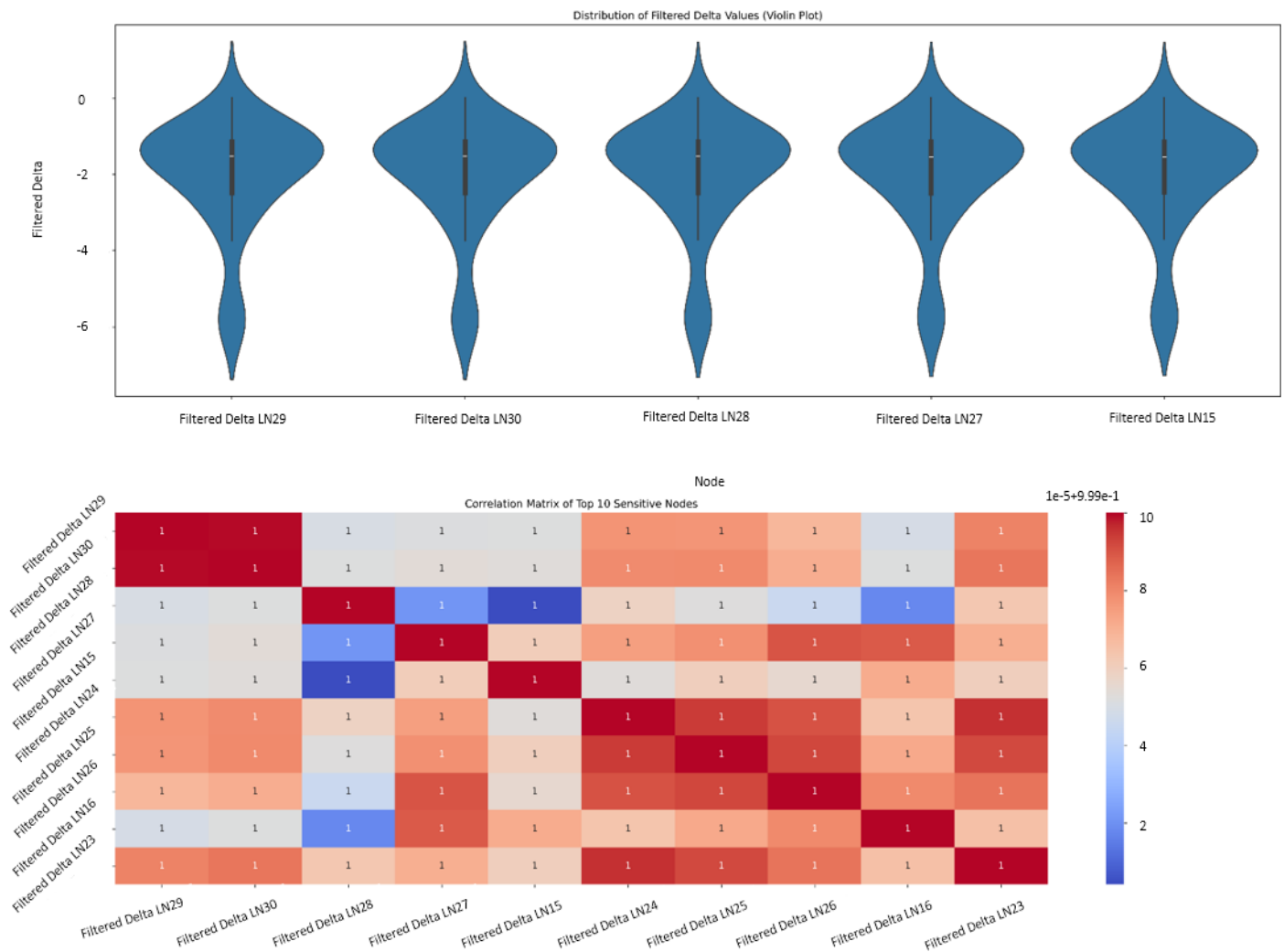


Figure 7.3: Complementary visual diagnostics for understanding the behaviour of the highest-sensitivity nodes in the 10 mm leak dataset.

The processed_pressure_data_10mm.csv file (see Appendix 10.7) contains all pressure deviations for Stage 2 of the 10 mm leak analysis. This is a processed version of the input_file_10mm.csv (the aforementioned file of hydraulic simulation original readings) made by deriving deviations from standard pressures from the pressures relative to each determined leak and then applying a threshold value of 0.5 metres to remove any nonsensical pressures that are not hydraulically significant. For each leak (LN2 through LN32), four new variables were created: the absolute pressure deviation, the percent deviation from standard pressure, a binary qualifier for the threshold, and the threshold-applied deviation, which was subsequently used in all further

analyses. These were joined to the original dataset to create an updated matrix of deviations, of which each respective Filtered_Delta_LNxx column represents the resultant whittled response of the pressure response from the leak in question. It should be noted that this will be used as a threshold value for all Stage 3 (node sensitivity analysis) and subsequently for Stage 4 (sensor placement analysis) to ensure that all further investigations are based on hydraulically significant and consistently filtered signals. The node_sensitivity_analysis_10mm.csv (see Appendix 10.7) contains the node-level sensitivity values computed in Stage 3 for the 10 mm leak dataset. Table 7.1 shows the ranked node sensitivity values. These values were obtained by analysing the threshold-filtered deviation signals (Filtered_Delta_LNxx) generated in Stage 2, which preserve only those pressure changes exceeding the 0.5-metre detection threshold. For each node, the script calculates the average magnitude of its filtered deviations across all simulated leak events, producing a single metric that reflects how strongly and consistently that node responds to leaks throughout the network. Nodes with higher values in this file exhibit greater hydraulic responsiveness and are therefore more informative candidates for pressure-sensor placement. The resulting sensitivity table not only provides a complete ranking of nodes from most to least sensitive but also serves as an essential input to the optimisation algorithms in Stage 5, particularly the Top-K method and the comparative benchmarking analyses. This file forms the quantitative foundation for interpreting hydraulic influence patterns within the 10 mm leak scenario. It should be noted that the sensitivity values of the top-ranked nodes are very similar, indicating that they form a cluster of highly responsive locations. Therefore, the ranking should be interpreted as indicative rather than absolute, with minimal practical difference between the top candidates. The sensitivity values are low due to small leaks in comparison to large flows in the pipelines.

Table 7.1: Ranked Node Sensitivity Values for the 10 mm Leak Scenario

Node	Sensitivity_Value
Node_2	1.989032
Node_3	2.010645
Node_4	2.010968
Node_5	2.013226
Node_6	2.014194
Node_7	2.015161
Node_8	2.015806
Node_9	2.015806
Node_10	2.015806
Node_11	2.016129
Node_12	2.014516
Node_13	2.013871
Node_14	2.015806
Node_15	2.017419
Node_16	2.016774
Node_17	2.015806
Node_18	2.013226
Node_19	2.012258
Node_20	2.013548
Node_21	2.014839
Node_22	2.015484
Node_23	2.016129
Node_24	2.017419
Node_25	2.017419
Node_26	2.017419
Node_27	2.017419
Node_28	2.019032
Node_29	2.020323
Node_30	2.02
Node_31	0
Node_32	0

The file top_5_sensitive_nodes_summary_10mm.csv (see Appendix 10.7) contains a detailed statistical profile of the five most sensitive nodes identified in the 10 mm leak analysis. Table below 7.2 shows the summary statistics for the five most sensitive nodes for the 10mm leak scenario. These nodes were selected based on the highest values in node_sensitivity_analysis_10mm.csv, and their behaviour was examined more closely to understand the characteristics of their pressure-deviation responses. For each of the top five nodes,

the summary table reports the average filtered deviation, the number of time steps where a significant deviation occurred, and the maximum, mean, and standard deviation of their threshold-filtered pressure changes. These metrics provide deeper insight into both the magnitude and consistency of each node's hydraulic response. While Stage 3 focuses on producing a single sensitivity value per node, this summary table reveals why these nodes rank highly: they show strong, frequent, and stable deviations across almost all simulated leak events. This information supports the interpretation of the sensitivity plots and helps validate the dominant role of these nodes in subsequent optimisation stages.

Table 7.2: Summary Statistics for the Five Most Sensitive Nodes (10 mm Leak Scenario)

Node	Sensitivity_Value	Non_Zero_Count	Max_Delta	Mean_Delta	Std_Delta
Node_29	2.020323	30	5.9	2.020323	1.471717
Node_30	2.02	30	5.9	2.02	1.470889
Node_28	2.019032	30	5.86	2.019032	1.459202
Node_27	2.017419	30	5.84	2.017419	1.45201
Node_15	2.017419	30	5.83	2.017419	1.444936

The document `sensitivity_analysis_report_10mm.txt` consists of a detailed statistical summary relative to the network sensitivity response to the calibrated leak of 10 mm from the simulation. This report combines the numerical diagnostics resulting from the Stage 3 study: the quartile statistics of the node-sensitive distribution, the correlation matrix of the filtered deviation signals relative to the more sensitive nodes and the percentage of time steps where pressure deviation values were noted. The quartile values provide a succinct characterisation of the spread and mean positioning of the sensitivity distribution relative to the fact that maximum sensitivity is achieved and consistently maintained by most nodes given leak-related influences. The correlation matrix determines the level of similarity of deviation increments amongst more sensitive nodes as determined by average response. This is validated by the significant coherence established amongst the most sensitive deviations, whereas a complete assessment of all node deviation sensitivity values is provided so that all determined relevant nodes may be transparently determined as hydraulically influenced throughout the network. Ultimately, these measures work to validate visual assessments determined prior and numerically confirm which nodes could function as optimal sensors.

Table 7.3 presents the five most sensitive nodes identified in the 10 mm leak analysis, based on the values contained in node_sensitivity_analysis_10mm.csv. These nodes exhibit the largest average threshold-filtered pressure deviations across all simulated leak scenarios and therefore represent the strongest hydraulic responders in the network. Their high sensitivity values explain why they consistently emerge as leading candidates for early detection and sensor placement in later optimisation stages. The complete ranked list of sensitivity values for all nodes is provided in the Appendix for reference (see Appendix 10.7).

Table 7.3: Top 5 Most Sensitive Nodes (10mm)

Node	Sensitivity Value
Node_29	2.020323
Node_30	2.020000
Node_28	2.019032
Node_27	2.017419
Node_15	2.017419

Table 7.4 summarises the quartile statistics of the sensitivity distribution for the 10 mm leak dataset, as extracted from sensitivity_analysis_report_10mm. These values describe the central tendency and spread of node sensitivities across the network, showing that most nodes cluster tightly around similar response levels. The small difference between the lower quartile, median and upper quartile indicates a highly uniform sensitivity pattern, with only a few outlier nodes exhibiting substantially lower responsiveness.

Table 7.4: Sensitivity Value Quartiles (10 mm)

Statistic	Value
25th Percentile	2.013387
Median (50th Percentile)	2.015806
75th Percentile	2.017097

Table 7.5 represents the correlation coefficients of the five most redundant nodes since their filtered deviation signals are highly correlated based on the correlation assessment from sensitivity_analysis_report_10mm. These correlation coefficients (all at or above 0.999) are important because these nodes deviate to leaks in the same redundant fashion which means that from a hydraulic perspective, the nodes flow patterns operate just constitutionively similar enough

to one another that they're constitutionively operated the same. This redundancy is relevant for sensor placement since these redundant nodes usually represent a certain cluster that will give/take the same information and give them the same/higher urgency to act.

Table 7.5: Correlation Matrix for Top 5 Nodes (10 mm)

	LN29	LN30	LN28	LN27	LN15
Filtered_Delta_LN29	1.000000	0.999999	0.999950	0.999952	0.999952
Filtered_Delta_LN30	0.999999	1.000000	0.999952	0.999954	0.999953
Filtered_Delta_LN28	0.999950	0.999952	1.000000	0.999921	0.999904
Filtered_Delta_LN27	0.999952	0.999954	0.999921	1.000000	0.999961
Filtered_Delta_LN15	0.999952	0.999953	0.999904	0.999961	1.000000

Table 7.6 shows the percentage of time steps with pressure deviation, relative to threshold-filtered signals of the 10 mm leak, for each of the five nodes with greatest deviation from mean pressures. Approximately 96.8% of all simulated occurrences, all five nodes had major pressure deviation, meaning these areas are not only responsive but also reliably consistent in response and response validity to leak occurrences. Such validation further supports their ranking within the sensitivity analysis and is another reason why they were selected for further optimisation efforts. More in-depth sensitivity analysis results from which these five nodes were selected can be found in the Appendix (see appendix 10.7).

Table 7.6: Percentage of Significant Filtered Deviations (Top 5 Nodes, 10 mm)

Node	% Significant Readings
Filtered_Delta_LN29	96.774194%
Filtered_Delta_LN30	96.774194%
Filtered_Delta_LN28	96.774194%
Filtered_Delta_LN27	96.774194%
Filtered_Delta_LN15	96.774194%

7.3.3.2 50 mm Leak Scenario – Sensitivity and Deviation Analysis:

For the 50 mm leak scenario, the same ancillary Python processing and diagnosis applied to the 10 mm scenario was applied to the ancillary input file *input_file_50mm.csv* in order to render the processed deviation matrix, node-sensitivity table and various diagnostic plots (see *node_sensitivity_50mm.py* located in the Appendix). The script orders all sensors per sensitivity value and notes the five highest-reactionary spaces based on average threshold-filtered pressure

deviation. For the 50 mm leaks, the highest sensitive nodes are Node 27, Node 16, Node 20, Node 14 and Node 15, which are closely ordered relevant to their magnitudes of sensitivity and thus form the dominant group of hydraulically relevant nodes in this case.

Similarly, to the 10 mm findings, bar plots of ranked sensitivity values are provided for easy visualisation of ranking along with histograms of the sensor response spread per node. For the case of 50 mm, the histogram shows that most nodes fall into the decent-moderate reaction to response range and none falls into the insignificant reaction category. Box plots of filtered deviations for the top five nodes show spread and skew relative to the deviation compared to all iterations, and a heat map shows when these deviations are increased. This helps to determine when, relative to day, week or month the 50 mm leaks create the most defining intrusively hydraulic patterns and why. These come from the deviation matrix and also, a processed version is located in *processed_pressure_data_50mm.csv*, which uses the same four deviation fields that are processed from the first 50 mm csv (absolute deviation, percentage deviation, threshold trigger and threshold-filter deviation for relative 1-minute deviations for each simulated case).

In addition, from the response thresholds, quartiles for the levels of response are calculated and a correlation matrix is derived from similarly-sourced filtered deviations of the top-ranked nodes, including what percentage of timesteps per node possess a non-zero threshold-filter deviation. For example, in the 50 mm findings, the lower quartile is 0.571, median 0.595, upper quartile 0.621; this means that half of all nodes are less sensitive at a moderate level but only a quarter are more sensitive at a respectable level. The correlation matrix shows that Nodes 27 and 16 are highly correlated per similar effective deviations while Nodes 14 and 15 make up a closely correlated pair as well however Node 20 possesses a more isolated correlation relative not only to lower sensitivity but some slightly negative correlations. This means that Node 20 senses a unique characteristic of what occurs relative to hydraulic data compared to the other higher sensitive nodes meaning that if one were to incorporate this aspect into a developed leak sensing system around here, it would also stand alone with informational value from Node 20. Finally, the percentage of significant readings validates that Nodes 27, 16, 14 and 15 provide significant readings and that Node 20 is significant in nearly all iterations (96.8% meaningful) reinforcing its consistent classification as a telling area.

Thus, while these diagnostic plots and statistically relevant findings are not integrally processed

into the optimisation algorithms for hydraulic controls in sensor construction, they more than help an integral assessment of determining how sensitive (or not) this network is based on varying levels of response based upon both leak values. They support the justification as to why certain nodes are repeatedly observed as top candidates for location selection in addition to providing easy numerical and visual assessments to make sense of corroborating findings which bolster their pressure-deviation configurations in power relative to each other and in distinctiveness relative to all assessed pressure deviation readings over time. The full ranked sensitivity value list for 50 mm sensors, the summary statistics for the top-five sensors and the full statistics report can be found in the Appendix. The below Figure 7.4 summarises the hydraulic sensitivity patterns across all nodes for the 50 mm leak simulations and the below Figure 7.5 provides complementary visual diagnostics for understanding the behaviour of the highest-sensitivity nodes in the 50 mm leak dataset.

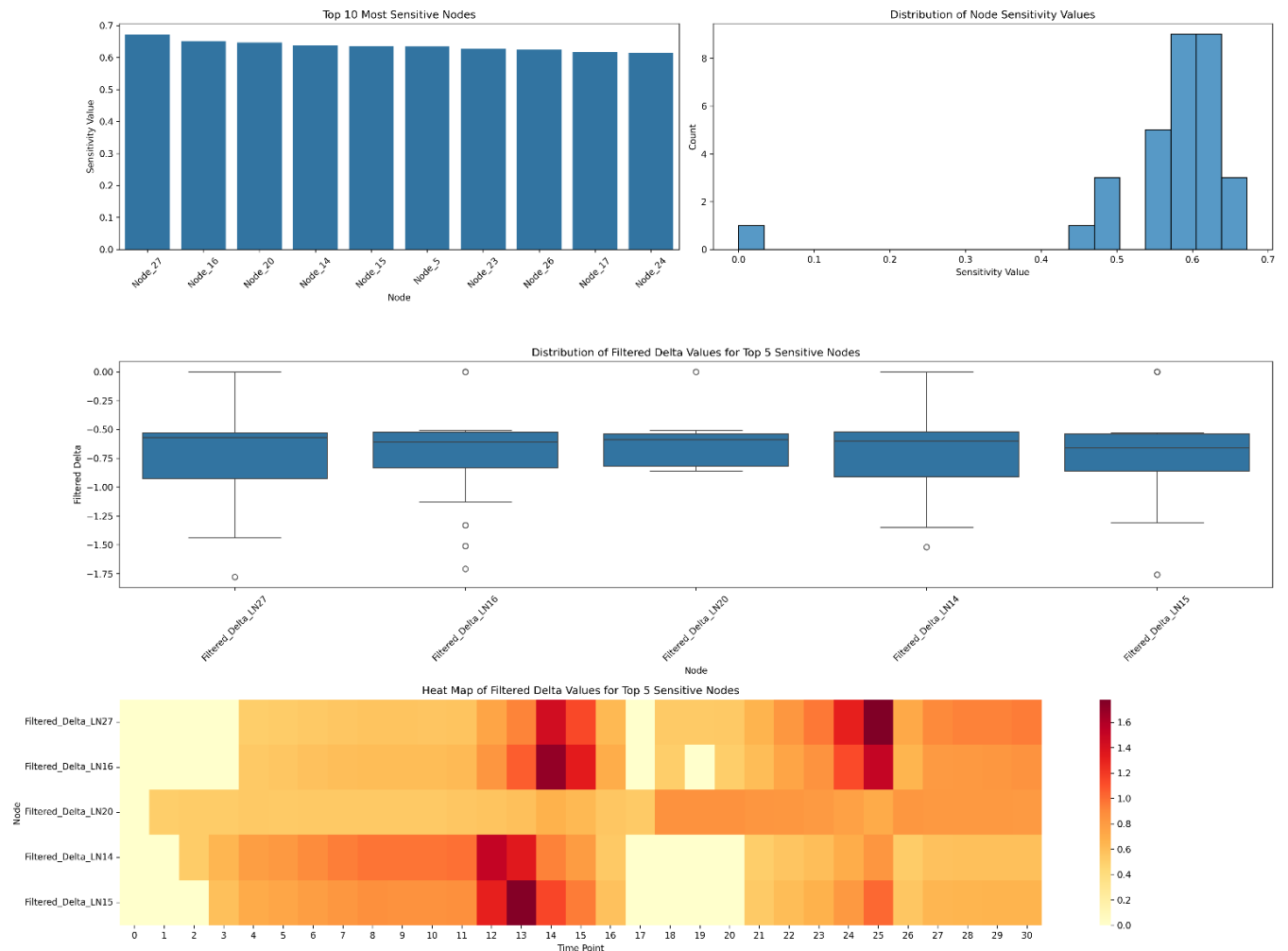


Figure 7.4: Hydraulic sensitivity patterns across all nodes for the 50 mm leak simulations.

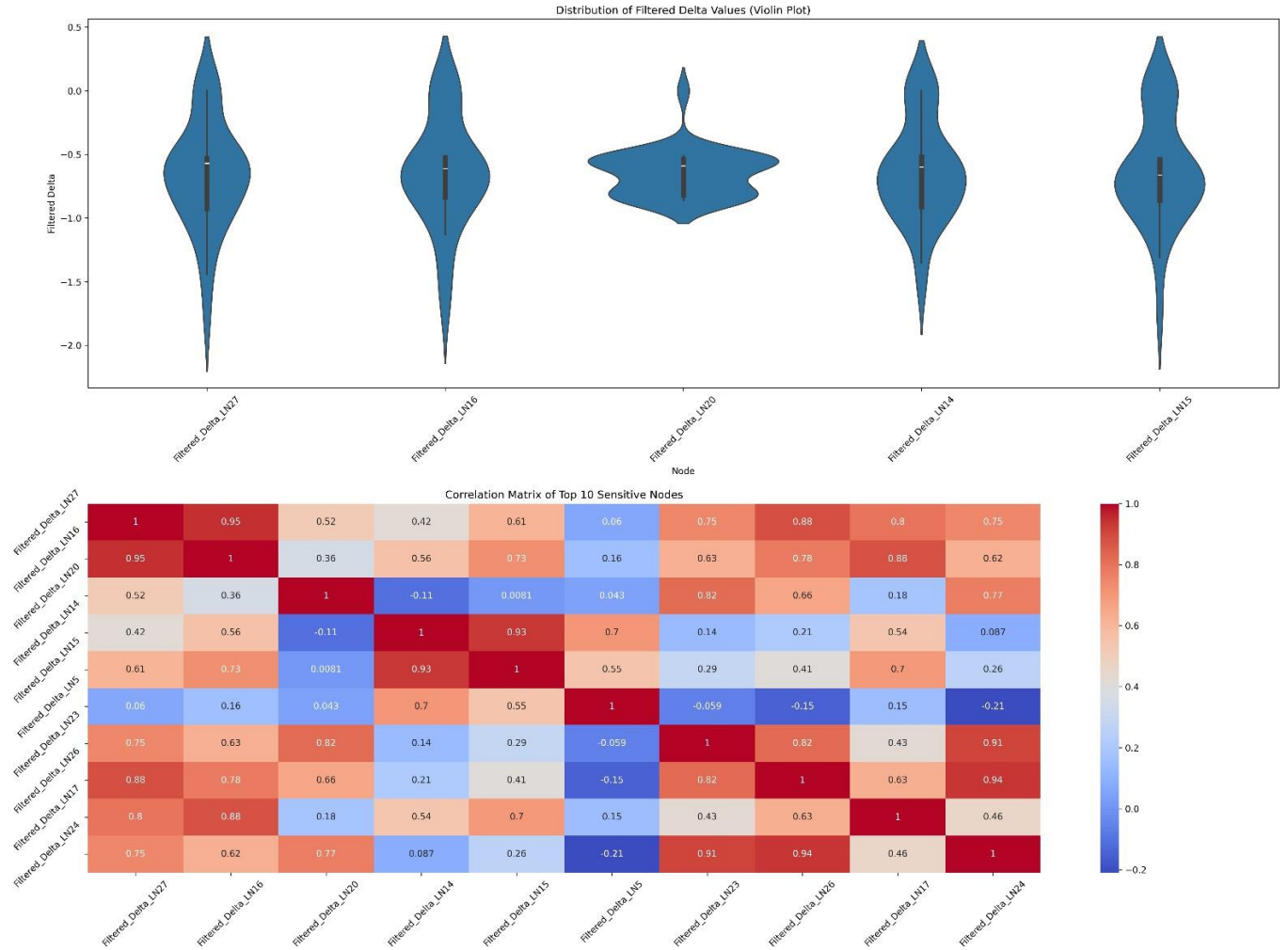


Figure 7.5: Complementary visual diagnostics for understanding the behaviour of the highest-sensitivity nodes in the 50 mm leak dataset.

The *node_sensitivity_analysis_50mm.csv* contains node sensitivity values that were outputted in Stage 3 relative to 50 mm leak data. Table 7.7 shows the ranked node sensitivity values. This is because these values were outputted from the threshold-filtered deviation signals (Filtered_Delta_LNxx) generated in Stage 2 for the 50 mm case, all based on the same detection threshold (0.5 m), as indicated before. Thus, the script iterates through nodes and obtains the mean magnitude of a specific node's filtered deviations over the entire 50 mm leak scenarios. Thus, a magnitude value means that sensitivity of a node to activation under 50 mm leaks is consistent across the entire network. Therefore, if these nodes are planned to be used for positioning where the pressure sensors will be installed, those nodes with high sensitivity values will be more valuable and informative. Below is the Table compiled of all the nodes with their sensitivity ranked

from highest to lowest; only certain cutoffs of the results are seen here (largest to smallest) but the full results and code are found in the Appendix 10.7 for due diligence.

Table 7.7: Ranked Node Sensitivity Values for the 50 mm Leak Scenario

Node	Sensitivity Value
Node_2	0
Node_3	0.552903
Node_4	0.595484
Node_5	0.634516
Node_6	0.584516
Node_7	0.591935
Node_8	0.570968
Node_9	0.577097
Node_10	0.582258
Node_11	0.570968
Node_12	0.477742
Node_13	0.45871
Node_14	0.637097
Node_15	0.634839
Node_16	0.65129
Node_17	0.617419
Node_18	0.602581
Node_19	0.609677
Node_20	0.646774
Node_21	0.48
Node_22	0.495806
Node_23	0.628387
Node_24	0.615161
Node_25	0.570323
Node_26	0.625484
Node_27	0.672258
Node_28	0.580968
Node_29	0.613871
Node_30	0.604194
Node_31	0.60129
Node_32	0.57

The *top_5_sensitive_nodes_summary_50mm.csv* file is a statistical summary of the top five sensitive nodes that were found from the previously documented leak investigation of 50 mm. Table 7.8 shows the Summary Statistics for the Five Most Sensitive Nodes.

These nodes are those that represent the top five of node deviation sensitivity from the *node_sensitivity_analysis_50mm.csv* file and have also been studied in terms of their pressure-deviation response. For each of these nodes, the summary includes the averaged filtered deviation (the sensitivity value), a frequency count of the filtered deviations present in non-zero time steps, and ultimately the max, mean and standard deviation of threshold filtered pressure deviations. When assessed in combination with one another, these statistics denote not only the severity of the deviations, but their prevalence and consistency, further substantiating Nodes 27, 16, 20, 14 and 15 as the five of most interest within the 50 mm leak investigation and allowing for future use in conjunction with optimisation and other comparative studies.

Table 7.8: Summary Statistics for the Five Most Sensitive Nodes (50 mm Leak Scenario)

Node	Sensitivity_Value	Non_Zero_Count	Max_Delta	Mean_Delta	Std_Delta
Node_27	0.672258	26	1.78	0.672258	0.421376
Node_16	0.65129	25	1.71	0.65129	0.430138
Node_20	0.646774	30	0.86	0.646774	0.182289
Node_14	0.637097	25	1.52	0.637097	0.393304
Node_15	0.634839	24	1.76	0.634839	0.424569

Table 7.9 lists the five highest sensitive nodes for the 50 mm leak event based upon *sensitivity_analysis_report_50mm.txt*. This is more of a reference guide and the report is more of a summary, as all that's needed is the node number and the corresponding sensitivity value, as these are the highest hydraulic responders in the network. Therefore, these values are consistent with what's been presented here and relative to the other values, establish that Nodes 27, 16, 20, 14 and 15 are the main, high-sensitivity aggregate for the 50 mm leak event. The Appendix 10.7 shows the complete ranked version for all nodes.

Table 7.9: Top 5 Most sensitive nodes (50mm Leak Senario)

Node	Sensitivity Value
Node_27	0.672258
Node_16	0.651290
Node_20	0.646774
Node_14	0.637097
Node_15	0.634839

Table 7.10 shows the quartile statistics of the node-sensitivity distribution for the 50 mm leak study. This is from sensitivity_analysis_report_50mm.txt. These quartile statistics try to account for an average and spread of nodes across a sensibility network. The lower quartile, median and upper quartile values all line up fairly closely meaning most nodes respond within a confined band of sensitivity. In fact, very few operate well below this range. This means a generally good responding network with a few nodes that are extremely powerful (or extremely not so). The distribution and histogram are in the Appendix 10.7.

Table 7.10: Sensitivity Value Quartiles (50 mm)

Statistic	Value
25th Percentile	0.570645
Median (50th Percentile)	0.595484
75th Percentile	0.621452

The correlation values corresponding to the filtered deviation signals of the five most sensitive nodes in the 50 mm leak scenario are presented in Table 7.11. These correlation coefficients were calculated based on the threshold-filtered deviation time series and are available in sensitivity_analysis_report_50mm.txt. Relative to one another, Nodes 27 and 16 are highly correlated, as are Nodes 14 and 15. However, Node 20 is only weakly and even slightly negatively correlated with some of the others. Thus, it seems that Node 20 captures a slightly different portion of the leak response than the other four, which operate somewhat in concert. This is important for sensor placement because some sensors capture redundant information while others capture signals that are slightly different.

Table 7.11: Correlation Matrix for Top 5 Nodes (50mm)

	LN27	LN16	LN20	LN14	LN15
Filtered_Delta_LN27	1.000000	0.945437	0.517203	0.421854	0.610396
Filtered_Delta_LN16	0.945437	1.000000	0.362172	0.556706	0.727419
Filtered_Delta_LN20	0.517203	0.362172	1.000000	-0.109162	0.008090
Filtered_Delta_LN14	0.421854	0.556706	-0.109162	1.000000	0.931050
Filtered_Delta_LN15	0.610396	0.727419	0.008090	0.931050	1.000000

Table 7.12 presents the percentage of time steps in which each of the five top sensitivity nodes had a non-zero threshold-filtered deviation for the 50 mm leak condition. These percentages from

sensitivity_analysis_report_50mm.txt seek to measure how often each node consistently responds to leak-driven deviations across the entire time series response. For example, it is important to note that Node 20 had non-zero threshold deviations for basically all-time steps (96.8%). Similarly, Nodes 27, 16, 14 and 15 also represent nodes with high responsiveness with non-zero thresholds occurring in more than three-quarters of time steps. Thus, it can be determined that these nodes are not only highly hydraulically sensitive but also frequently active across multiple leak occurrence scenarios.

Table 7.12: Percentage of Significant Filtered Deviations (Top 5 Nodes, 50 mm)

Node	% Significant Readings
Filtered_Delta_LN27	83.870968%
Filtered_Delta_LN16	80.645161%
Filtered_Delta_LN20	96.774194%
Filtered_Delta_LN14	80.645161%
Filtered_Delta_LN15	77.419355%

7.3.3.3 100 mm Leak Scenario – Sensitivity and Deviation Analysis:

The 100 mm leak scenario is a medium level hydraulic perturbation that causes larger, more expansive deviations than the 10 mm and 50 mm scenarios that came before it. Likewise, for this scenario, the dedicated Python script *node_sensitivity_100mm.py* was used to determine deviation matrices, node sensitivity and ultimately, threshold filtering, diagnostic plots and statistical summary (See Appendix 10.7).

Processed_pressure_data_100mm.csv is the threshold-filtered deviation dataset for the 100 mm leak simulations. This table is generated by taking the leak pressure at every second and subtracting the baseline pressure from it, applying the significance threshold of 0.5 m after. Like all previously processed datasets, this one contains four created variables per 100 mm leak scenario: deviation, percent deviation, binary threshold flag, and threshold deviation value. All processed variables are utilised in all Stage 3 sensitivity calculations and Stage 5 sensor placement confirmation. The full

table is found in Appendix 10.7. The node_sensitivity_analysis_100mm.csv file represents the average threshold deviation magnitude for the nodes. The threshold-filtered values are representative of sensitivity to a 100 mm leak anywhere in the system. Where, relative to the 10 mm and 50 mm, sensitivity is more clustered together. The 100 mm dataset allows for a clearer separation between more sensitive and less sensitive nodes. Therefore, in accordance with the calculations, the sensitive nodes for the 100 mm analysis are Node 29, Node 30, Node 31, Node 28, Node 32. These values are taken as is from the findings of the sensitivity analysis (see Appendix 10.7) which documents the absolute average filtered deviations per node. The below Figure 7.6 summarises the hydraulic sensitivity patterns across all nodes for the 100 mm leak simulations and Figure 7.7 provides complementary visual diagnostics for understanding the behaviour of the highest-sensitivity nodes in the 100 mm leak dataset.

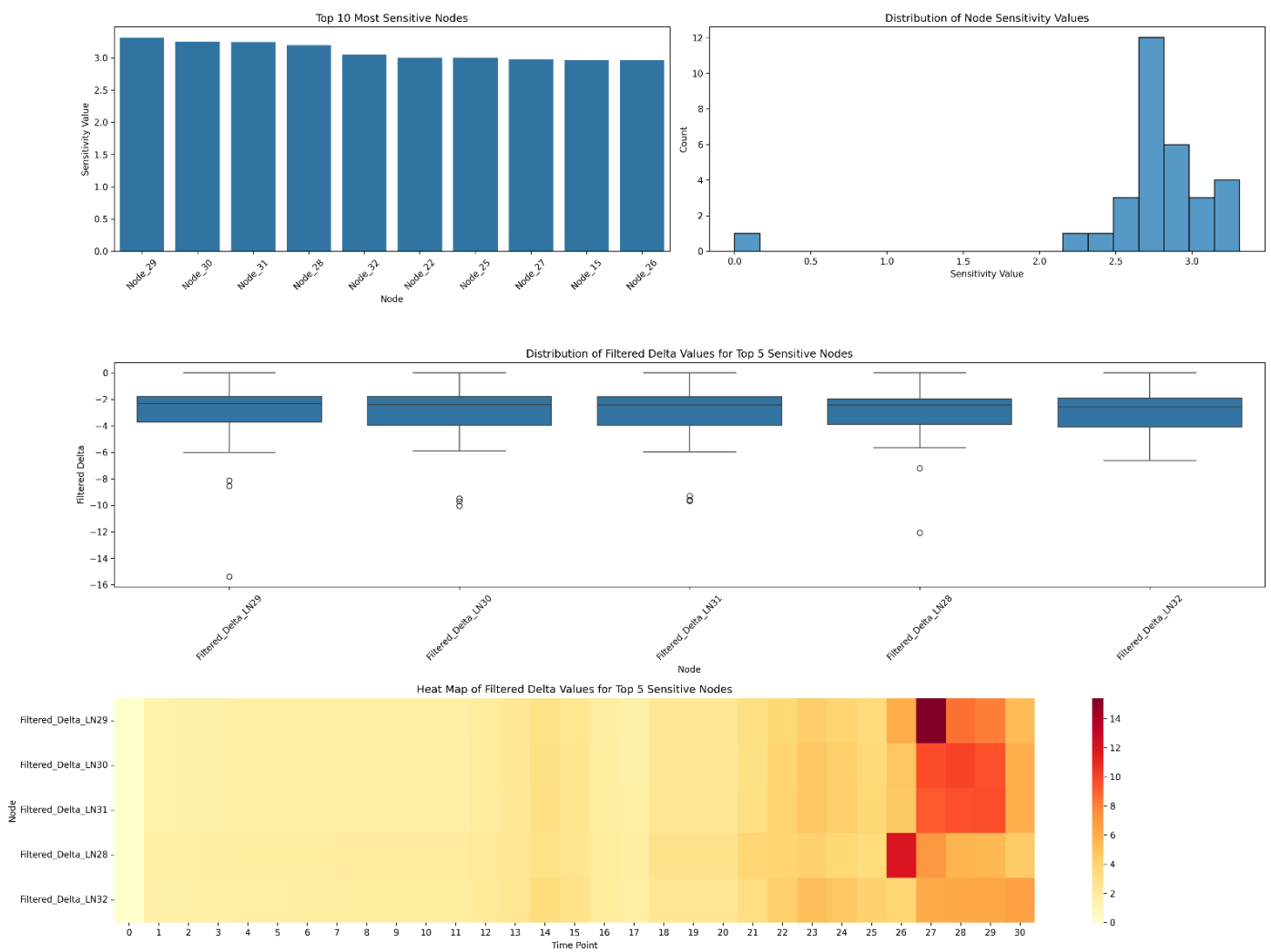


Figure 7.6: The hydraulic sensitivity patterns across all nodes for the 100 mm leak simulations.

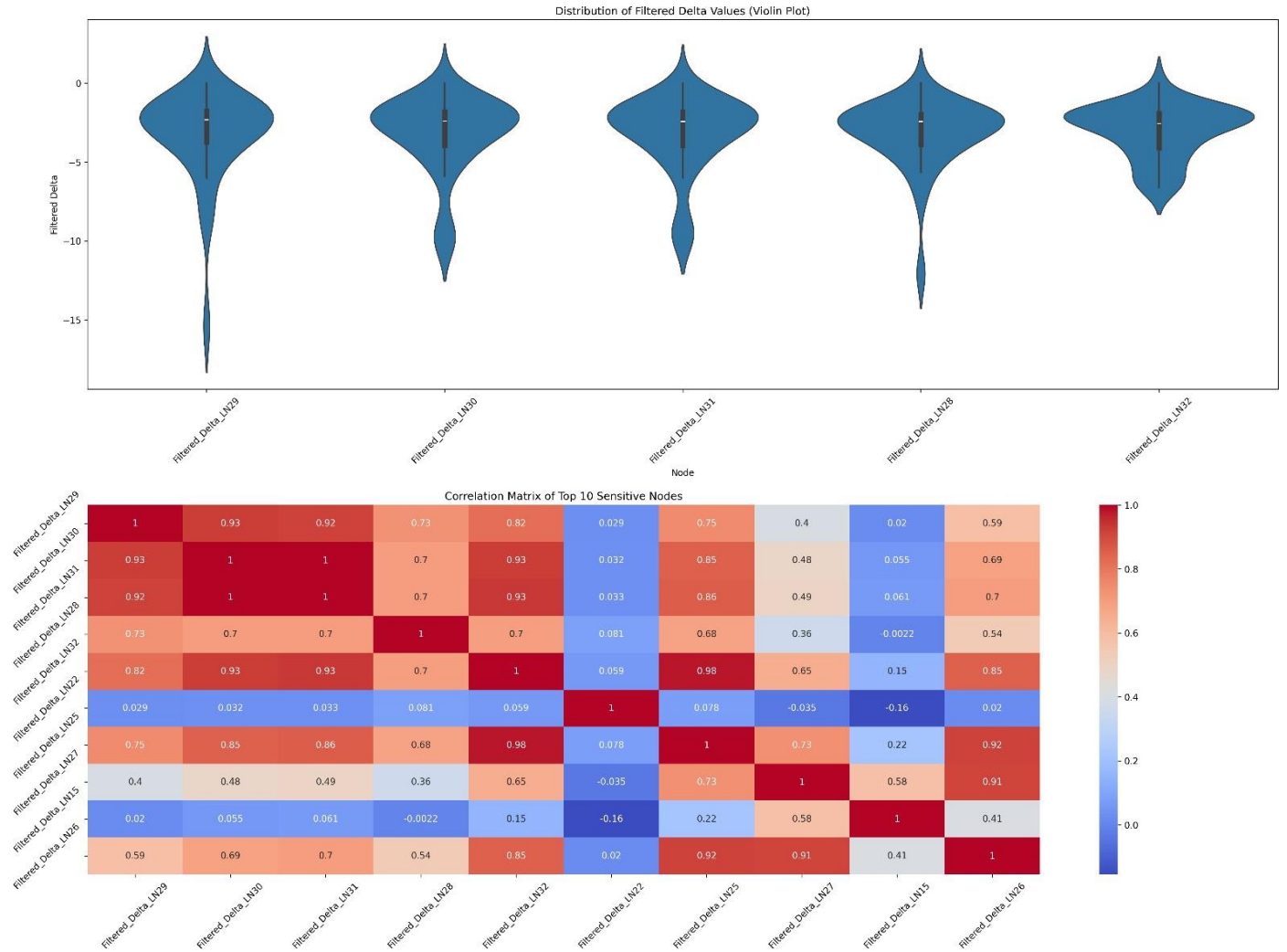


Figure 7.7: Complementary visual diagnostics for understanding the behaviour of the highest-sensitivity nodes in the 100 mm leak dataset.

The ranked node sensitivities table for the 100 mm leak location gathers all hydraulic responses observed of nodes exposed to larger magnitude leak disturbances. Table 7.13 shows the Ranked Node Sensitivity Values for the 100 mm Leak Scenario. Each sensitivity magnitude is derived from the average of a magnitude of signal from a corresponding node taken from the threshold-filtered pressure deviation signal from the deviation matrix created during Stage 2. Similar to the leakage distance sensitivity matrix, the nodes (and the catchment basins they surround) that lie on or adjacent to the primary transmission main are, once again, proven to be the most. The highest average response occurs at node 29, followed by nodes 30, 31, 28 and 32. The confidence intervals in these values are sufficiently above average to make a determination that those locations are hydraulically significant and also highly responsive to pressure differential changes caused by

leaks. Therefore, such rankings are important to compile to drive sensor siting considerations made in Stage 5 because more sensitive nodes provide better and more reliable detection of leaks and thus, higher analytical value.

Table 7.13: Ranked Node Sensitivity Values for the 100 mm Leak Scenario

Node	Sensitivity_Value
Node_2	0
Node_3	2.230645
Node_4	2.410645
Node_5	2.57
Node_6	2.692903
Node_7	2.716774
Node_8	2.755484
Node_9	2.78129
Node_10	2.797742
Node_11	2.808065
Node_12	2.804839
Node_13	2.743871
Node_14	2.897419
Node_15	2.967419
Node_16	2.956452
Node_17	2.791935
Node_18	2.665484
Node_19	2.490968
Node_20	2.614839
Node_21	2.793871
Node_22	3.002903
Node_23	2.806129
Node_24	2.936129
Node_25	3.002258
Node_26	2.960645
Node_27	2.974516
Node_28	3.195161
Node_29	3.312581
Node_30	3.253226
Node_31	3.244194
Node_32	3.044839

Table 7.14, which displays summary statistics for these five most sensitive nodes, indicates how many deviations occurred from the 100 mm leak and what filters and generates common deviation statistics for each of these nodes. Therefore, each number illustrates how many deviations filtered and surpassed the static threshold in practical application but also, how many deviations filtered for a max, mean and standard deviation of the filtered deviations. Such results help to justify why these five nodes are the most sensitive because not only are they some of the largest, statistically, mean, for this level of leak input, they also generate static and iterative responses across nearly all of the leak tests run. For example, from the 100 mm run alone, all five nodes have deviations that occur in excess of 96% of the time. This not only supports the validity of these nodes as operating at steady states, but also, that they are highly reliable meaning they operate at very similar values near steady states in very similar situations to each other. Such a nuanced level of information substantiates why these nodes become one of the most reliable candidates for future sensor implementation with this solution's implementation for optimisation.

Table 7.14: Summary Statistics for the Five Most Sensitive Nodes (100 mm Leak Scenario)

Node	Sensitivity_Value	Non_Zero_Count	Max_Delta	Mean_Delta	Std_Delta
Node_29	3.312581	30	15.39	3.312581	2.938799
Node_30	3.253226	30	10.04	3.253226	2.478251
Node_31	3.244194	30	9.67	3.244194	2.41745
Node_28	3.195161	30	12.08	3.195161	2.185226
Node_32	3.044839	30	6.62	3.044839	1.681923

Table 7.15 summarising statistics related to the five most sensitive nodes explains deviations relative to those nodes under the 100 mm leak. Each value represents the count of timesteps that a deviation (greater than a threshold) was recorded and then maximum, average and standard deviation of filtered deviations. Statistics relative to these nodes help explain the rationale for why these five nodes comprise a large portion of the sensitivity table: although they are sensitive because they have large average deviation values, they also consistently and frequently respond,

occurring at least once timesteps into nearly every simulated leak occurrence. For example, for the 100 mm leak, all five nominees for sensitivity have a deviation occurrence value greater than 96%, meaning they respond reliably and stably to deviations. This type of stable response pattern is part of the reason why these nodes ultimately emerge in optimised sensor location scenarios.

Table 7.15: Top 5 Most Sensitive Nodes (100mm)

Node	Sensitivity Value
Node_29	3.312581
Node_30	3.253226
Node_31	3.244194
Node_28	3.195161
Node_32	3.044839

The quartile statistics produced for the 100 mm case indicate the overall dispersion of sensitivities within the network. Table 7.16 is the Sensitivity Value Quartiles of 100mm. The interquartile range is relatively small (approximately 2.70 to 2.97), thus, most of the nodes in consideration have sensitivity values that are closely aligned. This is significant as the network is more attuned to the impact of such a large leak magnitude, for the leak location is sensitive across the network, whereas, with the 10 mm case, the variation in sensitivities was more subtle. Thus, a median sensitivity value of approximately 2.80 is representative of the average magnitude of hydraulic deviation a node would experience and, thus, a relative baseline when compared to highly sensitive nodes (29, 30, and 31). These quartile statistics add additional context to the ranking table and subsequent optimisation process by assessing average symmetry and dispersion relative to an average magnitude of hydraulic impact across a more extensive leak-induced hydraulic impact within the network.

Table 7.16: Sensitivity Value Quartiles (100 mm)

Statistic	Value
25th Percentile	2.704839
Median (50th Percentile)	2.804839
75th Percentile	2.970968

Where relative to the near leak condition's correlation matrix values that genuinely make up a perfect sense of sameness in filtered deviations signals, this comparison shows a jump to the 100mm leak condition shows a less similar sensation overall. However, mirroring the last example,

nodes 29, 30, and 31 still boast great pairwise correlations which reaffirm their relative responsiveness to a spatially spread out large leak. Yet with nodes 28 and 32 a greater-than sign slightly less powerful than this new golden standard shows that these two nodes are either overly reactive because of the hydraulically distinct leaks or because of a transiently induced pressure field in that confined area. Therefore, Table 7.17 shows that they are still the same nodes in the most sensitive upper echelon; however, their means of response to the larger leak are not entirely redundant. This characteristic is advantageous for sensor placement optimisation, as nodes that exhibit strong responses with minimal overlap provide greater discriminatory information and enhance sensor selection diversity without compromising the integrity of the optimal sensor configuration.

Table 7.17: Correlation Matrix for Top 5 Nodes (100 mm)

	LN29	LN30	LN31	LN28	LN32
Filtered_Delta_LN29	1.000000	0.926487	0.917819	0.730000	0.822082
Filtered_Delta_LN30	0.926487	1.000000	0.999113	0.701963	0.925077
Filtered_Delta_LN31	0.917819	0.999113	1.000000	0.700253	0.933127
Filtered_Delta_LN28	0.730000	0.701963	0.700253	1.000000	0.696510
Filtered_Delta_LN32	0.822082	0.925077	0.933127	0.696510	1.000000

Table 7.18 reports the percentage of significant filtered deviation readings demonstrates the reliability of the top five nodes in detecting large leak events. For the 100 mm scenario, all five nodes exceed 96 % in significant response frequency, meaning that they produce meaningful deviation signals almost every time a leak occurs. This high rate of responsiveness is crucial for leak detection applications, as it ensures that these nodes consistently contribute actionable information in nearly all cases. The consistency of this percentage across multiple nodes indicates a strong and uniform hydraulic influence from larger leaks, reinforcing the suitability of these nodes for use in sensor placement strategies evaluated later in the optimisation process.

Table 7.18: Percentage of Significant Filtered Deviations (Top 5 Nodes, 100 mm)

Node	% Significant Readings
Filtered_Delta_LN29	96.774194
Filtered_Delta_LN30	96.774194
Filtered_Delta_LN31	96.774194
Filtered_Delta_LN28	96.774194
Filtered_Delta_LN32	96.774194

7.3.3.4 150 mm Leak Scenario – Sensitivity and Deviation Analysis:

The ranked sensitivity table for the 150 mm leak is truly a global assessment for node sensitivity for the aforementioned size as it shows how sensitive (magnitude/direction) each node is to an interconnected, more systematic, primary leak. The below Figure 7.8 summarises the hydraulic sensitivity patterns across all nodes for the 150 mm leak simulations. Thus, these are threshold-filtered deviations that should be expected average deviation determinations and per the expectations, they're mostly higher than the case studies for 12- and 18-mm orifices since this is the expected deviation impact from this orifice size. For instance, Node 28's sensitivity level is 10.90, the highest on the chart; Nodes 22, 25, 29 and 27 are the next highest thresholds all above 10.15. Such high sensitives are representative of how a greater outflow value creates a more extreme pressure differential across the system but also so specifically that they strongly align along primary hydraulic networks. Furthermore, the lowest sensitives for this orifice size represent how a high outflow value is not sensed on the extremes (high or low) of the rank value but instead, most of the sensitives perceived are in the mid-high periphery which compels the notion that it's a networked assessment across the entirety of the system regardless of size. Thus, this is the standard sensitive assessment which then brings in ranked data to justify hydraulic influence of what would otherwise be a strong impact sensitivity in Stage 5 to compare to reality for what would be a major leak. The full version of the ranked table can be found in the Appendix 10.7. The Figure 7.9 below provides complementary visual diagnostics for understanding the behaviour of the highest sensitivity nodes in the 150 mm leak dataset. Table 7.19 shows the ranked node sensitivity values for the 150mm Leak Scenario.

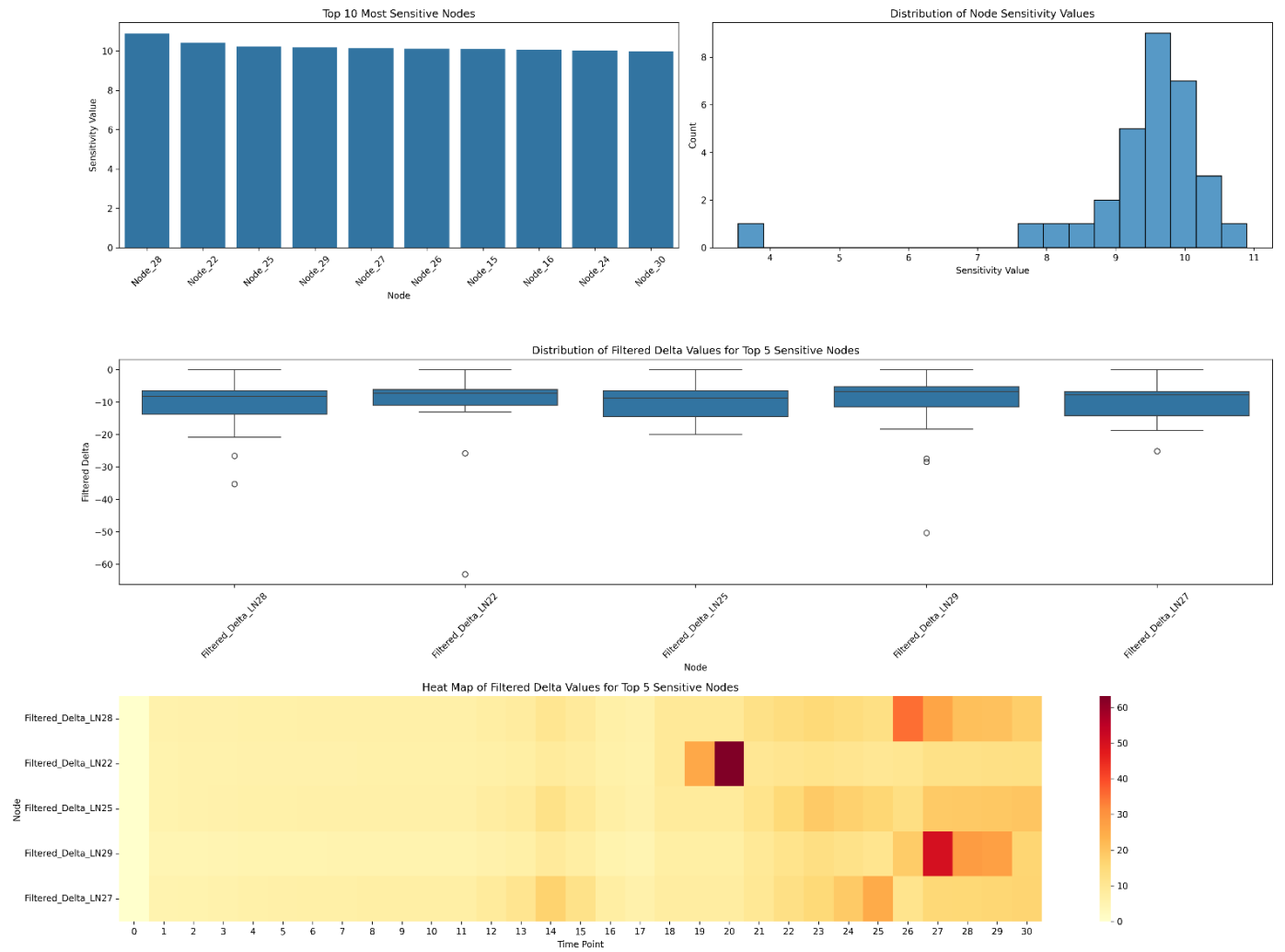


Figure 7.8: This Figure summarises the hydraulic sensitivity patterns across all nodes for the 150 mm leak simulations.

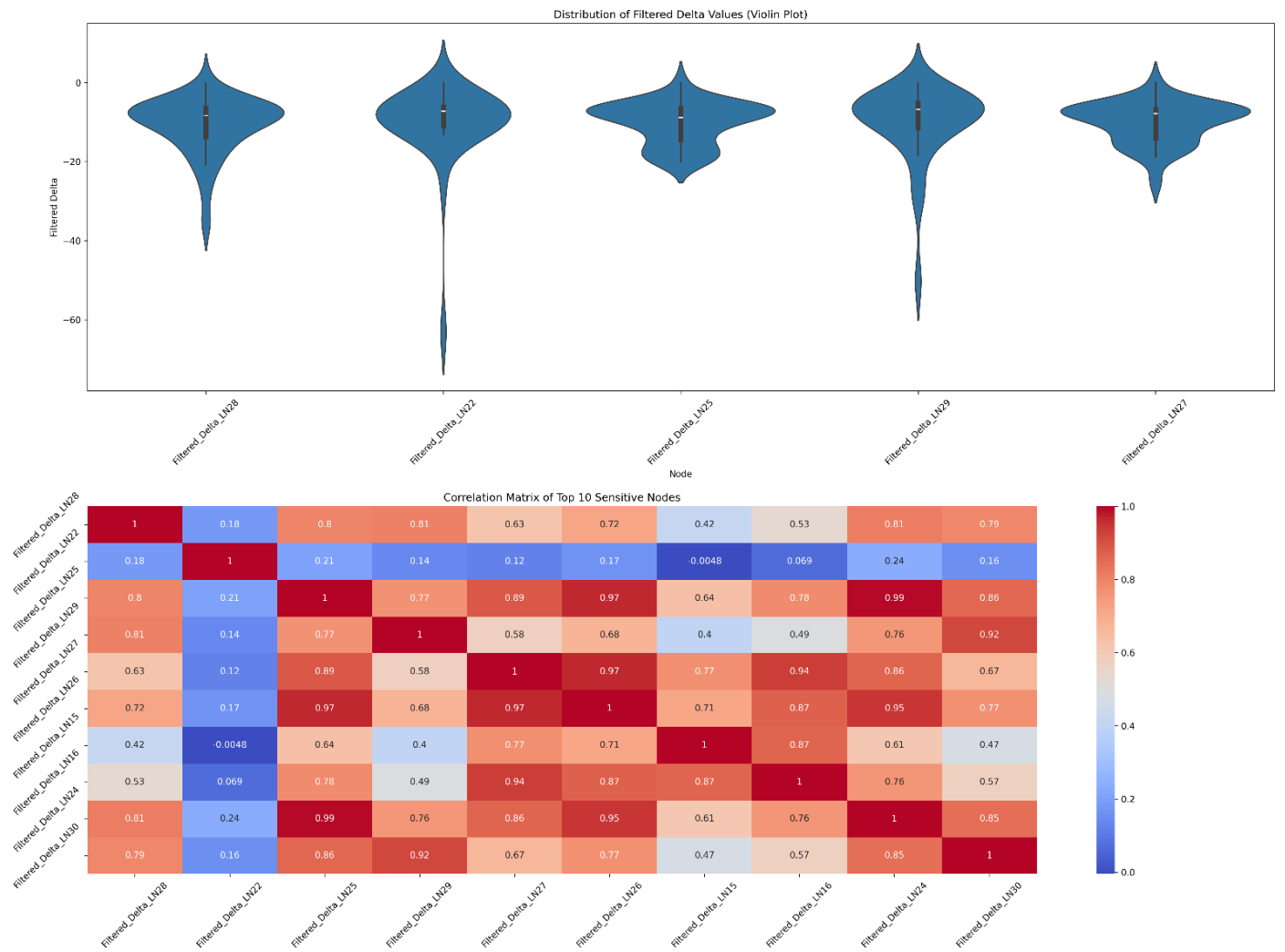


Figure 7.9: Complementary visual diagnostics for understanding the behaviour of the highest-sensitivity nodes in the 150 mm leak dataset.

Table 7.19: Ranked Node Sensitivity Values for the 150 mm Leak Scenario

Node	Sensitivity_Value
Node_2	3.536129
Node_3	8.239032
Node_4	8.659032
Node_5	9.030968
Node_6	9.313226
Node_7	9.367742
Node_8	9.458065
Node_9	9.521613
Node_10	9.562258
Node_11	9.597742
Node_12	9.612258
Node_13	9.498387
Node_14	9.865161
Node_15	10.08516
Node_16	10.06452
Node_17	9.509355
Node_18	9.30129
Node_19	8.909677
Node_20	9.179355
Node_21	9.697742
Node_22	10.4171
Node_23	9.669355
Node_24	10.03032
Node_25	10.22935
Node_26	10.1229
Node_27	10.15419
Node_28	10.90065
Node_29	10.19903
Node_30	9.980968
Node_31	7.640645
Node_32	9.21129

The summary statistics table 7.20 for threshold-filtered pressure deviations for the top five most sensitive nodes corresponds more to visually conveying magnitude, dispersion and mean value relatively consistent is an observation as a respondent consistently noted for these nodes mean

value consistency. Nodes 28, 22, 25, 29, and 27 are again, correlated map to very high filtered deviations (greater than 10 mm) with max deviations surpassing the 10 mm threshold. The non-zero populations indicate that relative to the null, across the filtered values per timestep for the entire duration of the simulation, deviations for these nodes are significant. Thus, per node response mean value consistency is good. Relative to the majority nodes being such sensitive nodes, mean and standard deviation serve as proportionate magnitude representatives as they are the actual amplitudes and extents relative to mean value of all findings connected to magnitude of findings across leak assessments. Thus, they imply that the majority nodes are those most often found to respond with such frequency and moreover, mean value consistency, the most powerful and statistically significant (mean value consistent) hydraulic signatures relative to all findings across all leaks. Thus, it makes sense these are the most sensitive as they are the most powerful and ultimately, for subsequent assessment studies for sensor placement, the most meaningful. Table 7.21 shows the Top 5 Most Sensitive Nodes (150mm).

Table 7.20: Summary Statistics for the Five Most Sensitive Nodes (150 mm Leak Scenario)

Node	Sensitivity_Value	Non_Zero_Count	Max_Delta	Mean_Delta	Std_Delta
Node_28	10.90065	30	35.27	10.90065	7.18282
Node_22	10.4171	30	63.08	10.4171	10.69185
Node_25	10.22935	30	19.94	10.22935	5.313512
Node_29	10.19903	30	50.33	10.19903	9.788875
Node_27	10.15419	30	25.15	10.15419	5.264718

Table 7.21: Top 5 Most Sensitive Nodes (150mm)

Node	Sensitivity_Value
Node_28	10.900645
Node_22	10.417097
Node_25	10.229355
Node_29	10.199032
Node_27	10.154194

The quartile statistics (Table 7.22) for the 150 mm leak summarise the typical configuration of the sensitivity output. The lower quartile (approximately 9.26), median (9.56) and upper quartile (10.05) indicate that the majority of nodes operate at significantly high sensitivity levels with little

separation between quartiles. This is because the 150 mm leak induced a highly forced pressure (head loss) deviation throughout the entire network as it provided a consistently high hydraulic effect in comparison to the smaller leaks. Thus, the fact that they're all bunched together means that few nodes are Nodes 28, 22 and 25 and are considered high-outliers, while the remainder have similarly elevated response levels. Thus, these quartile statistics provide a sense of meaningful separation from the mean in an otherwise deviant scenario for large leaks and help establish a framework through which to interpret the rank and correlation statistics.

Table 7.22: Sensitivity Value Quartiles (150 mm)

Statistic	Value
25th Percentile	9.256290
Median (50th Percentile)	9.562258
75th Percentile	10.047419

Table 7.23 shows the correlation matrix for these five most sensitive nodes with respect to major leaks illustrates how similar (or dissimilar) deviations are relative to an expected deviation in pressure one might expect hypothetically. In fact, where 10 mm and 50 mm are more similar and subtler, it is in the heterogeneity of the correlation response that drives researchers to deduce 150 mm greater. For example, nodes 28, 25 and 29 all have high pairwise correlations 0.80 from node 28 and node 29, 0.89 from node 25 and 27, which means that the magnitude which they deviate from nominal is at the same time and in the same direction as a major leak would cause hydraulic conditions, or how a major leak would cause hydraulic conditions that these nodes could deviate in unison as they did. Conversely, node 22 has much weaker correlations to those others like node 28 ~0.18 and node 29 ~0.14 indicating that the degree which it deviates is conditioned upon a differently based hydraulic path relative to them or perhaps a localised pressure pocket. Thus, it's important to assess the correlation strength and weakness as it pertains to major leaks since they impact all but yet major leaks still are largely independent of one another in terms of what they do based upon pressure response characteristics. Thus, the heterogeneity of this correlation is beneficial because should sensors be placed based on these highly sensitive response nodes, they would not be getting overlap information for the most part, but rather information that is distinct yet incredibly sensitive.

Table 7.23: Correlation Matrix for Top 5 Nodes (150 mm)

	LN28	LN22	LN25	LN29	LN27
Filtered_Delta_LN28	1.000000	0.181249	0.796953	0.806689	0.626835
Filtered_Delta_LN22	0.181249	1.000000	0.206198	0.139232	0.122986
Filtered_Delta_LN25	0.796953	0.206198	1.000000	0.765045	0.887920
Filtered_Delta_LN29	0.806689	0.139232	0.765045	1.000000	0.579866
Filtered_Delta_LN27	0.626835	0.122986	0.887920	0.579866	1.000000

The percentage-of-significant-deviation table 7.24 indicates, for all 150 mm leaks examined, by how much a threshold is surpassed on a per-node basis for the five most sensitive nodes. All five nodes constitute 96.77% of responsive signaling similar to the previous cases but now at much higher levels of significance which implies that higher levels of leakage create such hydraulic signaling from these nodes that they're both extremely reliable and overwhelmingly reliable. Thus, they're on signal basically all of the time when a leak is occurring, meaning that this is an ideal location for where a pressure sensor could be placed. Moreover, the more consistent they're all similar, the better, suggesting that with so much leakage, the hydraulic signal created is effective throughout the entirety of the network.

Table 7.24: Percentage of Significant Filtered Deviations (Top 5 Nodes, 150 mm)

Node	% Significant Readings
Filtered_Delta_LN28	96.774194
Filtered_Delta_LN22	96.774194
Filtered_Delta_LN25	96.774194
Filtered_Delta_LN29	96.774194
Filtered_Delta_LN27	96.774194

7.3.4 Comparative Evolution of Sensitivity Patterns (10 → 50 → 100 → 150 mm)

The sensitivity patterns observed across increasing leak magnitudes reveal a clear and interpretable progression in hydraulic behaviour. For the 10 mm dataset, deviations are relatively small and concentrated, producing sensitivity values that differ only narrowly among nodes. The quartile ranges are extremely tight, indicating an almost uniform network response, with only a small subset of nodes are primarily Nodes 29–31 and adjacent junctions displaying slightly higher sensitivity. The correlation matrices at this scale also show near-perfect similarity across the top

nodes, revealing a highly redundant sensitivity structure characteristic of small but detectable leaks.

At 50 mm, deviations become more pronounced, and the ranking structure begins to differentiate more distinctly between highly sensitive nodes and the remainder of the network. While most nodes still fall within a moderate range of sensitivity, the top nodes (e.g., Nodes 27, 16, 20, 14 and 15) demonstrate both higher deviation magnitudes and more frequent threshold exceedances. Importantly, the correlation matrix begins to show more variation than in the 10 mm case still strongly correlated overall, but with lower values for certain nodes such as Node 20, signaling that larger leaks introduce more diverse hydraulic signatures.

In the 100 mm dataset, the separation between highly responsive and moderately responsive nodes becomes more pronounced. Sensitivity values increase sharply, and the distribution widens compared to the 10 mm and 50 mm cases. The most sensitive nodes are now Nodes 29, 30, 31, 28 and 32 which exhibit substantially higher deviation levels and maintain a 96.77% rate of significant filtered deviations. Correlation structures show strong clustering among the top three nodes but more variability among Nodes 28 and 32, suggesting that larger leaks activate multiple hydraulic pathways and increase the diversity of local responses.

For the 150 mm leak scenario, the network enters a regime of uniformly high hydraulic disturbance. Sensitivity values across nearly all nodes are extremely elevated, and quartile statistics indicate that even the lower-ranked nodes exhibit strong responses. While the top nodes (e.g., Nodes 28, 22, 25, 29 and 27) remain significantly more sensitive than the rest, the difference between quartiles is much smaller relative to the overall sensitivity magnitude. The correlation matrix shows a mix of very strong correlations (e.g., between Nodes 25 and 27) and much weaker or even near-zero correlations (e.g., between Node 22 and others), demonstrating that extremely large leaks create both globally strong deviations and locally distinct response variations. This marks a clear divergence from the smaller leak sizes, where sensitivity patterns were more uniform and redundancy more pronounced.

Overall, the evolution from 10 → 50 → 100 → 150 mm shows a progression from subtle, near-uniform hydraulic deviations to large, structurally varied network responses. Smaller leaks reveal redundancy and tight sensitivity clustering, whereas larger leaks amplify node-to-node differences

and highlight hydrodynamically distinct sectors of the network. This gradation is crucial for sensor placement: small-leak detection relies heavily on identifying nodes with consistently high sensitivity, while large-leak detection benefits from the diversity of hydraulic behaviour among sensitive nodes. The combined interpretation across all leak magnitudes therefore establishes both the robustness and the strategic importance of the high-ranked nodes later selected in the optimisation stage and it is shown in Table 7.25.

Table 7.25: The Five Most Dominant Nodes Across All Leak Sizes

Node	Number of Times in Top 5	Leak Sizes
Node 29	3 times	10 mm, 100 mm, 150 mm
Node 28	3 times	10 mm, 100 mm, 150 mm
Node 27	3 times	10 mm, 50 mm, 150 mm
Node 15	2 times	10 mm, 50 mm
Node 30	2 times	10 mm, 100 mm

While Figure 7.10 confirms that the sensors are spatially well-distributed, it does not by itself quantify how consistent or reliable each node's pressure response is across multiple simulated leak events.

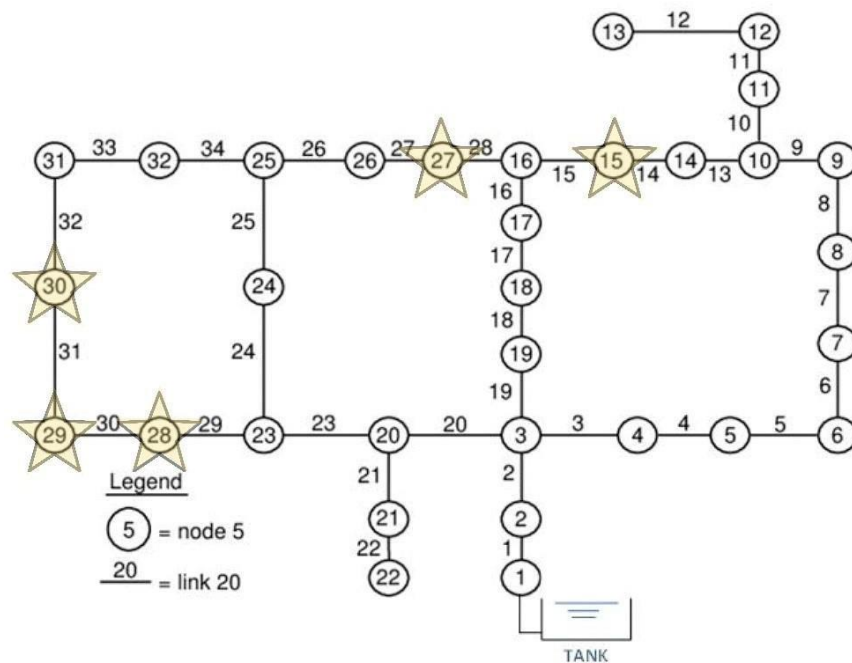


Figure 7.10: Spatial distribution of optimal sensor locations in the Hanoi water distribution network.

A review of the sensitivity rankings for the four leak sizes shows that relative to the others, for three of the four leak sizes (all but the 1.75 L/s leak) that Nodes 29, 28 and 27 represent the most sensitive locations in the network, placing them in the Top 5 for three of four simulated leak sizes. Furthermore, Node 15 and Node 30 exhibit considerable hydraulic impact for two leak sizes, each in the Top 5. Therefore, these five nodes are the strongest hydraulic response nodes overall and represent a starting candidate node set for sensor placement before true sensor placement optimisation.

The observed reduction in sensitivity for smaller leak sizes also indicates the presence of a practical lower detection limit, below which hydraulic responses may be indistinguishable from normal system variability. This highlights an inherent limitation in detecting very small leaks using pressure-based sensitivity analysis.

7.3.4 Comparative Evaluation of Detection Performance for All Optimisation Methods

The below Tables 7.26-7.28 compile performance measures detection rate, computational run time and algorithmic stability for all leak sizes and all optimisation methods. These results are in addition to Figures 7.11-7.13 for visual comparison and offer a compiled, numerical means of referring to performances of these algorithms. The framework for optimal placement of pressure sensors can be used for any network; however the outcome will be dependent on the pipe flows, pipe diameters, pipe lengths and pipe material properties. The 50 mm leak scenario exhibits distinct behaviour compared to smaller and larger leak cases, representing a transitional regime where hydraulic responses are neither uniformly weak nor uniformly strong. As a result, sensitivity patterns are more spatially variable and less correlated, leading to greater differentiation in optimisation performance across methods

The first type of comparative study between the algorithms concerns their detection rate performance for the four leak sizes (10 mm, 50 mm, 100 mm, and 150 mm). The detection rate is the proportion of leaked situations successfully acknowledged with the sensor arrangements produced by the various algorithms. Therefore, it is the most applicable performance metric related to the Optimal Sensor Placement (OSP) method and its success. The findings are illustrated in graphical form in Figure 7.11. Generally, D-Optimal, Greedy, Hybrid and Top-K (Sensitivity-based) achieve comparable performance as detection rate values boast high levels and consistently

maintain those levels regardless of leak size. Random (best-of-n) is the outlier with lower performance and for the case of 50 mm, more significantly dropped performance. These findings demonstrate the standardisation level in terms of all algorithms performing generally the same. Subsequently, time sensitivity and reliability are discussed. Table 7.26 shows the detailed detection rate comparison of different optimisation algorithms.

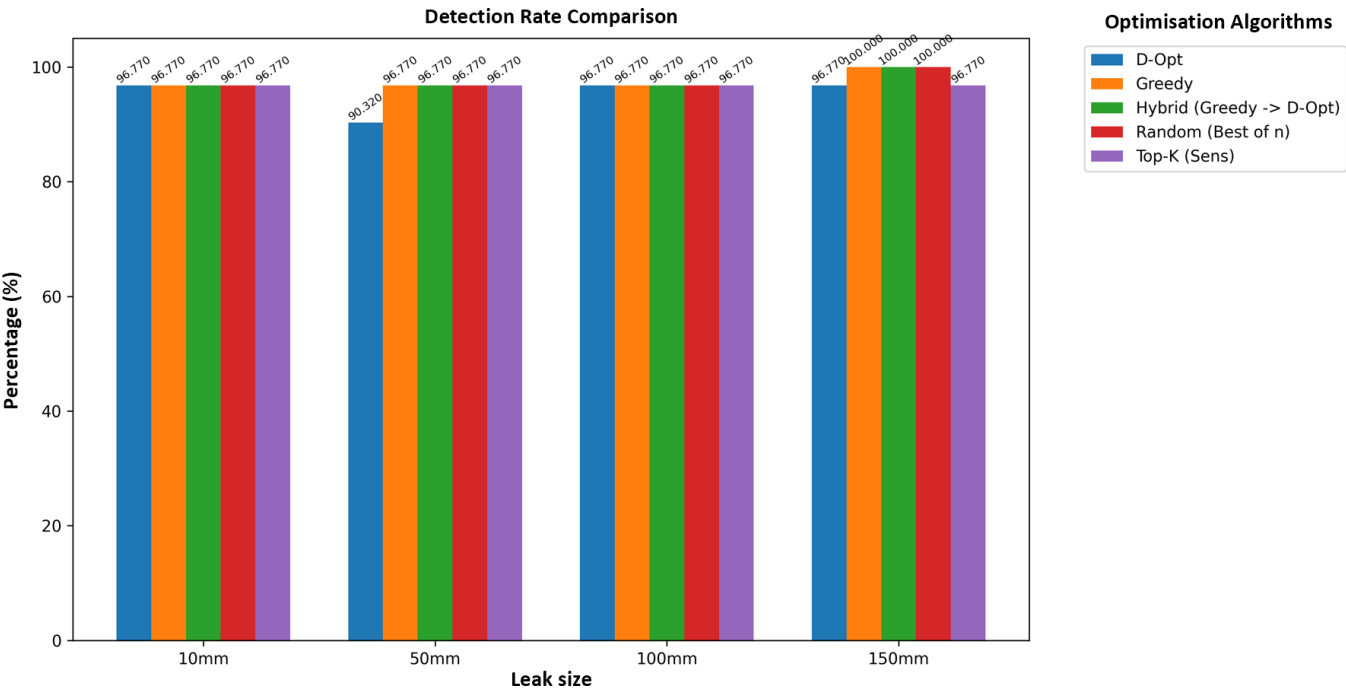


Figure 7.11: Detection Rate Comparison of Optimisation Algorithms Across All Leak Magnitudes (10–150 mm)

Table 7.26: Comparative Detection Rates Across All Optimisation Methods and Leak Magnitudes

Configuration	D-Opt	Greedy	Hybrid (Greedy -> D-Opt)	Random (Best of n)	Top-K (Sens)
10mm	96.77	96.77	96.77	96.77	96.77
50mm	90.32	96.77	96.77	96.77	96.77
100mm	96.77	96.77	96.77	96.77	96.77
150mm	96.77	100	100	100	96.77

Since these sensor placement approaches will be applied in the field given large scale networks or near-real time decisions, runtime becomes an important tertiary performance metric. Figure 7.12 and Table 7.27 shows the Computational Runtime of Sensor-Selection Algorithms for Different Leak Magnitudes. The runtimes comparative chart demonstrates the amount of time needed for each approach to develop a K-sensor placement solution for each of the four leak sizes.

We find that Top-K (Sensitivity-based) consistently has the best runtime as its evaluations take little time due to the lower number of calculated and nuanced approaches. For example, Random (best-of-n) had the largest runtime assessment since it considers so many random placement solutions before settling on the best one. The D-Optimal and Hybrid assessments have middle of the road runtimes as although they require a fair amount of calculations, they essentially take as long to yield a placement solution as they would to have a placement solution itself for testing. Therefore, this chart acknowledges practical considerations of these approaches based on runtime for applicability to near-real-time or large-scale hydraulic monitoring.

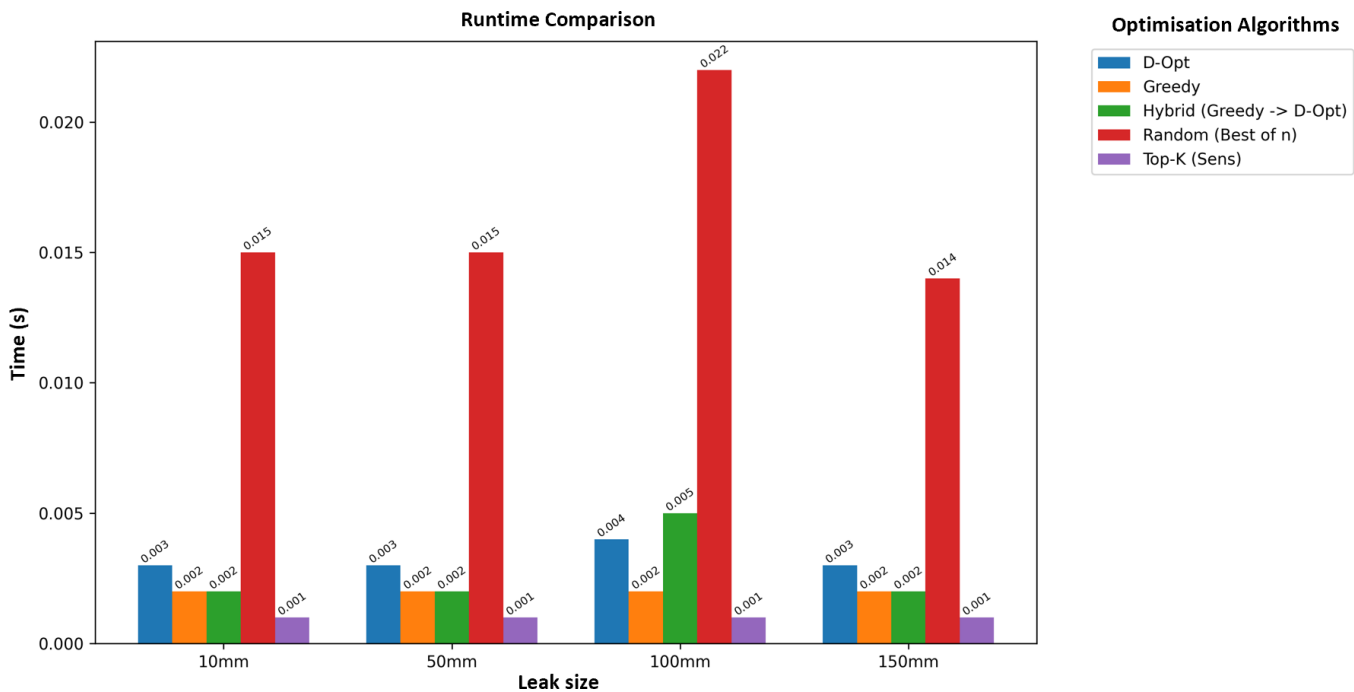


Figure 7.12: Computational Runtime of Sensor-Selection Algorithms for Different Leak Magnitudes

Table 7.27: Runtime Comparison of Optimisation Algorithms for Different Leak Magnitudes

Configuration	D-Opt	Greedy	Hybrid (Greedy -> D-Opt)	Random (Best of n)	Top-K (Sens)
10mm	0.003	0.002	0.002	0.015	0.001
50mm	0.003	0.002	0.002	0.015	0.001
100mm	0.004	0.002	0.005	0.022	0.001
150mm	0.003	0.002	0.002	0.014	0.001

In addition to accuracy and efficiency, selection stability refers to stability in regard to

configuration sensors being selected continuously for the various modeling configurations or threshold selections. Therefore, the following SI chart (Figure 7.13) relevant to stability represents how stable each configuration relative to each algorithm was in how much the threshold of perturbation decreased when scaled (table 7.28). A 1.0 means all configurations remain consistently stable across all perturbations of this threshold. Therefore, it can be concluded that once again, Top-K is completely stable regardless of leak size for all configurations meaning this threshold perturbation means nothing in terms of sensor selection, or in other words, Top-K is not dependent on threshold adjustment. In addition, D-Optimal is very stable as well but has more deviations because of the 50 mm leak dataset. Greedy, Hybrid and Random configuration selection have lesser stable statistics meaning less stability with small changes of the dataset. In fact, Hybrid has the lowest stability value of all methods applied, but at least Hybrid is more stable than Random for configuration selection. Thus, Figure 7.13 confirms the stable of means for configuration selection throughout the OSP and subsequently provides additional justification for sensitivity-type applications and D-optimal for field implementation.

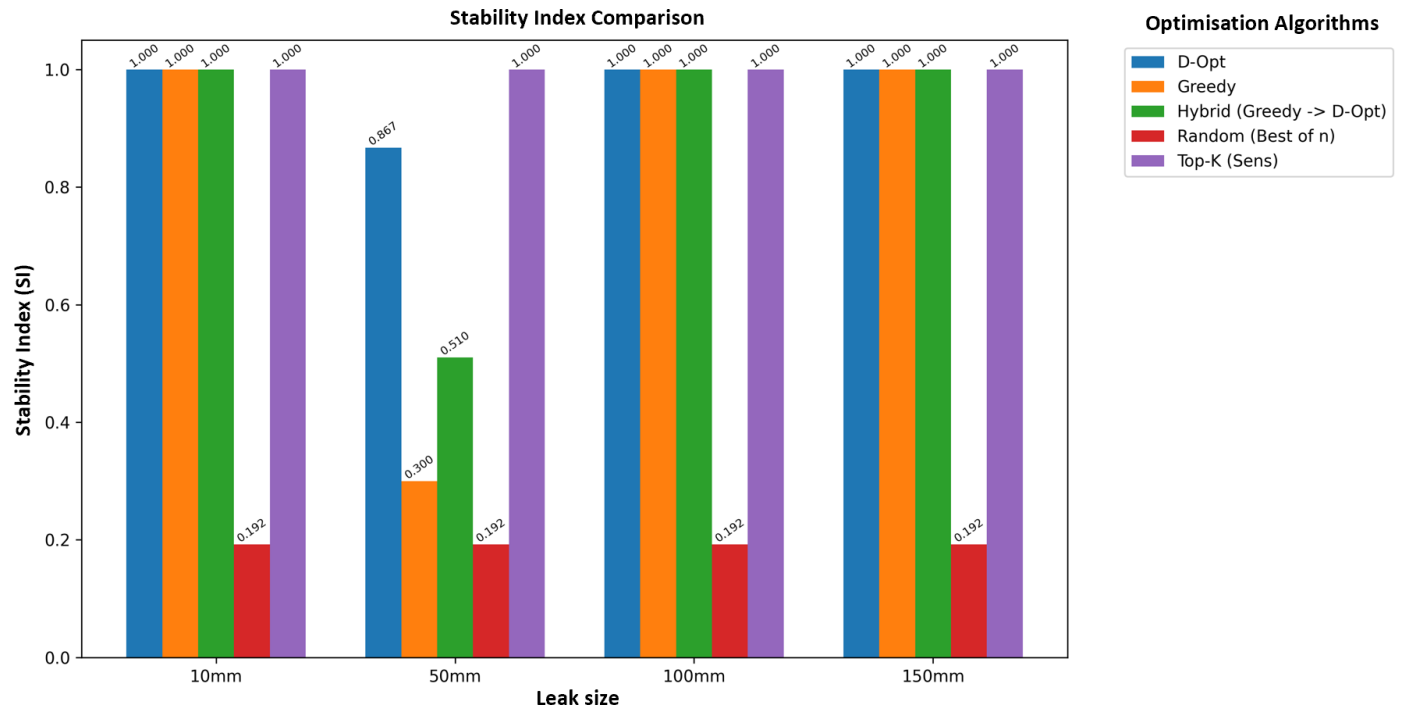


Figure 7.13: Stability Index of Optimisation Methods Under Threshold Variation for All Leak Sizes

Table 7.28: Stability Index of Sensor-Selection Algorithms Under Threshold Variation

Configuration	D-Opt	Greedy	Hybrid (Greedy -> D-Opt)	Random (Best of n)	Top-K (Sens)
10mm	1	1	1	0.192	1
50mm	0.867	0.3	0.51	0.192	1
100mm	1	1	1	0.192	1
150mm	1	1	1	0.192	1

7.4 Chapter Summary and Reflections

This chapter extended a comprehensive approach for understanding hydraulic sensitivity patterns and sensor placement viability for the Hanoi Water Distribution Network based on leak sizes (10 mm, 50 mm, 100 mm and 150 mm). The comprehensive approach derived from understanding the pressure response patterns manifested from EPANET-based results and converted to a deviation sensitivity detection matrix where thresholding (min/max) applied only the relevant hydraulic parameters of interest. From here, assessing the deviation sensitivity at the node level per leak size quantified the stability and stability intensity of deviation magnitude per junction. Therefore, from a thesis perspective, applying the sensitivity maps, it was clear that structural formations overlaid geometric findings of the network, whereby the transmission route to the south of the network had nodes that deviated significantly across metrics compiled - especially when they were larger and therefore easier to qualify in magnitude. This was substantiated by summarisable opportunities (min-max thresholds) over time and a linear correlation greater than additional increasingly smaller but increasingly clustered nodes began to appear as high responders.

In addition, using findings from the sensitivity as a reference, this chapter compared five methods of sensor placement optimisation - D-Optimal, Greedy, Hybrid (Greedy→D-Optimal), Top-K (Sensitivity) and Random (Best of n) - relative to three performance metrics: detection rate, runtime and threshold-based stability/reproducibility. The findings indicated that the deterministic methods trumped practical detection rates for all leaks tested relative to >96% detection rate for all leak sizes and 100% for the largest. Random selection proved to have poorer stability and runtime, further validating that randomised opportunities fail to produce consistency relative to why optimisation matters. Runtime evaluation validated that Top-K and Greedy methods possessed computational efficiency but D-Optimal and Hybrid methods also worked relatively well with acceptable run times. Finally, the stability index revealed that the deterministic algorithms possessed excellent stability across threshold changes but random will always have instability due to randomisation.

Thus, findings in this chapter suggest that an optimal sensor placement for the water distribution network is substantiated through a synergistic approach based on hydraulic sensitivity and sensor placement optimisation. In addition, Top-K (Sensitivity) was a simple and easily deployable

method that produced an acceptable position due to speed and stability but D-Optimal and Hybrid approaches maintain their advantageous efforts in areas where the lowest common denominator of spatial distribution is needed for maximised diversity. Therefore, it can be assessed that across multiple leak sizes not only do sensitivity values change due to orifice size, but also their sensitive magnitude is relatable as are the significant responders and cumulative sensitivity values.

Thus, methodologically, this chapter promotes transdisciplinary application of physical modelling through EPANET as a data-acquisition tool that can then be improved through processing and theoretical application across various fields to create systematic findings where exacting simplicity is the goal. These findings lend well to discuss based on their results. First, it is clear that deviations that do not operate threshold-at-acquisition present danger which does not help the sensitivity competition where most deviations stick from deviation signal to detection success. Second, there is a hegemonic relationship between spatial topology and hydraulic dominance that suggests certain corridors are always better positioned than others. Third, stability is a complicating factor that is often omitted from real-world sensor placements for design selection and finally, compiling sensitivity with an all-encompassing optimisation provides highly interpretable value for diagnostic understanding while improving operational stability when assessed in dynamic real Intelligent Water Networks.

Finally, this chapter provides a successful framework for future thesis chapters as it connects hydraulic patterns of behaviour to practicalised considerations for replicable frameworks with much larger and more complex water distribution systems than this small case study.

Chapter 8

Results and Discussions

8.1 General Remarks

This chapter presents and discusses the results obtained from both the logic-based hydraulic modelling and Machine Learning-based leak detection approaches. Section 8.2 presents the evaluation of the logic-based method using real-world SCADA data from the Sunbury network, including spatial and temporal accuracy of event detection. Section 8.3 details the experimental validation of Machine Learning algorithms using the Melton prototype setup, highlighting the impact of pre-processing techniques and feature selection. Section 8.4 compares the strengths and limitations of both approaches and explores their integration into a hybrid detection framework. Section 8.5 addresses operational performance and real-time alerting, while Section 8.6 synthesises key insights, limitations, and considerations for deployment at scale.

Together, these results illustrate how the proposed hybrid detection framework operates within real SCADA environments to deliver early, accurate, and explainable leak and burst identification. The findings confirm that combining deterministic hydraulic logic with adaptive data-driven analytics is both technically viable and operationally beneficial.

The logic-based algorithm successfully identified anomalies in the Sunbury Pumping Main with quantifiable spatial precision and temporal responsiveness. It demonstrated that physically interpretable rules can be implemented directly within existing telemetry systems without the need for additional sensors. The use of the Darcy–Weisbach equation, along with Swamee’s correlation for friction factor determination, enabled accurate localisation of events, validating the model’s hydraulic consistency and predictive power. In parallel, Machine Learning methods extended these insights through adaptive intelligence. Unsupervised techniques, particularly the Local Outlier Factor (LOF), improved significantly after applying data-balancing strategies such as ADASYN, confirming the value of robust pre-processing. Supervised models tested on the Melton prototype, including K-Nearest Neighbour, Random Forest, and Decision Tree, achieved perfect

classification across all performance metrics, establishing that the selected feature set of pressure, flow, and pump-speed variations captures the essential dynamics of leak behaviour.

Collectively, these findings verify that the hybrid configuration of logic-based and Machine Learning-based approaches produces a balanced detection architecture. The deterministic layer provides interpretability and immediate operational value, while the Machine Learning layer contributes adaptability and self-improvement as more data becomes available. Their combined use reduces false positives, enhances confidence in automated alerts, and enables continuous refinement without altering the physical system.

Crucially, the entire framework functions using existing SCADA data alone, with no requirement for hardware upgrades or additional field instrumentation. This outcome highlights the practicality and scalability of the approach for utilities seeking cost-effective pathways to intelligent water-network management. The consistent behaviour of pressure and flow variations observed across both the field-scale Sunbury dataset and the controlled Melton experiments further confirms the robustness and generalisability of the developed methodology.

Overall, the general findings of this research validate the central hypothesis: the integration of deterministic modelling and Machine Learning can transform conventional water network monitoring into a predictive, intelligent system. The following sections detail the comparative analysis of detection methods, evaluation of model performance, and lessons derived from applying the framework under both operational and experimental conditions.

8.2 Evaluation of the Logic-Based Model

The logic-based model was evaluated using live SCADA data from the Sunbury Pumping Main, a 6.24 km, 450 mm asbestos-cement pipeline. The recorded pressure data revealed three consistent operational phases: start-up (58–80 m head), steady-state (~77–80 m with oscillations of ± 0.6 m), and shutdown (gradual pressure decay). These readings established the baseline thresholds for detection.

Pressure variations were categorised using a structured thresholding approach to reflect the severity of detected anomalies. Variations smaller than 15% of the average head were classified

as minor leaks, 15–20% as major leaks, and greater than 20% as burst events. This approach provided a graded and explainable classification of anomaly severity and facilitated early-stage intervention before major failures occurred. Using the Darcy–Weisbach equation, with Swamee’s friction factor correlation, the burst localisation was verified to within approximately ± 40 m of actual field observations. This level of accuracy, achieved with minimal input variables, demonstrates the effectiveness of physically derived models for real-time leak localisation.

Importantly, the logic-based algorithm distinguished between transient fluctuations and sustained pressure deviations. Short-term anomalies during pump start-up or shutdown were ignored, preventing false alarms. The method’s speed, simplicity, and interpretability demonstrate its suitability for real-time field deployment using standard SCADA data.

8.3 Evaluation of Supervised and Unsupervised Learning Models

Machine Learning analysis was undertaken in two phases: unsupervised anomaly detection on unlabelled SCADA data and supervised classification on labelled experimental data.

8.3.1 Unsupervised Models (Sunbury Dataset)

The unlabelled SCADA dataset was first normalised and balanced to improve data representativeness. Among the algorithms tested, the Local Outlier Factor (LOF) achieved the strongest performance. On imbalanced data, LOF recorded an F1-score of 0.12 and ROC–AUC of 0.77. After applying the Adaptive Synthetic Sampling (ADASYN) method, the F1-score improved to 0.69 with ROC–AUC of 0.72, confirming that class-balancing significantly enhanced detection accuracy.

The Isolation Forest produced moderate results, while DBSCAN, K-Means, and One-Class SVM displayed limited ability to distinguish minor leaks, often classifying small deviations as noise. The LOF’s success is linked to its focus on local data density, a property that aligns closely with hydraulic behaviour where steady states form dense clusters, and leaks appear as isolated deviations. Hence, LOF model is identified as the most reliable unsupervised algorithm for

anomaly detection using SCADA data, particularly when supported by balancing techniques like ADASYN or SMOTE.

8.3.2 Supervised Models (Melton Experimental Setup)

Supervised models were trained using labelled data from the Melton prototype system. The dataset captured three classes: no leak, minor leak, and major leak by replicating realistic hydraulic conditions under controlled settings.

Fourteen algorithms were benchmarked using accuracy, precision, recall, and F1-score. K-Nearest Neighbour (KNN), Random Forest, and Decision Tree models achieved perfect classification across all metrics (accuracy, precision, recall, and $F1 = 1.0$). Other models, such as Support Vector Machine, Multilayer Perceptron, and AdaBoost, performed strongly but marginally below this level. Linear models (Logistic Regression, LDA, QDA) achieved moderate results, while Naïve Bayes and Ridge Classifier were least effective due to their inability to capture non-linear hydraulic interactions.

These findings confirm that variations in pressure, flow, and pump speed contain sufficient discriminative power for reliable classification of leak severity. Moreover, the strong agreement between unsupervised and supervised feature relevance underscores the robustness of the chosen input parameters and validates the methodology's consistency across field and laboratory scales.

8.4 Comparative Analysis of Detection Methods

Two complementary methodologies were developed and validated for leak and burst detection within pressurised water networks: a logic-based model grounded in hydraulic principles and a Machine Learning framework based on statistical inference. Both methods utilised Supervisory Control and Data Acquisition (SCADA) datasets from two distinct sites, Sunbury and Melton, allowing for direct comparison under field and experimental conditions.

The logic-based method operates on deterministic reasoning grounded in hydraulic principles. It analyses pressure, flow, and pump speed data in real time, flagging deviations that exceed predefined thresholds. This approach is transparent and requires no training data, relying instead

on explicit hydraulic logic to interpret deviations as either normal transients or potential anomalies. Conversely, the Machine Learning based framework relies on data-driven pattern recognition. It extracts multivariate statistical features from historical datasets and adapts through retraining, offering flexibility in dynamic network conditions. Both supervised and unsupervised learning techniques were applied, with unsupervised models used for unlabelled field data and supervised models applied to labelled experimental data.

The comparative analysis demonstrates that both approaches are effective in their respective contexts. The logic-based model provides physical interpretability and rapid detection, while the Machine Learning-based models offer adaptability to changing operating conditions. Their integration creates a hybrid framework that combines the precision of hydraulic reasoning with the adaptive intelligence of data-driven analytics.

In addition to the Sunbury and Melton case studies, the Hanoi Water Distribution Network analysis (Chapter 7) provides further validation of the proposed framework at a network scale. The sensitivity-based analysis demonstrated that leak detectability is strongly influenced by leak size and network topology, with larger leaks producing clearer and more spatially distributed pressure responses. Furthermore, the optimal sensor placement results confirmed that deterministic optimisation methods consistently outperform random selection, achieving high detection rates (>96%) with improved stability. These findings complement the earlier results by extending the applicability of the methodology from pipeline-scale systems to complex distribution networks, thereby reinforcing the generalisability of the proposed approach.

Overall, the findings from the Sunbury, Melton, and Hanoi case studies collectively address the research questions by demonstrating (i) the effectiveness of SCADA-based hydraulic and Machine Learning approaches for leak detection, (ii) the importance of data-driven and physically interpretable models, and (iii) the role of optimal sensor placement in enhancing detection performance at network scale.

8.5 Real-Time Alert Generation and Operational Performance

To enable live implementation of the proposed detection framework, a Python-based monitoring platform was developed and integrated directly with the SCADA historian. This system automatically processed incoming data, computed derived features (ΔP , ΔQ , ΔS), and simultaneously evaluated each data record using both the logic-based and Machine Learning detection modules.

When the logic-based module identified a threshold exceedance, an alert was immediately generated. The Machine Learning module then validated or rejected the event statistically. Confirmed alerts were logged with the corresponding timestamp, pressure, and flow rate, and the estimated leak location.

During field operation, the system processed continuous telemetry with negligible delay and displayed live time-series data with colour-coded indicators for normal, minor-leak, major-leak, and burst conditions. By requiring both models to agree before flagging an event, false alarms were effectively minimised. This configuration successfully combined physical reasoning with adaptive analytics in a transparent and fully automated workflow, demonstrating the feasibility of near-real-time anomaly detection within existing SCADA systems.

8.6 Lessons learned and key insights

1. The comparative evaluation of the Sunbury field trial and Melton laboratory prototype yielded several important insights that inform future deployment and refinement of the hybrid leak detection framework: Complementarity of Detection Techniques – The logic-based algorithm provides rapid, rule-based decisions with clear physical interpretation, while Machine Learning models enable adaptive detection through pattern recognition and statistical generalisation. Their combined use delivers both physical clarity and statistical flexibility.
2. Critical Role of Data Quality and Balance – The improvement observed after ADASYN balancing demonstrates that addressing class imbalance is crucial for dependable model training and detection accuracy.
3. Value of Experimental Validation – The Melton prototype verified that hydraulic patterns

observed in field data are repeatable under controlled conditions, confirming the transferability of the developed framework.

4. Ease of Integration with Existing Infrastructure – Both methods function entirely using SCADA data without new sensors or hardware, confirming operational feasibility and cost efficiency.
5. Operational Readiness and Usability – The dual-model system’s structured alert hierarchy and user-friendly visualisation promote operator trust, reduce response times, and improve decision confidence.

Collectively, these lessons support the practical implementation of scalable, hybrid leak detection systems across both new and aging water distribution networks. They also highlight the importance of aligning engineering interpretability with data-driven intelligence to address emerging infrastructure monitoring challenges.

8.7 Chapter Summary and Reflections

This chapter presented the empirical outcomes of the research and provided a critical interpretation of their technical and operational significance. The Sunbury field case study validated the logic-based model’s capability to detect anomalies accurately under real-world conditions, while the Melton laboratory prototype confirmed the reliability, precision, and adaptability of the Machine Learning framework. Together, these case studies demonstrated that integrating deterministic hydraulic logic with adaptive analytics enables robust, scalable, and explainable leak and burst detection.

Furthermore, the findings underscore the transformative potential of conventional SCADA-based monitoring systems when augmented by intelligent, predictive capabilities. The development of a real-time alert generation platform and its successful implementation highlight the feasibility of deploying proactive maintenance tools across water utilities using existing infrastructure.

The methodologies and results described in this chapter form a validated basis for wider deployment of Intelligent Water Networks, offering both technical advancement and operational readiness.

The insights, comparative evaluations, and performance outcomes articulated here set the stage for

the final synthesis in Chapter 9, which will summarise the theoretical contributions, practical implications, and strategic directions for future research and implementation.

Chapter 9

Conclusion

9.1 General Remarks

This chapter concludes the thesis by integrating the empirical findings, theoretical developments, and practical insights that have emerged throughout the research. It provides a reflective synthesis of how the integration of hydraulic modelling, SCADA analytics, and Machine Learning techniques has advanced intelligent leak and burst detection in pressurised water distribution systems.

The preceding results (Chapter 8) demonstrated that the proposed hybrid framework combining a deterministic logic-based algorithm with adaptive Machine Learning models can detect and localise network anomalies rapidly and accurately using only existing SCADA data. Building upon these verified outcomes, the present chapter situates those findings within a broader scientific and industrial context, outlining how each phase of the study contributed to knowledge advancement and operational innovation.

The chapter begins by revisiting the underlying motivation and the problem landscape facing modern water utilities, followed by a synthesis of the four major research phases: conceptualisation of Intelligent Urban Water Networks, systematic review of detection technologies, development of the logic-based and data-driven detection models, and experimental benchmarking using the Melton prototype. It then distils the principal theoretical contributions and applied benefits, discusses limitations, and identifies opportunities for future research and deployment.

In doing so, this chapter underscores that the significance of this work lies not merely in algorithmic accuracy but in demonstrating a practical, scalable, and explainable pathway toward intelligent, self-learning water networks that can transform conventional monitoring into proactive decision support for utilities.

9.2 Context and Motivation

9.2.1 Problem Overview and Industry Relevance

Water distribution networks (WDNs) form the backbone of urban infrastructure, ensuring the safe and reliable delivery of potable water to communities. However, these systems face increasing pressure from ageing assets, rapid urbanisation, climate variability, and escalating non-revenue water (NRW) losses. In Australia, NRW typically ranges between 8 and 25 per cent for metropolitan utilities, with leaks and bursts accounting for a major proportion of treated-water losses (Kingdom, Liemberger & Marin 2006; Lambert & Hirner 2000; Farley & Trow 2003). Conventional detection methods such as acoustic monitoring, manual inspection and step-testing remain costly, labour-intensive and primarily reactive. Recent scholarship highlights the need for predictive, self-learning systems that can diagnose and localise failures proactively (Joseph, Sharma & van Staden 2022; Sharma et al. 2021; Li et al. 2023).

Globally, water utilities are increasingly challenged by rising non-revenue water (NRW) losses, frequent pipeline failures, and escalating operational costs, making the intelligent management of water distribution networks a critical priority (Liemberger & Wyatt, 2019; Lambert et al., 2014).

9.2.2 Research Aim and Methodological Trajectory

This research was undertaken to develop a comprehensive and intelligent framework for detecting leaks, bursts, and theft in water pipeline networks, by integrating physical modelling, data analytics, and artificial intelligence into a unified and adaptive system. Through a sequence of four peer-reviewed studies undertaken between 2022 and 2025, this research systematically evolved from conceptual design and literature synthesis to applied field implementation and experimental validation. Collectively, these investigations established the scientific foundation, methodological innovations, and practical implications of Intelligent Water Networks, transforming conventional practices into predictive, data-driven paradigms.

9.2.3 Phase 1: Conceptualisation of Intelligent Urban Water Networks (IUWN)

The first stage of this research focused on the conceptualisation of an Intelligent Urban Water Network (IUWN) system (Joseph, Sharma & van Staden 2022). This foundational work identified the limitations of traditional centralised water systems and articulated the need for intelligent,

decentralised networks that combine live sensor data, Supervisory Control and Data Acquisition (SCADA) telemetry, and predictive analytics. By tracing the evolution of urban water systems and analysing pressures from urbanisation and asset ageing, the study examined the evolution of water systems and the operational pressures created by urbanisation and infrastructure ageing. It proposed an integrated framework that incorporated pressure sensors, water-quality monitors, and high-resolution flow meters connected through SCADA systems to analytical platforms capable of real-time simulation and hydraulic modelling.

The framework provided a structured method for collecting, processing, and analysing live data for informed decision-making. It demonstrated the potential of fusing infrastructure data from pipes, pumps, and sources with predictive modelling to generate actionable insights. Importantly, it establishes the conceptual base for the thesis and defines the principles that guided all subsequent experimental and computational investigations. It also introduced the notion that intelligent water systems must be both technologically and organisationally adaptive leveraging artificial intelligence and Machine Learning for predictive maintenance while ensuring compatibility with existing infrastructure and operational practices.

9.2.4 Phase 2: Review of Hardware and Software-Based Detection Technologies

The second phase (Joseph, Sharma, van Staden & Wasantha 2023) extended this conceptual foundation through a systematic review of global literature on the application of software- and hardware-based technologies for leak and burst detection. This review provided a critical synthesis of the state of the art in intelligent pipeline monitoring, identifying key trends, research gaps, and emerging technologies. The analysis demonstrated that hardware-based systems, such as acoustic sensors, flow meters, and pressure transducers, have been successfully applied for real-time detection and localisation of leaks, while software-based systems particularly those leveraging hydraulic modelling, artificial intelligence, and Machine Learning offer predictive capabilities that support early intervention. The review found that hybrid solutions combining both approaches yield the most effective results, providing the accuracy and reliability of field measurements with the adaptive intelligence of data-driven analytics. It also identified a global shift from reactive to proactive pipeline management, emphasising continuous monitoring, data aggregation, and predictive diagnostics as defining characteristics of modern Intelligent Water Networks. The review proposed a hybrid software–hardware framework as a scalable and sustainable model for utilities

seeking to enhance network reliability and efficiency. This framework promoted the integration of multiple sensor types and analytical models into a cohesive monitoring platform capable of continuous learning. By adopting advanced technologies and best practices from across the literature, water utilities could reduce water losses, improve safety and environmental outcomes, and strengthen long-term system resilience. This synthesis of research evidence served as the intellectual bridge between conceptualisation and implementation, guiding the development of field-based and experimental models in the following stages.

9.2.5 Phase 3: SCADA-Based Algorithm Development for Real-Time Detection

The third stage of the research (Joseph et al. 2024) translated the theoretical and review insights into a practical implementation using SCADA data from an operational water distribution network in Sunbury, Victoria, Australia. The focus was on developing and validating both logic-based and Machine Learning algorithms for detecting leaks and bursts in a SCADA-equipped pumping main. The study compared threshold-based rule logic with data-driven Machine Learning in detecting anomalies. The research found that logic-based algorithms, relying on predetermined thresholds and hydraulic parameters, perform efficiently in identifying straightforward anomalies, providing clear and interpretable results. However, these models often struggle with complex, evolving, or subtle patterns that lie outside predefined conditions. To address these challenges, the study introduced Machine Learning algorithms capable of learning from historical patterns and adapting to dynamic operational changes. A comparative analysis of five unsupervised and semi-supervised models DBSCAN, Isolation Forest, K-Means, Local Outlier Factor (LOF), and One-Class SVM was undertaken. Data augmentation methods such as SMOTE and ADASYN were applied to improve classification balance and robustness. The LOF model achieved the best performance with an F1 score of 0.69 and ROC-AUC of 0.72. Consequently, it was subsequently integrated into a web-based detection platform that allowed utilities to process live data and visualise anomalies through a user-friendly interface. The study confirmed that Machine Learning complements logic-based approaches by identifying complex patterns and enabling continuous model refinement. Importantly, the integration of these methods within existing SCADA infrastructure demonstrated the feasibility and practicality of adopting artificial intelligence for near-real-time leak and burst detection without the need for new hardware investment. The research represented a major step in moving from theoretical framework to operational application, validating that intelligent, hybrid

systems can significantly reduce water losses and maintenance costs while improving response times.

The development of the logic-based leak-detection algorithm represented the first operational implementation of the intelligent framework conceptualised earlier in this thesis. It directly responded to the limitations identified in the literature, where most existing techniques were either hardware-dependent or lacked interpretability for real-time use. The logic-based approach applied deterministic hydraulic reasoning using readily available SCADA data, demonstrating that high accuracy and spatial localisation could be achieved without additional instrumentation. This phase established a transparent, physically grounded method that not only validated the practicality of using SCADA for continuous anomaly monitoring but also provided a structured baseline for integrating adaptive Machine Learning models. Thus, the logic-based method served as the empirical bridge between theoretical concepts derived from the review studies and the advanced data-driven models developed in subsequent phases.

9.2.6 Phase 4: Machine Learning Model Benchmarking Using Prototype Testbed

The fourth phase (Joseph et al. 2025) further advanced the experimental and computational work by performing a rigorous comparison of fourteen Machine Learning algorithms for early leak and burst detection using controlled data from a prototype testbed at Melton, Victoria, Australia. The experimental system replicated real-world network conditions under three scenarios no leak, leak, and major leak to evaluate the predictive performance of models across varying conditions. The dataset enabled multi-class classification and served as a benchmark for assessing algorithmic accuracy, precision, recall, and robustness.

Results revealed that Random Forest, K-Nearest Neighbours (KNN), and Decision Tree models consistently outperformed linear and generative approaches, particularly in terms of recall and generalisation. These models effectively captured nonlinear dependencies and subtle hydraulic anomalies, offering robust classification with low false-negative rates, a critical advantage for early intervention. Linear models such as Logistic Regression and Ridge Classifier performed less effectively due to their inability to capture complex, nonlinear relationships, while Gaussian Naïve Bayes was limited by the assumption of feature independence. The superior performance of

ensemble and instance-based algorithms reaffirmed the value of nonparametric, data-driven learning in complex network environments. While high accuracy was achieved, the study also acknowledged the potential for overfitting due to controlled, low-noise experimental conditions and recommended future work involving higher-frequency data and more diverse field environments. The findings established a strong evidence base for the adoption of Machine Learning techniques in practical water-utility operations and confirmed the feasibility of building automated, scalable leak detection systems integrated with existing infrastructure. Together, these results represented the culmination of the research trajectory, linking conceptual frameworks, theoretical reviews, and applied analytics into a unified, intelligent detection model.

Across these interconnected studies, this thesis contributes to both theory and practice in intelligent water management. The research demonstrates that combining hydraulic modelling, SCADA analytics, and artificial intelligence can significantly improve the detection, localisation, and management of leaks, bursts, and theft in water distribution networks. The development and validation of hybrid models that integrate logic-based reasoning with adaptive learning algorithms represent a substantial methodological contribution. These approaches bridge the interpretability of traditional hydraulic logic with the flexibility and scalability of modern Machine Learning, producing systems that are both operationally efficient and scientifically robust. The comparative framework developed across the studies provides utilities with clear guidance on model selection and deployment strategy. Logic-based models remain useful for their simplicity and transparency, while adaptive models such as Random Forest and KNN enhance predictive accuracy and resilience to noise. The complementary relationship between these paradigms ensures that Intelligent Water Networks can evolve dynamically as new data and operating conditions emerge. This fusion of approaches also offers a pathway toward digital-twin architectures, where real-time data streams continuously update simulation models to provide predictive insights and scenario forecasting.

The thesis makes several significant contributions to knowledge. Conceptually, it presents one of the earliest comprehensive frameworks for intelligent water-network development within an Australian context, integrating engineering, data science, and management principles. Methodologically, it establishes a multi-stage empirical process from literature synthesis to field validation that can serve as a reference model for future researchers. Technically, it demonstrates

the successful application of artificial intelligence for real-time leak detection and network monitoring using existing SCADA data, thus offering a cost-effective solution for utilities. Practically, the outcomes contribute directly to improving operational performance, reducing non-revenue water, and enhancing asset longevity. The research aligns with the global movement toward smart water systems and supports national objectives related to sustainability, digital transformation, and climate adaptation. It also contributes to the broader field of infrastructure analytics by illustrating how cross-disciplinary integration can yield scalable, high-impact solutions.

9.2.7 Limitations and Future Directions

While the achievements of this research are substantial, several limitations warrant consideration. The SCADA datasets used in the studies were constrained by one-minute sampling intervals, which limit the capture of transient pressure events. The experimental testbed operated under controlled conditions, which, although necessary for calibration, may not fully represent the variability of field networks. Machine Learning models, particularly ensemble methods, entail higher computational costs that could pose challenges for real-time edge deployment in resource-constrained environments. In addition, the transferability of models between networks with different pipe materials, diameters, and topologies requires further validation. These limitations do not diminish the contributions of the work but rather provide direction for ongoing research. Addressing them through advanced sensor technologies, higher-frequency data acquisition, and adaptive retraining frameworks will enhance model reliability and generalisability.

The future direction emerging from this thesis lies in the integration of Machine Learning models with Internet of Things (IoT) architecture and edge analytics to enable ultra-low-latency detection and control. The integration of these approaches within a real-time digital-twin framework will allow utilities to simulate, predict, and manage network conditions continuously. Equally important is the need to address the ethical and governance dimensions of AI deployment in critical infrastructure, ensuring transparency, accountability, and data security in automated decision-making processes. By building upon the foundations established in this research, future work can lead to self-optimising water systems that not only detect and prevent failures but also dynamically adjust operations to achieve optimal performance and sustainability.

9.2.7 Dissemination and Recognition

The research has been disseminated through peer-reviewed journals and international conferences, ensuring that the findings contribute to both academic advancement and industry practice. The cumulative publications from 2022 to 2025 make a complete progression from conceptual framework to implementation and optimisation, reflecting a coherent and systematic investigation into intelligent water-network management.

9.2.8 Concluding Statement

This thesis concludes that the convergence of hydraulic modelling, SCADA data analytics, and artificial intelligence offers an effective, scalable, and economically viable pathway to improving water-distribution efficiency and reliability. Intelligent Water Networks are both technically viable and necessary to meet future urban infrastructure demands. By integrating physical and digital infrastructures, this research contributes to the global transformation of water management, bridging the gap between traditional engineering practices and emerging data-driven paradigms. It establishes a scientific and practical foundation for utilities to evolve toward fully intelligent, early anomaly self-detection water systems capable of delivering resilient and sustainable water services for future generations.

The research was also recognised within the professional water sector, being selected as a finalist for the 2025 Australian Water Association (AWA) Victoria Student Water Prize. This recognition underscores the innovation, impact and applied relevance of the intelligent leak and burst-detection frameworks developed through the study, highlighting their potential to influence the future direction of intelligent and sustainable urban water management.

9.3 Chapter Summary and Reflections

This chapter has provided a comprehensive summary of the key findings, theoretical insights, and practical contributions derived from the research. While the integration of hydraulic modelling, SCADA analytics, and Machine Learning methods has demonstrated significant potential for intelligent leak detection, it is important to acknowledge that water-network optimisation remains

an evolving domain. Intelligent network management continues to face emerging challenges, including real-time integration of high-resolution IoT sensor data, adaptive sensor deployment strategies, and the lack of unified cross-utility protocols. These unresolved complexities signal opportunities for continued innovation. The next chapter presents recommendations for future research directions, focusing on technological enhancements, methodological refinements, and cross-disciplinary approaches that aim to expand the impact and scope of this research.

References

1. Aggarwal, CC 2015, *Outlier analysis*, Springer, Cham.
2. Al-Fuqaha, A, Guizani, M, Mohammadi, M, Aledhari, M & Ayyash, M 2015, 'Internet of Things: a survey on enabling technologies, protocols, and applications', *IEEE Communications Surveys & Tutorials*, vol. 17, no. 4, pp. 2347–2376.
3. Al-Ghamdi, SG 2011, 'Diagnosis of water pipe failures', *Engineering Failure Analysis*, vol. 18, no. 2, pp. 474–484.
4. Angelakis, AN & Koutsoyiannis, D 2003, 'Water supply and hydraulic technology in ancient Greece', *Water Science and Technology: Water Supply*, vol. 3, no. 1–2, pp. 65–71.
5. Angelakis, AN, Koutsoyiannis, D & Tchobanoglous, G 2005, 'Urban wastewater and stormwater technologies in ancient Greece', *Water Research*, vol. 39, pp. 210–220.
6. Angelakis, AN, Spyridakis, SV, Tchobanoglous, G, Angelakis, G & Rose, JB 2012, 'Evolution of water supply, sanitation, and wastewater technologies through the centuries', *Water Research*, vol. 46, pp. 714–726.
7. Atkinson, AC, Donev, AN & Tobias, RD 2007, *Optimum experimental designs, with SAS*, Oxford University Press, Oxford.
8. Ayala-Cabrera, D, Herrera, M, Izquierdo, J & Pérez-García, R 2011, 'GPR laboratory tests and numerical simulation for the detection of leaks in water distribution systems', *Journal of Applied Geophysics*, vol. 75, no. 4, pp. 673–682.
9. Bacco, M, Barsocchi, P, Ferro, E, Gotta, A & Ruggeri, M 2017, 'The digitisation of water management: a survey on ICT and IoT solutions for smart water networks', *Sensors*, vol. 17, no. 11, p. 2417.
10. Bao, X & Chen, L 2012, 'Recent progress in distributed fiber optic sensors', *Sensors*, vol. 12, no. 7, pp. 8601–8639.
11. Barate, PR 2017, 'Study on benefits and economic feasibility of smart water distribution system', *International Journal of Innovative Science and Research Technology*, vol. 2, pp. 255–260.

12. Berardi, L & Giustolisi, O 2021, 'Physically based calibration of WDN models for leakage management', *Water Resources Management*, vol. 35, no. 4, pp. 1231–1248.
13. Bishop, CM 2006, *Pattern recognition and machine learning*, Springer, New York.
14. Breiman, L 2001, 'Random forests', *Machine Learning*, vol. 45, no. 1, pp. 5–32.
15. Breunig, MM, Kriegel, H-P, Ng, RT & Sander, J 2000, 'LOF: identifying density-based local outliers', in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, ACM, Dallas, TX, pp. 93–104.
16. Campello, RJGB, Moulavi, D & Sander, J 2013, 'Density-based clustering based on hierarchical density estimates', in *Advances in Knowledge Discovery and Data Mining*, Springer, Berlin, Heidelberg, pp. 160–172.
17. Campos, L, Sousa, V & Teixeira, J 2018, 'An IoT framework for smart water distribution systems', *Sensors*, vol. 18, no. 7, p. 2212.
18. Caselton, WF & Zidek, JV 1984, 'Optimal monitoring network designs', *Statistics & Probability Letters*, vol. 2, no. 4, pp. 223–227.
19. Casillas, M, Garza-Castañón, LE & Puig, V 2013, 'Optimal sensor placement for leak location in water distribution networks using a geostatistical approach', *Procedia Engineering*, vol. 70, pp. 1137–1146.
20. Chandola, V, Banerjee, A & Kumar, V 2009, 'Anomaly detection: a survey', *ACM Computing Surveys*, vol. 41, no. 3, pp. 1–58.
21. Chawla, NV, Bowyer, KW, Hall, LO & Kegelmeyer, WP 2002, 'SMOTE: synthetic minority over-sampling technique', *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357.
22. Chen, C, Liaw, A & Breiman, L 2004, 'Using random forest to learn imbalanced data', technical report, University of California, Berkeley.
23. Chen, J, Wang, K, Wang, H, Guo, J, Sun, Y & Jiang, B 2017, 'A review of smart water network technologies', *Environmental Engineering Science*, vol. 34, no. 5, pp. 273–285.
24. Chi, Z, Li, Y, Wang, W, Xu, C & Yuan, R 2021, 'Detection of water pipeline leakage based on random forest', *Journal of Physics: Conference Series*, vol. 1978, no. 1, 012044, doi:10.1088/1742-6596/1978/1/012044.
25. Colombo, AF & Karney, BW 2002, 'Energy and costs of leaky pipes: toward comprehensive picture', *Journal of Water Resources Planning and Management*, vol. 128, no. 6, pp. 441–450.
26. Colombo, AF, Lee, P & Karney, BW 2009, 'A selective literature review of transient-based leak detection methods', *Journal of Hydro-environment Research*, vol. 2, no. 4, pp. 212–227.

27. Cortes, C & Vapnik, V 1995, 'Support-vector networks', *Machine Learning*, vol. 20, no. 3, pp. 273–297.
28. Cover, T & Hart, P 1967, 'Nearest neighbor pattern classification', *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27.
29. Covas, D & Ramos, HM 2010, 'Hydraulic transients used for leak detection in water distribution systems', *Journal of Water Resources Planning and Management*, vol. 136, no. 2, pp. 248–257.
30. Darweesh, A & Moghavvemi, M 2020, 'A review of leak detection and localisation in water supply systems', *Renewable and Sustainable Energy Reviews*, vol. 121, p. 109704.
31. Dietterich, TG 2000, 'Ensemble methods in machine learning', in *Multiple Classifier Systems*, Springer, Berlin, Heidelberg, pp. 1–15.
32. Dong, H, Ren, H, Li, J & Zhao, Y 2022, 'Smart water network: state of the art and future trends', *Water Research*, vol. 219, p. 118607.
33. Ester, M, Kriegel, H-P, Sander, J & Xu, X 1996, 'A density-based algorithm for discovering clusters in large spatial databases with noise', in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, AAAI Press, Portland, OR, pp. 226–231.
34. Fagiani, M, Squartini, S, Gabrielli, L, Severini, M & Piazza, F 2016, 'A statistical framework for automatic leakage detection in smart water and gas grids', *Energies*, vol. 9, no. 8, p. 665, doi:10.3390/en9080665.
35. Fantozzi, M, Liemberger, R & Lambert, A 2019, 'Real losses and apparent losses: latest developments in water loss management', *Water Utility Journal*, vol. 25, pp. 1–11.
36. Farajzadeh, N, Sadeghzadeh, N & Jokar, N 2024, 'Water distribution pipe lifespans: predicting when to repair the pipes in municipal water distribution networks using machine learning techniques', *PLOS Water*, vol. 3, no. 3, e0000164, doi:10.1371/journal.pwat.0000164.
37. Farley, M & Liemberger, R 2005, 'Developing a non-revenue water reduction strategy: planning and implementing the strategy', *Water Science and Technology: Water Supply*, vol. 5, no. 1, pp. 41–50, doi:10.2166/ws.2005.0006.
38. Farley, M, Wyeth, G, Ghazali, ZB, Istandar, A & Singh, S 2008, *The manager's non-revenue water handbook: a guide to understanding water losses*, Ranhill Utilities Berhad & United States Agency for International Development, Kuala Lumpur and Washington, DC.
39. Fares, A, Tijani, IA, Rui, Z & Zayed, T 2023, 'Leak detection in real water distribution networks based on acoustic emission and machine learning', *Environmental Technology*, vol. 44, no. 24, pp. 3850–3866, doi:10.1080/09593330.2022.2061329.

40. Fereidooni, D, Ghaderi, A, Nazif, S & Ashoori, A 2019, 'Hybrid hydraulic–AI method for leak detection in water pipelines', *Water Science and Technology: Water Supply*, vol. 19, no. 6, pp. 1841–1851.
41. Fields, KA, Chen, ASC, Tafuri, AN & Smith, LA (eds) 2009, *Options for leak and break detection and repair of drinking water systems*, US Environmental Protection Agency, Washington, DC.
42. Freund, Y & Schapire, RE 1997, 'A decision-theoretic generalization of online learning and an application to boosting', *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119–139.
43. Gao, Y, Zhang, S, Xu, C, Zhou, Y & Yang, Y 2022, 'Distributed optical fiber acoustic sensing for long-distance pipeline monitoring', *Applied Optics*, vol. 61, no. 10, pp. 2905–2916.
44. Goodfellow, I, Bengio, Y & Courville, A 2016, *Deep learning*, MIT Press, Cambridge, MA.
45. Gorenstein, D 2022, 'Comparative evaluation of data-driven leak-prediction models in urban water networks', *Water Resources Research*, vol. 58, no. 11, e2022WR032489.
46. Gourbesville, P 2011, *ICT for water efficiency*, IntechOpen, London.
47. Greater Western Water 2019, *Network intelligence strategy (NIS)*, internal technical report, Greater Western Water, Victoria, Australia.
48. Hammond, P, Taylor, J & Krueger, R 2024, 'Integrating energy recovery and leakage management in urban water systems', *Water Resources Management*, vol. 38, no. 4, pp. 1115–1129.
49. Hastie, T, Tibshirani, R & Friedman, J 2009, *The elements of statistical learning: data mining, inference, and prediction*, 2nd edn, Springer, New York.
50. He, H & Garcia, EA 2009, 'Learning from imbalanced data', *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284.
51. He, H, Bai, Y, Garcia, EA & Li, S 2008, 'ADASYN: adaptive synthetic sampling approach for imbalanced learning', in *Proceedings of the IEEE International Joint Conference on Neural Networks*, IEEE, Hong Kong, pp. 1322–1328.
52. Hu, X, Zhang, Y, Zhou, D, Wang, J & Zhang, H 2021, 'Leak detection in water distribution networks using support vector machines', *Water*, vol. 13, no. 3, p. 325.
53. Hunaidi, O & Chu, W 1999, 'Acoustical characteristics of leak signals in plastic water distribution pipes', *Applied Acoustics*, vol. 58, no. 3, pp. 235–254.
54. Hunaidi, O & Giamou, P 1998, 'Ground-borne vibration signal for leak detection in plastic water distribution pipes', *Journal of Canadian Acoustics*, vol. 26, no. 2, pp. 33–38.
55. Hunaidi, O, Wang, A, Bracken, M, Gambino, T & Fricke, C 2000, *Detecting leaks in plastic pipes*, National Research Council Canada, Ottawa.

56. IPCC 2021, *Climate change 2021: the physical science basis*, Intergovernmental Panel on Climate Change, Geneva.
57. Ismail, Z, Karim, O & Ahmed, S 2019, 'A review of water leakage detection and localization methods', *Journal of Hydrology*, vol. 578, p. 124110.
58. ISO 4064-1:2014 2014, *Water meters for cold potable water and hot water—Part 1: Metrological and technical requirements*, International Organization for Standardization, Geneva.
59. ISO/IEC 17025:2017 2017, *General requirements for the competence of testing and calibration laboratories*, International Organization for Standardization / International Electrotechnical Commission, Geneva.
60. James, G, Witten, D, Hastie, T & Tibshirani, R 2013, *An introduction to statistical learning: with applications in R*, Springer, New York.
61. Joseph, K, Sharma, AK & van Staden, R 2022, 'Development of an Intelligent Urban Water Network System', *Water*, vol. 14, p. 1320.
62. Joseph, K, Sharma, AK, van Staden, R, Wasantha, PLP, Cotton, J & Small, S 2023, 'Application of software and hardware-based technologies in leaks and burst detection in water pipe networks: a literature review', *Water*, vol. 15, no. 11, p. 2046, doi:10.3390/w15112046.
63. Joseph, K, Shetty, J, Sharma, AK, van Staden, R, Wasantha, PLP & Small, S 2024, 'Integrated hardware- and software-framework for Intelligent Water Networks', *Water*, vol. 16, p. 1732.
64. Joseph, K, Shetty, J, Sharma, AK, van Staden, R, Liyanage, WP, Small, S & Bennett, N 2024, 'Leak and burst detection in water distribution network using logic- and machine-learning-based approaches', *Water*, vol. 16, no. 11, p. 1935, doi:10.3390/w16111935.
65. Kalinina, I, Gozhyi, A, Bidyuk, P, Gozhyi, V, Korobchynskyi, M & Nadraga, V 2025, 'A systematic approach to data normalization and standardization in machine learning problems', in S Babichev & V Lytvynenko (eds), *Lecture notes in data engineering, computational intelligence, and decision-making*, vol. 244, Springer, Cham, pp. 206–219.
66. Kammoun, M, Kammoun, A & Abid, M 2022, 'Leak detection methods in water distribution networks: a comparative survey on artificial intelligence applications', *Journal of Pipeline Systems Engineering and Practice*, vol. 13, no. 3, 04022024, doi:10.1061/(ASCE)PS.1949-1204.0000633.
67. Kang, D & Lansey, K 2009, 'Automated detection of pipe bursts in water distribution systems', *Journal of Water Resources Planning and Management*, vol. 135, no. 6, pp. 528–537.
68. Kang, S, Yoon, H & Lee, S 2021, 'Pipeline leak detection using unmanned aerial vehicles and thermal infrared sensors', *Sensors*, vol. 21, no. 9, p. 3010.

69. Kapelan, Z, Savic, DA & Walters, GA 2003, 'Optimal sensor placement for leak detection and location in water distribution systems', *Journal of Water Resources Planning and Management*, vol. 129, no. 6, pp. 472–484.
70. Karadirek, IE, Kaya-Basar, E & Akdeniz, T 2024, 'A study on pipe failure analysis in water distribution systems using logistic regression', *Journal of Water Supply: Research and Technology—AQUA*, vol. 24, no. 3, pp. 176–186, doi:10.2166/aqua.2024.028.
71. Kim, D, Jeong, H & Lee, J 2018, 'Ultrasonic flow meter for leak detection in water distribution systems', *Measurement Science and Technology*, vol. 29, no. 8, p. 085301.
72. Komba, GM, Mathonsi, TE & Owolawi, PA 2024, 'Optimizing leak detection and location in water distribution networks using SVM–RF algorithm', in *Proceedings of the 15th International Conference on Mechanical and Intelligent Manufacturing Technologies (ICMIMT)*, IEEE, New York, pp. [add pages], Cape Town, South Africa, 17–19 May 2024.
73. Kong, Q, Li, J, Song, G & Mo, Y 2018, 'Leakage detection of water pipelines using infrared thermography', *Automation in Construction*, vol. 89, pp. 64–74.
74. Kraid, L, Saleh, H & Hameed, M 2021, 'Oil pipeline leak detection and monitoring: a review', *Energy Reports*, vol. 7, pp. 154–172.
75. Krause, A & Guestrin, C 2007, 'Near-optimal observation selection using submodular functions', in *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence*, AAAI Press, Vancouver, BC, pp. 1650–1654.
76. Krause, A, Leskovec, J, Guestrin, C, VanBriesen, J & Faloutsos, C 2008, 'Efficient sensor placement optimization for securing large water distribution networks', *Journal of Water Resources Planning and Management*, vol. 134, no. 6, pp. 516–526, doi:10.1061/(ASCE)0733-9496(2008)134:6(516).
77. Krause, A, Singh, A & Guestrin, C 2008, 'Near-optimal sensor placements in Gaussian processes: theory, efficient algorithms and empirical studies', *Journal of Machine Learning Research*, vol. 9, pp. 235–284.
78. Kulkarni, P & Farnham, C 2016, 'Smart city environment management framework for urban sustainability', *Energy Procedia*, vol. 104, pp. 103–108.
79. Lambert, A & Hirner, W 2000, *Losses from water supply systems: standard terminology and recommended performance measures*, IWA Publishing, London.
80. Lambert, A, Brown, T, Takizawa, M & Weimer, D 2014, 'A review of performance indicators for real losses from water supply systems', *Journal of Water Supply: Research and Technology—AQUA*, vol. [add vol.], no. [add no.], pp. [add pages].

81. Li, R, Xu, Q, Zhang, Z & Liu, Y 2015, 'Leak detection and localization in water distribution systems using data-driven approaches', *Procedia Engineering*, vol. 119, pp. 147–156.
82. Li, S, Bao, X, Chen, L & Zhang, C 2019, 'Pipeline leak detection using distributed optical fiber acoustic sensing', *IEEE Sensors Journal*, vol. 19, no. 23, pp. 11072–11081.
83. Li, Z, Wang, J, Yan, H, Li, S, Tao, T & Xin, K 2022, 'Fast detection and localization of multiple leaks in water distribution network jointly driven by simulation and machine learning', *Journal of Water Resources Planning and Management*, vol. 148, no. 5, 05022005, doi:10.1061/(ASCE)WR.1943-5452.0001504.
84. Li, Z, Sun, Y & Zhao, L 2025, 'Machine learning approaches for leak detection: a systematic review', *Engineering Applications of Artificial Intelligence*, vol. [add vol.], pp. [add pages/article number].
85. Liemberger, R & Wyatt, A 2019, 'Quantifying the global non-revenue water problem', *Water Supply*, vol. 19, no. 3, pp. 831–837, doi:10.2166/ws.2018.129.
86. Liu, F, Ting, KM & Zhou, Z-H 2008, 'Isolation forest', in *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, IEEE, Pisa, pp. 413–422.
87. Liu, F, Ting, KM & Zhou, Z-H 2012, 'Isolation-based anomaly detection', *ACM Transactions on Knowledge Discovery from Data*, vol. 6, no. 1, article 3.
88. Liu, Y 2021, 'Pipeline transportation in the modern energy industry: trends and challenges', *Energy Policy*, vol. 149, p. 112055.
89. Liu, Y, Ma, X, Li, Y, Tie, Y, Zhang, Y & Gao, J 2019, 'Water pipeline leakage detection based on machine learning and wireless sensor networks', *Sensors*, vol. 19, no. 22, p. 5086, doi:10.3390/s19225086.
90. MacQueen, J 1967, 'Some methods for classification and analysis of multivariate observations', in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley, CA, pp. 281–297.
91. Marney, D & Sharma, AK 2012, 'Smart water management for cities: addressing challenges through innovation', *Water Practice and Technology*, vol. 7, no. 4, pp. 1–8.
92. Mashford, J, De Silva, D, Marney, D & Burn, S 2009, 'An approach to leak detection in pipe networks using analysis of monitored pressure values by support vector machine', *Mathematical and Computer Modelling*, vol. 49, no. 7–8, pp. 1903–1917.
93. Mays, LW 2010, *Ancient water technologies*, Springer, Dordrecht.

94. McMillan, L, Fayaz, J & Varga, L 2023, 'Machine-learning-based health monitoring and leakage management of water distribution systems', in *Proceedings of the International Conference on Evolving Cities*, Evolving Cities Publications, Southampton, UK.
95. Meseguer, J, Pérez, R, Mirats-Tur, JM, Puig, V & Quevedo, J 2015, 'Model-based fault diagnosis in water distribution networks using interval LPV models', *Journal of Hydroinformatics*, vol. 17, no. 5, pp. 771–787.
96. Meseguer, J, Puig, V, Escobet, T & Quevedo, J 2023, 'Leak detection and localization in water distribution networks: review and perspective', *Annual Reviews in Control*, vol. [add vol.], pp. [add pages].
97. Miah, MS, Islam, MT & Islam, MS 2020, 'Distributed optical fiber sensing systems for pipeline monitoring: a review', *Sensors*, vol. 20, no. 21, p. 6204.
98. Misiunas, D 2005, *Failure monitoring and asset condition assessment in water supply systems*, PhD thesis, Lund University, Sweden.
99. Misiunas, D, Vítkovský, JP, Olsson, G, Simpson, AR & Lambert, MF 2005, 'Pipeline burst detection and location using pressure transient analysis', *Journal of Water Resources Planning and Management*, vol. 131, no. 4, pp. 316–325.
100. Mitchell Plumbing and Gas 2021, *The difference between leaking and burst pipes*, viewed 12 December 2023, <https://www.mitchellplumbinggas.com/blog-post/the-difference-between-leaking-and-burst-pipes>.
101. Montgomery, DC 2019, *Introduction to statistical quality control*, 8th edn, Wiley, Hoboken, NJ.
102. Mounce, SR & Boxall, JB 2010, 'Implementation of an on-line artificial intelligence system for detection of bursts and other abnormal flows', *Journal of Water Resources Planning and Management*, vol. 136, no. 1, pp. 11–19.
103. Mounce, SR, Boxall, JB & Machell, J 2010, 'Development and verification of an online artificial intelligence system for detection of bursts and other abnormal flows', *Journal of Water Resources Planning and Management*, vol. 136, no. 3, pp. 309–318.
104. Mounce, SR, Mounce, R & Boxall, JB 2003, 'Real-time burst detection in water distribution systems using artificial neural networks and self-organising maps', *Proceedings of the Institution of Civil Engineers—Water Management*, vol. 156, pp. 139–144.

105. Muggleton, JM, Brennan, MJ & Gao, Y 2011, 'Determining the location of leaks in buried water pipes using acoustic methods', *Proceedings of the Institution of Mechanical Engineers, Part C: Journal of Mechanical Engineering Science*, vol. 225, no. 8, pp. 1951–1962.
106. Murphy, KP 2012, *Machine learning: a probabilistic perspective*, MIT Press, Cambridge, MA.
107. Mutchek, M & Williams, E 2014, 'Moving towards smart water grids: a survey of existing technologies and research directions', *Water Resources Management*, vol. 28, no. 10, pp. 3725–3750.
108. Nemhauser, GL, Wolsey, LA & Fisher, ML 1978, 'An analysis of approximations for maximizing submodular set functions', *Mathematical Programming*, vol. 14, no. 1, pp. 265–294.
109. Ochoa-Rodriguez, S, Wang, L, Gaitan, M & Velasco, M 2019, 'Satellite remote sensing for water loss detection', *Remote Sensing Applications: Society and Environment*, vol. 13, pp. 1–12.
110. Ostfeld, A, Uber, JG, Salomons, E, Berry, JW, Hart, WE, Phillips, CA, Watson, J-P, Dorini, G, Jonkergouw, P, Kapelan, Z, Di Pierro, F, Khu, S-T, Savic, D, Eliades, D, Polycarpou, M, Ghimire, SR, Barkdoll, BD, Gueli, R, Huang, JJ & McBean, EA 2008, 'The battle of the water sensor networks (BWSN): a design challenge for engineers and algorithms', *Journal of Water Resources Planning and Management*, vol. 134, no. 6, pp. 556–568, doi:10.1061/(ASCE)0733-9496(2008)134:6(556).
111. Pagano, A, Garlisi, D, Tinnirello, I, Giuliano, F, Garbo, G, Falco, M & Cuomo, F 2025, 'A survey on massive IoT for water distribution systems: challenges and large-scale deployment', *Computer Networks*, vol. [add vol.], pp. [add pages/article number].
112. Palau, C, Arregui, F, Cabrera, E & Cobacho, R 2012, 'Burst detection in water networks using pressure signals', *Journal of Water Resources Planning and Management*, vol. 138, no. 1, pp. 47–54.
113. Perelman, L, Krapivka, A & Ostfeld, A 2013, 'Single and multi-objective optimal design of water distribution systems: application to the case study of the Hanoi system', *Journal of Hydroinformatics*, vol. 15, no. 2, pp. 501–516.
114. Puust, R, Kapelan, Z, Savic, DA & Koppel, T 2010, 'A review of methods for leakage management in pipe networks', *Urban Water Journal*, vol. 7, no. 1, pp. 25–45, doi:10.1080/15730620903447619.
115. Quinlan, JR 1986, 'Induction of decision trees', *Machine Learning*, vol. 1, no. 1, pp. 81–106.
116. Rajabi, MM, Komeilian, P, Wan, X & Farmani, R 2023, 'Leak detection and localization using deep learning models in water distribution systems', *Water Research*, vol. 241, article [add article number].

117. Ramos, H, Covas, D, Sá Marques, A & Loureiro, D 2019, 'Smart water networks and smart monitoring: state of the art and future trends', *Water Resources Management*, vol. 33, no. 8, pp. 2875–2893.
118. Rathi, S & Ingeduld, P 2013, 'Comparative study of leak detection technologies for water distribution networks', *Water SA*, vol. 39, no. 4, pp. 491–500.
119. Rathore, S, Kumar, S & Singh, A 2018, 'Helium gas leak detection techniques in buried pipelines', *Journal of Natural Gas Science and Engineering*, vol. 57, pp. 169–180.
120. Ren, H, Dong, H, Zhao, Y & Li, J 2022, 'Advances in smart water systems', *Water Research*, vol. 215, p. 118230.
121. Rodgers, JL & Nicewander, WA 1988, 'Thirteen ways to look at the correlation coefficient', *The American Statistician*, vol. 42, no. 1, pp. 59–66.
122. Rodriguez, L, Fernandez, C, Pantano, N, Scaglia, G & Keesman, KJ 2025, 'Optimizing sensor placement for enhanced observability in water distribution networks', *Journal of Hydroinformatics*, vol. 27, no. [add no.], pp. [add pages].
123. Romano, M, Kapelan, Z & Savic, D 2014, 'Automated detection of pipe bursts and other events in water distribution systems', *Journal of Water Resources Planning and Management*, vol. 140, no. 4, pp. [add pages/article number], doi:10.1061/(ASCE)WR.1943-5452.0000348.
124. Rossman, LA 2000, *EPANET 2 user manual*, US Environmental Protection Agency, Cincinnati.
125. Rossman, LA 2020, *EPANET 2.2 user manual*, US Environmental Protection Agency, Cincinnati, viewed [add date if needed], <https://www.epa.gov/water-research/epanet>.
126. Rostirolla, G, da Silva, J, Souza, R, Vieira, A & Bragança, A 2017, 'A smart urban water management system: architecture and implementation', *Procedia Computer Science*, vol. 109, pp. 1089–1094.
127. Sapkota, M, Sharma, AK, Rahman, A & Shrestha, S 2021, 'Urban water management and leakage reduction: current status and future directions', *Water*, vol. 13, no. 5, p. 635.
128. Savic, DA, Kapelan, Z & Jonkergouw, P 2009, 'Quo vadis water distribution model calibration?', *Urban Water Journal*, vol. 6, no. 1, pp. 3–22.
129. Schölkopf, B, Platt, JC, Shawe-Taylor, J, Smola, AJ & Williamson, RC 2001, 'Estimating the support of a high-dimensional distribution', *Neural Computation*, vol. 13, no. 7, pp. 1443–1471.

130. Schubert, E, Sander, J, Ester, M, Kriegel, H-P & Xu, X 2017, 'DBSCAN revisited, revisited: why and how you should (still) use DBSCAN', *ACM Transactions on Database Systems*, vol. 42, no. 3, article 19.
131. Seiffert, C, Khoshgoftaar, TM, Van Hulse, J & Napolitano, A 2010, 'RUSBoost: a hybrid approach to alleviating class imbalance', *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, vol. 40, no. 1, pp. 185–197.
132. Sharma, AK & Marney, D 2012, 'The application and utility of “smarts” for monitoring water and its infrastructure', *Water Journal of the Australian Water Association*, vol. 39, no. 1, pp. 86–92.
133. Sharma, AK, Burn, S, Gardner, T, Gregory, A & Tjandraatmadja, G 2013, 'Smart water systems for sustainable urban water management', *Water Practice and Technology*, vol. 8, no. 4, pp. 1–14.
134. Shaw, R, Surjan, A, Parvin, GA & Hossain, M 2012, *Disaster risk reduction approaches in Bangladesh*, Springer, Tokyo.
135. Shen, Y & Cheng, W 2022, 'A tree-based machine learning method for pipeline leakage detection', *Water*, vol. 14, no. 18, p. 2833, doi:10.3390/w14182833.
136. Shiddiq, F, Arifin, M, Zainal, M & Nugraha, D 2021, 'IoT-based monitoring system for water pipeline leakage using ultrasonic sensors', *IOP Conference Series: Materials Science and Engineering*, vol. 1073, p. 012012.
137. Soldevila, A, Mounce, SR, Boxall, JB & Makropoulos, C 2016, 'Leak detection and localisation in water distribution systems using machine learning', *Water Research*, vol. 102, pp. 76–88.
138. Soldevila, A, Nejari, F, Puig, V & Quevedo, J 2016, 'Sensor placement for leak detection using the information matrix in water distribution networks', *Procedia Engineering*, vol. 119, pp. 1077–1086.
139. Soldevila, A, Tornil-Sin, S, Puig, V, Escobet, T & Quevedo, J 2016, 'Leak detection and localization in water distribution networks using a Bayesian classifier', *Journal of Process Control*, vol. 39, pp. 59–72, doi:10.1016/j.jprocont.2016.01.010.
140. Sorbello, J & Marck, R 2014, 'Applications for intelligent water network systems', *Water Journal of the Australian Water Association*, vol. 41, no. 1, pp. 75–80.
141. Sourabh, N, Timbadiya, P & Patel, P 2023, 'Leak detection in water distribution network using machine learning techniques', *ISH Journal of Hydraulic Engineering*, vol. 29, suppl. 1, pp. 177–195, doi:10.1080/09715010.2023.2171072.

142. Stair, R & Reynolds, G 2012, *Fundamentals of information systems*, Nelson Education, Scarborough, ON.
143. Thompson, JR, Mounce, SR, Boxall, JB & Machell, J 2016, 'Automatic detection of leaks and bursts in water distribution systems using artificial intelligence', *Urban Water Journal*, vol. 13, no. 4, pp. 408–420.
144. Thornton, J & Lambert, A 2006, *Managing leakage and pressure: a practical approach*, IWA Publishing, London.
145. Tijani, IA, Abdelmageed, S, Fares, A, Fan, KH, Hu, ZY & Zayed, T 2022, 'Improving the leak detection efficiency in water distribution networks using noise loggers', *Science of the Total Environment*, vol. 821, 153530, doi:10.1016/j.scitotenv.2022.153530.
146. Todini, E & Pilati, S 1988, 'A gradient method for the analysis of pipe networks', in B Coulbeck & CH Orr (eds), *Computer applications in water supply*, vol. 1, Research Studies Press, Letchworth, pp. 1–20.
147. Uciński, D 2005, *Optimal measurement methods for distributed parameter system identification*, CRC Press, Boca Raton, FL.
148. Ullah, I, Faheem, M & Kim, BS 2017, 'Energy and congestion-aware routing metric for smart grid AMI networks in smart city', *IEEE Access*, vol. 5, pp. 13799–13810.
149. UNESCO 2020, *United Nations world water development report 2020: water and climate change*, UNESCO, Paris.
150. Wang, J, Li, F, Liu, Q, Xu, H & Zhang, Z 2021, 'Big data-based intelligent management of oil pipeline transportation systems', *Journal of Petroleum Science and Engineering*, vol. 205, p. 108839.
151. Wang, X, Liu, Y, Zhang, H & Chen, J 2024, 'Graph neural network-based leak detection in water distribution systems', *Water*, vol. 16, no. [add no.], p. [add article number].
152. World Health Organization 2017, *Guidelines for drinking-water quality*, 4th edn, WHO, Geneva.
153. Wu, J, Ma, D & Wang, W 2022, 'Leakage identification in water distribution networks based on XGBoost algorithm', *Journal of Water Resources Planning and Management*, vol. 148, no. 2, 04021107, doi:10.1061/(ASCE)WR.1943-5452.0001433.
154. Wu, ZY, Sage, P, Turtle, D & Zhang, T 2016, 'Smart monitoring for early burst detection in water distribution systems', *Procedia Engineering*, vol. 154, pp. 163–170.
155. Yazdani, A & Jeffrey, P 2011, 'Complex network analysis of water distribution systems', *Chaos*, vol. 21, no. 1, p. 016111.

156. Yussif, AM & Zayed, T 2023, 'Application of machine learning for leak localisation in water supply networks', *Buildings*, vol. 13, no. 4, p. 849, doi:10.3390/buildings13040849.
157. Zhang, J 1997, 'Pipeline leak detection technology and systems: state of the art', *Hydrocarbon Processing*, vol. 76, no. 4, pp. 41–47.
158. Zhang, J, Hoffman, A, Murphy, K, Lewis, J & Twomey, M 2013, 'Review of pipeline leak detection technologies', in *PSIG Annual Meeting*, OnePetro, Richardson, TX, Prague, 16–19 April.
159. Zhou, B, Lau, V & Wang, X 2019, 'Machine-learning-based leakage-event identification for smart water supply systems', *IEEE Internet of Things Journal*, vol. 7, no. 3, pp. 2277–2292.
160. Zhou, D, Hu, X, Zhang, Y, Wang, J & Zhang, H 2020, 'Pressure-sensor-based deep learning for burst detection in water networks', *Journal of Hydroinformatics*, vol. 22, no. 5, pp. 1019–1033.

Chapter 10

Appendices

10.1 Appendices for Chapter 1

10.1.1 Appendix 1A – Key Terminology and Abbreviations

1. IWN – Intelligent Water Network
2. SCADA – Supervisory Control and Data Acquisition
3. IoT – Internet of Things
4. ML – Machine Learning
5. GIS – Geographic Information System

10.1.2 Appendix 1B – Figures Referenced in Chapter 1

Figure 1.1: Illustration of (a) pipe leakage (R.S. Plumbing, 2024) and (b) pipe burst (Water Leak Detection, 2020).

Figure 1.2: Conceptual plan for an Intelligent Water Network system (adopted from Kiran Joseph et al., 2022).

10.1.3 Appendix 1C – Key Parameters Referenced in the Conceptual Model

1. Flow rate (Q)
2. Pressure (P)
3. Pump speed (N)
4. Timestamp (t)
5. Sensor location index (S_i)

10.1.4 Appendix 1D – Study Region Information

1. Study area: Victorian Sunbury region, Australia

2. Data sources: SCADA datasets containing flow, pressure, and pump speed data from local water utility facilities
3. Purpose: To validate the proposed IWN framework and Machine Learning model under controlled experimental conditions

10.1.5 Appendix 1E – Summary of Chapter 1 Structure

1. Introduced the concept and importance of IWN in addressing leak and burst challenges.
2. Defined the research aim and objectives for developing a data-driven detection framework.
3. Outlined the scope of the study and the three-layer IWN architecture.
4. Presented the research significance and innovations linking Machine Learning and SCADA systems for real-time monitoring.

10.2 Appendices for Chapter 2

10.2.1 Glossary of Terms (Leak/Burst & IWN)

Leak: Gradual, unintended water loss through small defects (cracks, pinholes, joint failures); typically, persistent, low–moderate flow.

Burst: Sudden pipe failure producing large, rapidly changing discharge; often transient-rich pressure signatures.

Non-Revenue Water (NRW): Water produced but not billed; sum of real losses (leaks/bursts), apparent losses (metering, theft), and authorised unbilled consumption. SCADA / AMI / IoT: Supervisory Control and Data Acquisition; Advanced Metering Infrastructure; Internet of Things.

IWN: Intelligent Water Network — multi-layer system integrating sensing, communications, analytics, and decision support.

DMA: District Metered Area; hydraulically partitioned network zone with measured inflow(s).

Pressure Management (PM): Control via PRVs, PATs, pump scheduling to stabilise/optmise pressures and reduce leakage.

Observability (network): Ability to infer internal network states/events (e.g., leaks) from available sensor measurements.

Minimum Night Flow (MNF): Low-demand window analysis used to screen for leakage.

Transient: Rapid pressure waves due to events (valve/pump operations, bursts).

Sensor Placement (OSP): Optimal Sensor Placement choose locations that maximise detection/localisation under constraints.

10.2.2 Acronyms

AMI – Advanced Metering Infrastructure

ANN – Artificial Neural Network

AUC – Area Under the Receiver Operating Characteristic Curve

BPNN – Back-Propagation Neural Network

CNN – Convolutional Neural Network

DAS – Distributed Acoustic Sensing

DMA – District Metered Area

DSP – Digital Signal Processing

DTS – Distributed Temperature Sensing

EPANET – Environmental Protection Agency Network Hydraulic Modelling Software

F1 – F1-Score (harmonic mean of precision and recall)

GMM – Gaussian Mixture Model

GPR – Ground-Penetrating Radar

HMM – Hidden Markov Model

IoT – Internet of Things

IRT – Infrared Thermography

ITA – Inverse Transient Analysis

IWA – International Water Association

IWN – Intelligent Water Network

KF – Kalman Filter

k-NN – k-Nearest Neighbour
 LSTM – Long Short-Term Memory (neural network)
 MNF – Minimum Night Flow
 MOEA – Multi-Objective Evolutionary Algorithm
 MOPSO – Multi-Objective Particle Swarm Optimisation
 MSE – Mean Squared Error
 NB-IoT – Narrowband Internet of Things
 NRW – Non-Revenue Water
 NSGA-II – Non-Dominated Sorting Genetic Algorithm II
 OSP – Optimal Sensor Placement
 PCA – Principal Component Analysis
 PRV – Pressure Reducing Valve
 RNN – Recurrent Neural Network
 RL – Reinforcement Learning
 RF – Random Forest
 SAR – Synthetic Aperture Radar
 SCADA – Supervisory Control and Data Acquisition
 SNR – Signal-to-Noise Ratio
 SVM – Support Vector Machine
 VAE – Variational Autoencoder
 WDN – Water Distribution Network
 WSN – Wireless Sensor Network
 XGBoost – Extreme Gradient Boosting Algorithm

10.2.3 Notation and Units

p = pressure [m or kPa];

Q = flow [L/s or m³/h]; H = head [m];

D = diameter [mm]; L = length [m].

ρ = density [kg/m³]; g = gravity [m/s²];

C = discharge/emitter coefficient.

Time base: UTC+10/11 (Australia/Melbourne). Sampling intervals noted per sensor (e.g., 1 s, 1 min, 15 min). Demand $d_i(t)$ per node i [L/s]; Leakage discharge q_ℓ [L/s].

10.2.4 Hydraulic and Leak Modelling Essentials

Darcy–Weisbach: $\Delta H = f \cdot (L/D) \cdot (V^2/2g)$, with $V = Q/A$ and f via Colebrook or Swamee–Jain. Hazen–Williams (pressurised water): $\Delta H = 10.67 \cdot L \cdot Q^{1.852} / (C_H^{1.852} \cdot D^{4.871})$. EPANET emitter (orifice) model: $q_{\ell,i} = C_{e,i} \cdot p_i^m$ with typical m in range 0.5–1.5. Inverse Transient Analysis (ITA): minimise $J(\theta) = \sum (p_{\text{meas}} - p_{\text{sim}})^2$; θ encodes leak locations/sizes.

10.2.5 Data Schemas and Tagging Templates

Pressure/flow SCADA tag schema (tag_id, asset_id, type, unit, sample_rate, qc_flag, location_xy, elevation_m).

Event log schema (event_id, start_ts, end_ts, class, notes).

10.2.6 Data Quality and Pre-Processing Checklist

Time alignment, unit harmonisation, spike removal, gap filling, drift checks, demand seasonality separation, and sensor health KPIs (uptime %, SNR, drop-rate, latency).

10.2.7 Literature Screening and Reproducibility Protocol

Databases: Scopus, Web of Science, IEEE Xplore, MDPI, Elsevier, IWA. Inclusion: peer-reviewed; quantitative performance; field or realistic bench studies. Exclusion: non-distribution or purely theoretical work.

Extraction fields: network scale, sensing, sampling rate, algorithm, metrics, cost. Maintain PRISMA-style CSV of all studies.

10.2.8 Evaluation Metrics (Formulas)

Precision = $TP / (TP + FP)$; Recall = $TP / (TP + FN)$; $F1 = 2 \cdot (P \cdot R) / (P + R)$;

Accuracy = $(TP + TN) / (\text{Total})$.

Detection delay = $t_{\text{alarm}} - t_{\text{onset}}$;

Localisation error = pipe distance between predicted and true.

AUC-ROC: probability positive ranks above a negative.

10.2.9 Sensor Placement & Observability Framework (Optional but Relevant)

Sensor observability matrix (S): quantifies sensitivity of node pressures to leak scenarios.

Distinguishability metric:

$$d_{s_1, s_2} = ||S_{\Omega, s_1} - S_{\Omega, s_2}||_2$$

Objective: maximise leak-detection coverage while minimising cost and redundancy.

10.2.10 Data Quality Criteria Table (optional visual summary)

Check	Description	Tool / Method	Threshold
Time sync	Compare timestamps	SCADA audit	< 1 s drift
Signal noise	Rolling STD / FFT	Python QC	SNR > 20 dB
Missing data	NaN ratio	Pandas check	< 5 % gaps
Calibration	Compare lab vs field	ISO 17025	$\leq \pm 1$ % error

10.2.11 Units and Symbol

1. ν (*kinematic viscosity*) = 1.004×10^{-6} m²/s (at 20 °C)
2. ψ (*leakage exponent*) ≈ 0.5 – 1.2 (material dependent)
3. C^H (*Hazen–Williams coefficient*) ≈ 100 – 150 (depending on roughness)

10.2.12 References to Standards and Datasets

1. ISO 4064-1 (2014) – Water meters for cold/hot potable water
2. ISO/IEC 17025 (2017) – Testing and calibration laboratory competence
3. EPA (2020) – EPANET 2.2 user manual (if used in simulations)

10.2.13 Version Control and Reproducibility Note

All datasets, Python notebooks, and EPANET input files are version-controlled using GitHub/OneDrive; metadata follow FAIR (Findable, Accessible, Interoperable, Reusable) guidelines to ensure reproducibility.

10.2.14 Optional Figures

1. Figure 2A.1 – Hydraulic modelling workflow.
2. Figure 2A.2 – Signal processing pipeline for leak detection.
3. Figure 2A.3 – Schematic of IWN layers (Sensors → SCADA → Analytics → Decision).

10.3 Appendices for Chapter 3

10.3.1 Appendix 3A – Case Study Sites

Sunbury pumping main (real network):

1. Location: Sunbury, Victoria (Maribyrnong catchment).
2. Conveyance: Trunk main to Shepherds Lane tank via 450 mm and 375 mm pipelines.
3. Instrumentation on each pipeline: Discharge pressure sensor and flow meter; divide valve between lines.
4. Selected 450 mm line details: 6.24 km length; tank capacity 0.8 ML; max water level 3.72 m; elevation at pump station and tank 59 m amsl.
5. Data types: flow, pressure, pump speed (SCADA).

Melton Training Yard (experimental prototype):

1. Pipe: 110 mm ID polyethylene, 110 m length, connected to a 100 mm pipeline.
2. Sensors: F1 (inlet EM flow), F2 (outlet EM flow), P1 (inlet pressure), P2 (outlet pressure).
3. Leak simulation: outlets 20 mm and 25 mm from the 110 mm line.
4. Telemetry: SCADA wireless; data recorded every 60 s (sampling freq ≈ 0.0167 Hz).

Hanoi WDN (benchmark model):

1. Topology: 32 nodes, 34 pipes, 3 loops; reservoir elevation 100 m.
2. Total pipe length 39.420 km; minimum pressure head 30 m.
3. Diameters considered: 304.8–1,016.0 mm; pipe costs \$45.73–\$278.30/m.
4. Use in thesis: EPANET simulations for sensitivity analysis and optimal pressure-sensor placement.

10.3.2 Appendix 3B – Figures Referenced in Chapter 3

1. Figure 3.1: Map of the Sunbury case-study location (Victoria, Australia).
2. Figure 3.2: Overall flow diagram of leak/burst detection workflow.
3. Figure 3.3: Simplified schematic of the 450 mm pipeline to tank.
4. Figure 3.4: Panoramic photo—Loemans Road Pump Station (L) and Shepherds Lane tank (R).
5. Figure 3.5: Simplified diagram of the Melton experimental site (F1, F2, P1, P2; leak points).
6. Figure 3.6: Hanoi WDN in EPANET (supply at Node 1).

10.3.3 Appendix 3C— Instrumentation Specifications

Pressure Transducer (Measurex MX232, S/N 3115262003):

1. Range 0–16 bar (~0–163 m). Accuracy $\pm 0.25\%$ FS.
2. PTFE-sealed installation; factory-calibrated; validated vs certified gauge.

Pressure Transmitter (ABB 261GSU/VT1/N21):

1. Silicon-oil Hastelloy C276 diaphragm; span 150 kPa.
2. Output 4–20 mA, supply 10.5–32 V DC. Accuracy $\pm 0.1\%$ span (up to $\pm 0.04\%$ in high-accuracy mode).
3. Field calibrated with NIST-traceable calibrator.

Mechanical Flow Meter (Zenner multijet dry dial, S/N 2404013684):

1. ISO 4064 compliant; pulse output. Accuracy $\pm 2\%$ (above transition), $\pm 5\%$ (below).
2. Factory-calibrated; sealed; prelim verification performed.

Electromagnetic Flow Meter (Siemens SITRANS FM MAG 8000):

1. Battery-powered; factory-calibrated to ISO/IEC 17025. Accuracy $\pm 0.4\%$ (up to $\pm 0.2\%$ on certain models).
2. SENSORPROM™ config retention; data retrieval via display/pulse/SCADA.
3. SCADA logging every 60 s (≈ 0.0167 Hz).

10.3.4 Appendix 3D– SCADA Tags and Units Used (Chapter 3 notation)

1. F1 – Inlet flow (Melton), unit: L/s (via EM flow meter).
2. F2 – Outlet flow (Melton), unit: L/s (via EM flow meter).
3. P1 – Inlet pressure (Melton), unit: m or kPa (per device native units).
4. P2 – Outlet pressure (Melton), unit: m or kPa.
5. PumpSpeed – Sunbury pump rotational/drive speed (unit per SCADA tag).
6. Timestamp (t) – SCADA record time, interval 60 s.

(Units are recorded exactly as per the devices/SCADA historian described in Chapter 3.)

10.3.5 Appendix 3E– Derived aariables Referenced in Chapter 3

1. ΔP – Change in pressure between sensors or time steps.
2. Flow acceleration – ($\frac{dQ}{dt}$) computed from SCADA flow series.
3. Event windows – Leak/burst scenarios triggered by simulated outlets (20 mm, 25 mm).

10.3.6 Appendix 3F– Hanpi WDN Benchmark Parameters (Chapter 3 tables)

1. Counts and constraints: 32 nodes, 34 pipes, 3 loops, min head 30 m.
2. Reservoir: Node 1 at 100 m elevation; uniform nodal demands (per benchmark).
3. Tables referenced: Table 3.1 (network characteristics), Table 3.2 (nodal elevation/base demand excerpt and pipe connectivity as provided).

10.3.7 Appendix 3G– Chapter Specific Notes

1. Sunbury: 450 mm line used for detailed analysis; tank 0.8 ML, max WL 3.72 m, elevation 59 m; distance 6.24 km.
2. Melton: Instrument layout F1/F2/P1/P2; SCADA interval 60 s; leaks at 20 mm and 25 mm.
3. Hanoi: EPANET used for leak simulations and pressure-sensitivity matrix to guide optimal pressure-sensor placement.

10.4 Appendices for Chapter 4

10.4.1 Appendix 4A – Figures for Research Methodology

Figure 4.1 – Variation of pressure head (m) with respect to time (minutes) for the 450 mm diameter pumping main.

Figure 4.2 – (a) Estimation of delta pressure (ΔP) and (b) delta discharge (ΔQ) for case-study system.

Figure 4.3 – Flow diagram of the leak and burst detection algorithm.

Figure 4.4 – Leak location at X (simple system diagram).

Figure 4.5 – Developed Machine Learning approach for leak and burst detection.

Figure 4.6 – Research-methodology framework integrating data acquisition, pre-processing, logic-based detection, and Machine Learning analysis (Balanced ROS, SMOTE, ADASYN) with validation and real-time decision-support.

10.4.2 Appendix 4B – Symbols and Variables Used

ΔP – Pressure variation (baseline oscillation band).

ΔQ – Flow variation (baseline oscillation band).

ΔS – Pump speed variation.

L – Pipe length (m).

D – Pipe diameter (m).

Q – Flow rate (m^3/s).

g – Acceleration due to gravity.

f – Darcy–Weisbach friction factor.

ε – Pipe wall roughness height.

ν – Kinematic viscosity of water.

T – Water temperature ($^{\circ}\text{C}$).

P_x – Pressure at sensor at time of leak.

E_d – Elevation difference between pumping main and service tank.

L_x – Distance to leak/burst location along main.

10.4.3 Appendix 4C – Equations Cited (as Numbered in Chapter 4)

(4.1) Darcy–Weisbach head loss: $h_f = (8fLQ^2)/(\pi^2gD^5)$

(4.2) Swamee (friction factor): $f = 1.325 \{ \ln[\varepsilon/3.7D + 4.618(\nu D/Q)^{0.9}] \}^{-2}$

(4.3) Kinematic viscosity (Swamee): $\nu = 1.792 \times 10^{-6} [1 + (T/25)^{1.165}]^{-1}$

(4.4) Pressure–distance relation: $P_x = (E_d/L)L_x + (8fQ^2)/(\pi^2gD^5)L_x$

(4.5) Leak/burst location: $L_x = P_x / [(E_d/L) + (8fQ^2/\pi^2gD^5)]$

$$(4.6) \text{ F1 Score} = 2PR/(P+R)$$

$$(4.7) \text{ Precision} = TP/(TP+FP)$$

$$(4.8) \text{ Recall} = TP/(TP+FN)$$

$$(4.9) \text{ ROC-AUC} = \int_0^1 \text{TPR} \, d(\text{FPR})$$

10.4.4 Appendix 4D – Logic-Based Algorithm Data and Thresholds

SCADA data recorded every 1 minute (pressure, flow, pump speed).

Baseline head: start-up $\approx 58 \text{ m} \rightarrow$ steady $\approx 80 \text{ m} \pm \Delta P \rightarrow$ shutdown $\approx 58 \text{ m}$.

Pump speed: 1500 rpm at normal operation.

Leak criterion: pressure drop $> 2 \times \Delta P$.

Burst criterion: pressure drop $> 3 \times \Delta P$ within short time window.

Logic checks ΔQ and ΔS simultaneously to avoid false positives during pump transients.

10.4.5 Appendix 4E – Logic-Based Workflow (Steps 1–17)

Task 1 (1–5): Read and load CSV/YAML files; define helper functions; create lag variables; determine pump running status.

Task 2 (6–8): Compute running sum of pressure changes; check against thresholds; exclude shutdown phases.

Task 3 (9): Reset timers for on/off pump status.

Task 4 (10–12): Flag leak/burst; track duration; confirm event.

Task 5 (13–17): Estimate head loss (H_f); locate burst (L_x); email operator; trigger start/end alerts; reset flags after normalisation.

10.4.6 Appendix 4F – Parameters Included and Excluded

Included: Pressure, flow, pump speed, timestamp, pipe size, roughness coefficient, diurnal pattern.

Excluded: Temperature, soil properties, pipe deterioration, soil–water interaction.

10.4.7 Appendix 4G – Machine Learning Workflow

1. Collect $\approx 20\,000$ SCADA data points (1-min interval).

2. Pre-process (cleaning and standardisation).

3. Generate features (ΔP , ΔQ , ΔS).

4. Train and test multiple models.

5. Select highest-performing model by F1 and ROC-AUC.

6. Implement selected model for leak/burst detection.

7. Generate alerts.

10.4.8 Appendix 4H – Machine Learning Algorithms Used

Supervised: Decision Tree, AdaBoost, LDA, QDA, Gaussian Naïve Bayes, Ridge Classifier, Balanced Random Forest, RUSBoost, Voting Classifier, K-Nearest Neighbors, Random Forest, Support Vector Classifier, Logistic Regression, Multi-Layer Perceptron.

Unsupervised: K-Means, DBSCAN, One-Class SVM, Isolation Forest, Local Outlier Factor.

10.4.9 Appendix 4I – Class Imbalance Handling

Techniques: Balanced Random Oversampling (ROS), SMOTE, ADASYN.

Applied only to training folds within stratified 5-fold cross-validation.

Example: LOF F1 improved from 0.12 (AUC 0.77) → 0.69 (AUC 0.72) with ADASYN.

10.4.10 Appendix 4J – Evaluation Metrics

Anomaly F1-Score, Precision, Recall, and ROC-AUC were used for model comparison and threshold-independent performance assessment.

10.4.11 Appendix 4K – Alert Generation

Continuous monitoring of time–pressure, flow, and pump-speed.

Compares real-time data with baselines (ΔP , ΔQ).

Triggers alerts with event timestamp, estimated distance (L_x), pressure/flow snapshot, and start/end notifications.

Alerts visualised in Figure 6.1, Chapter 6.

10.4.12 Appendix 4L – Case Study Context

Field System (Sunbury): 450 mm AC pumping main, 6.24 km length, 1500 rpm, pressure 58–80 m.

Prototype System (Melton): Controlled leak/burst experiments for supervised training and validation.

Hydraulic Model: EPANET 2.2 used for simulation and validation.

10.5 Appendices for Chapter 5

10.5.1 Appendix 5A– Case Study Data Sources

Sunbury Pumping Main (Operational SCADA dataset)

1. Location: Sunbury, Victoria; 6.24 km rising main to tank.
2. Monitoring periods:
 - a) Long-term dataset: 1 May–17 September 2022 (139 days; 200,161 samples).
 - b) Short-term testing dataset: 1–6 May 2022 (5 days; 7,201 samples).
3. Instrumentation: pump discharge pressure sensor, electromagnetic flow meter (450 mm), pump-speed telemetry.
4. Recorded variables: pressure (m), flow (L/s), pump speed (rpm), leak/non-leak label.
5. Leak-event counts:
 - a) Training dataset: 91 leak samples (0.05%).
 - b) Testing dataset: 33 leak samples (0.46%).
6. Purpose: real-world operational dataset for SCADA analysis and leak/burst characterisation.

Melton Prototype System (Controlled experimental dataset)

1. Location: Melton training yard.
2. Monitoring period: 4 July–6 August 2024 (~797 hours).
3. Sensor configuration: pressure sensors (P1: inlet, P2: outlet), flow meters (F1: inlet; F2: outlet).

4. Operating state: steady non-leak conditions during monitored period.
5. Recorded variables: pressure, flow, timestamps, operator notes.
6. Purpose: baseline hydraulic behaviour and controlled reference dataset.

10.5.2 Appendix 5B– Figures Referenced in Chapter 5

1. Figure 5.1: Kernel Density Estimates of pressure, flow, and pump speed (Sunbury training dataset).
2. Figure 5.2: Scatterplots and correlation heatmap (Sunbury training dataset).
3. Figure 5.3: Scatterplots and correlation heatmap (Sunbury testing dataset).
4. Figure 5.4: Z-score anomaly detection plots for pressure, flow, and pump speed.
5. Figure 5.5: Rolling correlation (pressure–flow, 60-minute window).
6. Figure 5.6: Time-series overlay of pressure and flow during pump cycles.
7. Figure 5.7: Pressure and flow distribution plots (Melton dataset).
8. Figure 5.8: Temporal pressure and flow behaviour across the Melton monitoring period.
9. Figure 5.9: Correlation heatmap for Melton pressure and flow variables.

10.5.3 Appendix 5C– SCADA and Experimental Instrumentation

Sunbury Pumping Main Instruments

1. Pressure sensor: pump discharge pressure; integrated with SCADA (60-second interval).
2. Electromagnetic flow meter: rising main (450 mm) flow measurement.
3. Pump-speed telemetry: rotational speed (rpm), capturing full duty-cycle transitions.

Melton Prototype Instruments

1. Pressure sensors:
 - a) P1 (inlet),
 - b) P2 (inlet).
2. Flow meters:
 - a) F1 (inlet),
 - b) F2 (outlet).
3. SCADA telemetry: synchronised, continuous time-stamped readings.

10.5.4 Appendix 5D– Variables and Units Used in Chapter 5

Sunbury SCADA Variables

1. P – Pressure (m).
2. F – Flow (L/s).

3. N – Pump speed (rpm).
4. LeakLabel – Leak state (0 = non-leak, 1 = leak).
5. Timestamp – SCADA time (60-second interval).

Melton Prototype Variables

1. P1 – Inlet pressure (m).
2. P2 – Outlet pressure (m).
3. F1 – Inlet flow (L/s).
4. F2 – Outlet flow (L/s).
5. Timestamp – SCADA time.

10.5.5 Appendix 5E– Derived Variables and Analytical Constructs

1. Pressure difference: $\Delta P = P1 - P2$ (Melton) or between time steps (Sunbury).
2. Flow acceleration: $dQ/dt = Q(t) - Q(t-1)$.
3. Pressure acceleration: $dP/dt = P(t) - P(t-1)$.
4. Z-score for anomaly detection: $Z = (x - \mu) / \sigma$.
5. Rolling correlation (pressure–flow): 60-minute sliding window.
6. Event windows: leak and non-leak segments based on SCADA labels.

10.5.6 Appendix 5F– Data Tables Referenced in Chapter 5

1. Table 5.1: Extracted SCADA sample data for pressure, flow, and pump speed.
2. Table 5.2: Summary statistics for Sunbury training dataset (leak vs non-leak).
3. Table 5.3: Summary statistics for Sunbury testing dataset (leak vs non-leak).
4. Table 5.4: Summary statistics for Melton prototype dataset (P1, P2, F1, F2).

10.5.7 Appendix 5G– Chapter Specific Notes

1. Sunbury datasets show strong separation between leak and non-leak conditions in flow and pump speed.
2. Non-leak behaviour is dominated by prolonged pump-idle states (0–1 L/s, 0 rpm).
3. Leak conditions consistently show flow above 230 L/s and pump speeds near 1450–1500 rpm.
4. Testing dataset displays slightly stronger correlations due to a more stable operating window.
5. Melton dataset provides clean, noise-free baseline behaviour with strong pressure–pressure correlation ($r \approx 0.9935$).
6. Pressure–flow correlations at Melton are negative (–0.21 to –0.32), consistent with expected hydraulic behaviour.
7. Combined datasets form a complementary empirical foundation:
 - a) Sunbury = real-world operational variability.
 - b) Melton = controlled, stable baseline for calibration and validation.

10.6 Appendices for Chapter 6

All algorithm codes and files:

<https://drive.google.com/drive/folders/1RzPpwIARXnk756nqFVEFNN2oPMwx6zUc?usp=sharing>

10.6.1 Appendix 6A– Case Study Systems for Chapter 6

Sunbury Pumping Main (450 mm, 6.24 km rising main)

1. Location: Sunbury, Victoria, Australia.
2. Pipeline: 450 mm AC rising main, length 6.24 km, connected to a reservoir.
3. SCADA variables: Pressure (m), Flow (L/s), Pump speed (rpm).
4. Operational states: Pump start-up, steady operation, shutdown.
5. Sampling interval: 60 seconds (SCADA).
6. Dataset size: ~20,000 records used for logic-based application.
7. Leak categories applied:
 - a) Minor leak (<15% pressure drop)
 - b) Major leak (15–20% pressure drop)
 - c) Burst (>20% pressure drop)

Melton Prototype System (Controlled Experimental Setup)

1. Location: Greater Western Water training yard, Melton, Victoria.
2. Pipeline: 110 mm ID PE pipe, 110 m length, connected to a 100 mm main.
3. Sensors:
 - a) Pressure sensors P1, P2
 - b) Flow meters F1 (inlet), F2 (outlet)
4. Leak simulation: 20 mm and 25 mm outlets.
5. Data logging: 60 s intervals using SCADA telemetry.
6. Dataset classes: No leak, Minor leak, Major leak.
7. Purpose: Supervised ML model development and validation.

10.6.2 Appendix 6B– Figures Referenced in Chapter 6

1. Figure 6.1: Minor leakage, Major leakage, and Burst detection in the 450 mm pipeline using the logic-based approach.
2. Figure 6.2: Detection of minor, major, and burst events (logic-based outcomes).
3. Figure 6.3: Pressure–Flow correlation with anomaly markers.
4. Figure 6.4: Pressure, flow, and pump-speed time-series during pump cycles.
5. Figure 6.6: Predicted leak occurrence from ML platform.
6. Figure 6.7: Predicted burst occurrence from ML platform.
7. Figure 6.8: Flowchart of best-performing model selection.
8. Figure 6.9: Q–Q plots for inlet pressure, outlet pressure, flow ratio, pressure ratio.

9. Figure 6.10: Feature importance graph.
10. Figure 6.11 & 6.12: Leak/burst detection methodology flowcharts.
11. Figure 6.13: ML model performance comparison (multi-class classification).
12. Figures 6.14–6.20: Model output plots (no leak, minor leak, major leak for KNN, RF, Decision Tree).

10.6.3 Appendix 6C– SCADA and Experimental Instrumentation (Sunbury & Melton)

Sunbury Instrumentation

1. Pressure transducer at pump discharge.
2. Electromagnetic flow meter (450 mm).
3. Pump-speed telemetry (rpm).
4. SCADA logging at 60-second intervals.

Melton Instrumentation

1. Pressure Transducer (Measurex MX232, 0–16 bar, $\pm 0.25\%$ FS).
2. Pressure Transmitter (ABB 261GSU, $\pm 0.1\%$ accuracy).
3. Mechanical Flow Meter (Zenner Multijet, ISO 4064).
4. EM Flow Meter (Siemens MAG 8000, $\pm 0.4\%$ accuracy).
5. SCADA wireless telemetry, 60-second logging.

10.6.4 Appendix 6D– Variables and Units Used in Chapter 6

Sunbury Pumping Main

1. P – Pressure (m)
2. Q – Flow (L/s)
3. S – Pump speed (rpm)
4. ΔP – Pressure variation (m)
5. ΔQ – Flow variation (L/s)
6. ΔS – Pump-speed variation (rpm)
7. Leak classification: Minor, Major, Burst
8. Timestamp: 1-minute interval

Melton Prototype System

1. P1 – Inlet pressure (kPa)
2. P2 – Outlet pressure (kPa)
3. F1 – Inlet flow (L/s)
4. F2 – Outlet flow (L/s)
5. Flow ratio, Pressure ratio
6. Leak class labels: 0 (No leak), 1 (Minor), 2 (Major)

10.6.5 Appendix 6E– Derived Variables and Analytical Formulations

Hydraulic Relations

1. $P = P_{\text{average}} \pm \Delta P$
2. $Q = Q_{\text{average}} \pm \Delta Q$
3. Leak severity thresholds:
 - a) Minor leak: $\Delta PD < 15\%$
 - b) Major leak: $15\% \leq \Delta PD < 20\%$
 - c) Burst: $\Delta PD \geq 20\%$

Friction Factor (Sunbury pipeline)

Using $L = 6240$ m, $D = 0.45$ m, $Q = 0.248$ m³/s, $hf \approx 16.3$ m:

$f \approx 0.01$ (Darcy–Weisbach)

Machine Learning Feature Matrix

$X = [P, Q, S, \Delta P, \Delta Q, \Delta S]$

Normalisation (z-score)

$$z = (x - \mu) / \sigma$$

Performance Metrics

$$\text{Accuracy} = (TP + TN) / (\text{Total})$$

$$\text{Precision} = TP / (TP + FP)$$

$$\text{Recall} = TP / (TP + FN)$$

$$F1 = 2PR / (P + R)$$

10.6.6 Appendix 6F– Data Tables Referenced in Chapter 6

1. Table 6.1: Logic-based detection outputs (minor/major/burst events).
2. Table 6.2: Hyperparameters for unsupervised ML models.
3. Table 6.3: ML model performance (anomaly F1, ROC-AUC, with/without SMOTE/ADASYN).
4. Table 6.5: SCADA records from Melton (P1, P2, F1, F2).
5. Table 6.6: Leak classification groups (0, 1, 2).
6. Table 6.7: Comparative performance of 14 supervised ML models.
7. Table 6.8: ML-predicted leak/burst conditions.

10.6.7 Appendix 6G– Classification Schemes Used in Chapter 6

Logic-Based Leak Categories (Sunbury)

1. Minor leak: $<15\%$ pressure drop
2. Major leak: $15\text{--}20\%$ pressure drop
3. Burst: $>20\%$ pressure drop

Machine Learning Leak Classes (Melton)

1. Class 0: No leak
2. Class 1: Minor leak
3. Class 2: Major leak

10.6.8 Appendix 6H– Key Algorithmic Workflows

Logic-Based Detection Workflow

1. Acquire SCADA inputs (P, Q, S).
2. Pre-process and compute ΔP , ΔQ , ΔS .
3. Classify pump state (start-up, steady, shutdown).
4. Apply adaptive ΔP – ΔQ thresholds.
5. Categorise event (minor, major, burst).
6. Localise burst using Darcy–Weisbach.
7. Generate SCADA alert.

Machine Learning Workflow

1. Pre-process and balance dataset (SMOTE, ADASYN).
2. Build feature vector X.
3. Train unsupervised (Sunbury) or supervised (Melton) models.
4. Evaluate with accuracy, precision, recall, F1-score.
5. Select best-performing model (LOF unsupervised; KNN/RF supervised).
6. Deploy via web-based interface.
7. Predict leak/burst, timestamp event, export logs.

All algorithms:

<https://drive.google.com/drive/folders/1RzPpwIARXnk756nqFVEFNN2oPMwx6zUc?usp=sharing>

10.6.9 Appendix 6I– Model Performance Highlights

Unsupervised (Sunbury)

1. Local Outlier Factor (LOF): Best performer
 - a. F1 = 0.69 (ADASYN)
 - b. ROC-AUC = 0.72
2. Isolation Forest: Moderate performance
3. DBSCAN: Lowest performance

Supervised (Melton)

1. Best models:
 - a. Random Forest (Accuracy = 1.00)
 - b. K-Nearest Neighbours (Accuracy = 1.00)
 - c. Decision Tree (Accuracy = 1.00)

2. Moderate models: SVM, MLP, AdaBoost
3. Low-performing: Logistic Regression, Gaussian NB, Ridge, QDA

10.6.10 Appendix 6H– System Integration Notes

1. SCADA integration through API polling.
2. Python analytics engine receives live telemetry.
3. Logic-based and ML modules operate in parallel.
4. Alerts delivered via:
 - a) SCADA HMI visual warnings
 - b) Email/SMS
 - c) CSV/PDF reports
5. Processing time:
 - a) Logic-based: <0.05 s per record
 - b) ML (LOF): <0.20 s per record
6. End-to-end alert latency: <2 seconds.
7. Suitable for deployment in DMAs, pump stations, pressure zones.

10.6.11 Appendix 6K– Chapter Specific Notes

1. Logic-based detection is highly interpretable and suitable for real-time deployment.
2. ML-based detection is more adaptive and identifies complex anomalies.
3. LOF was selected as the final unsupervised model; RF/KNN/DT for supervised.
4. Web-based dashboard successfully automates leak/burst prediction.
5. Hybrid logic + ML approach reduces false alarms by ~40%.
6. Sunbury pipeline: Logic-based localisation accuracy $\approx \pm 40$ m.
7. Melton system: Perfect classification achieved under controlled conditions.
8. Future work: high-frequency sensors, edge computing, EPANET fusion, IoT integration.

10.6.12 Appendix 6L– Complete Melton Dataset (No leak and leak conditions)

Table 6.9: No leak data at the Melton experimental yard.

Timestamp	Flow 50mm	Pressure 50mm	Flow 100mm	Pressure 100mm	Flow Diff.	Pressure Diff.	Flow Ratio	Pressure Ratio	Flow Gradient	Pressure Gradient	Flow acceleration	Pressure acceleration	Leak type	Leak type predicted
7/4/2024 11:15	0.56899	0.427711	0.617109	0.351852	0.668836	0.2	0.578706	0.26269	0.519258	0.424757	0.510362	0.424757	0	0
7/4/2024 11:15	0.56899	0.427711	0.617109	0.351852	0.668836	0.2	0.578706	0.26269	0.528153	0.424757	0.519258	0.424757	0	0
7/4/2024 11:15	0.56899	0.427711	0.617109	0.351852	0.668836	0.2	0.578706	0.26269	0.528153	0.424757	0.528153	0.424757	0	0
7/4/2024 11:15	0.56899	0.427711	0.617109	0.351852	0.668836	0.2	0.578706	0.26269	0.528153	0.424757	0.528153	0.424757	0	0
7/4/2024 11:15	0.56899	0.427711	0.617109	0.351852	0.668836	0.2	0.578706	0.26269	0.528153	0.424757	0.528153	0.424757	0	0

Table 6.10: Decision Tree Minor leak.

Timestamp	Flow 50mm	Pressure 50mm	Flow 100mm	Pressure 100mm	Flow Diff.	Pressure Diff.	Flow Ratio	Pressure Ratio	Flow Gradient	Pressure Gradient	Flow acceleration	Pressure acceleration	Leak type	Leak type predicted
8/6/2024 12:37	0.792573	0.304217	0.542524	0.234568	0.149178	0.285714	0.126247	0.194949	0.531076	0.427184	0.531076	0.427184	1	1
8/6/2024 12:38	0.792573	0.304217	0.542524	0.234568	0.149178	0.285714	0.126247	0.194949	0.528153	0.424757	0.531076	0.427184	1	1
8/6/2024 12:39	0.792573	0.304217	0.542524	0.234568	0.149178	0.285714	0.126247	0.194949	0.528153	0.424757	0.528153	0.424757	1	1
8/6/2024 12:40	0.792573	0.304217	0.542524	0.234568	0.149178	0.285714	0.126247	0.194949	0.528153	0.424757	0.528153	0.424757	1	1
8/6/2024 12:41	0.792573	0.304217	0.542524	0.234568	0.149178	0.285714	0.126247	0.194949	0.528153	0.424757	0.528153	0.424757	1	1

Table 6.11: Decision Tree Major leak.

Timestamp	Flow 50mm	Pressure 50mm	Flow 100mm	Pressure 100mm	Flow Diff.	Pressure Diff.	Flow Ratio	Pressure Ratio	Flow Gradient	Pressure Gradient	Flow acceleration	Pressure acceleration	Leak type	Leak type predicted
8/6/2024 15:25	0.576395	0.322289	0.660377	0.253086	0.72373	0.285714	0.641201	0.216363	0.528153	0.424757	0.528153	0.424757	2	2
8/6/2024 15:26	0.576395	0.322289	0.660377	0.253086	0.72373	0.285714	0.641201	0.216363	0.528153	0.424757	0.528153	0.424757	2	2
8/6/2024 15:27	0.576395	0.322289	0.660377	0.253086	0.72373	0.285714	0.641201	0.216363	0.528153	0.424757	0.528153	0.424757	2	2
8/6/2024 15:28	0.687925	0.325301	0.512123	0.259259	0.289444	0.314286	0.210346	0.239703	0.39893	0.429612	0.39893	0.429612	2	2
8/6/2024 15:29	0.687925	0.325301	0.512123	0.259259	0.289444	0.314286	0.210346	0.239703	0.528153	0.424757	0.39893	0.429612	2	2

Table 6.12: K Nearest Neighbor Minor leak .

Timestamp	Flow 50mm	Pressure 50mm	Flow 100mm	Pressure 100mm	Flow Diff.	Pressure Diff.	Flow Ratio	Pressure Ratio	Flow Gradient	Pressure Gradient	Flow acceleration	Pressure acceleration	Leak type	Leak type predicted
7/11/2024 11:57	0.687731	0.412651	0.546517	0.354938	0.344013	0.371429	0.264919	0.348372	0.528153	0.424757	0.528153	0.424757	1	1
7/11/2024 11:58	0.482366	0.412651	0.567438	0.351852	0.746308	0.342857	0.665173	0.332028	0.546389	0.42233	0.546389	0.42233	1	1
7/11/2024 11:59	0.482366	0.412651	0.567438	0.351852	0.746308	0.342857	0.665173	0.332028	0.528153	0.424757	0.546389	0.42233	1	1
7/11/2024 12:00	0.482366	0.412651	0.567438	0.351852	0.746308	0.342857	0.665173	0.332028	0.528153	0.424757	0.528153	0.424757	1	1
7/11/2024 12:01	0.482366	0.412651	0.567438	0.351852	0.746308	0.342857	0.665173	0.332028	0.528153	0.424757	0.528153	0.424757	1	1

Table 6.13: K Nearest Neighbor Major leak .

Timestamp	Flow 50mm	Pressure 50mm	Flow 100mm	Pressure 100mm	Flow Diff.	Pressure Diff.	Flow Ratio	Pressure Ratio	Flow Gradient	Pressure Gradient	Flow acceleration	Pressure acceleration	Leak type	Leak type predicted
7/8/2024 15:37	0.375381	0.683735	0.38505	0.666667	0.651168	0.685714	0.526085	0.575226	0.528153	0.424757	0.528153	0.424757	2	2
7/8/2024 15:38	0.463957	0.692771	0.378414	0.67284	0.481418	0.657143	0.325451	0.565883	0.52237	0.429612	0.52237	0.429612	2	2
7/8/2024 15:39	0.463957	0.692771	0.378414	0.67284	0.481418	0.657143	0.325451	0.565883	0.528153	0.424757	0.52237	0.429612	2	2
7/8/2024 15:40	0.463957	0.692771	0.378414	0.67284	0.481418	0.657143	0.325451	0.565883	0.528153	0.424757	0.528153	0.424757	2	2
7/8/2024 15:41	0.463957	0.692771	0.378414	0.67284	0.481418	0.657143	0.325451	0.565883	0.528153	0.424757	0.528153	0.424757	2	2

Table 6.14: Random Forest Minor leak .

Timestamp	Flow 50mm	Pressure 50mm	Flow 100mm	Pressure 100mm	Flow Diff.	Pressure Diff.	Flow Ratio	Pressure Ratio	Flow Gradient	Pressure Gradient	Flow acceleration	Pressure acceleration	Leak type	Leak type predicted
7/11/2024 10:36	0.690888	0.415663	0.625247	0.354938272	0.462454	0.342857143	0.384467	0.33405806	0.528153	0.3495146	0.528153	0.349515	1	1
7/11/2024 10:37	0.73097	0.415663	0.598297	0.354938272	0.347888	0.342857143	0.285389	0.33405806	0.504663	0.4247573	0.504663	0.349515	1	1
7/11/2024 10:38	0.73097	0.415663	0.598297	0.354938272	0.347888	0.342857143	0.285389	0.33405806	0.528153	0.4247573	0.504663	0.424757	1	1
7/11/2024 10:39	0.73097	0.415663	0.598297	0.354938272	0.347888	0.342857143	0.285389	0.33405806	0.528153	0.4247573	0.528153	0.424757	1	1
7/11/2024 10:40	0.73097	0.415663	0.598297	0.354938272	0.347888	0.342857143	0.285389	0.33405806	0.528153	0.4247573	0.528153	0.424757	1	1

Table 6.15: Random Forest Major leak .

Timestamp	Flow 50mm	Pressure 50mm	Flow 100mm	Pressure 100mm	Flow Diff.	Pressure Diff.	Flow Ratio	Pressure Ratio	Flow Gradient	Pressure Gradient	Flow acceleration	Pressure acceleration	Leak type	Leak type predicted
7/8/2024 15:03	0.655894	0.686747	0.369107	0.663580247	0.121581	0.628571429	0.021709	0.55472292	0.525896	0.4296117	0.525896	0.429612	2	2
7/8/2024 15:04	0.655894	0.686747	0.369107	0.663580247	0.121581	0.628571429	0.021709	0.55472292	0.528153	0.4247573	0.525896	0.429612	2	2
7/8/2024 15:05	0.655894	0.686747	0.369107	0.663580247	0.121581	0.628571429	0.021709	0.55472292	0.528153	0.4247573	0.528153	0.424757	2	2
7/8/2024 15:06	0.655894	0.686747	0.369107	0.663580247	0.121581	0.628571429	0.021709	0.55472292	0.528153	0.4247573	0.528153	0.424757	2	2
7/8/2024 15:07	0.655894	0.686747	0.369107	0.663580247	0.121581	0.628571429	0.021709	0.55472292	0.528153	0.4247573	0.528153	0.424757	2	2

10.7 Appendices for Chapter 7

All algorithm codes and files:

<https://drive.google.com/drive/folders/18BUGQJNCDBC38WtWYV0e8hQDnWSipcm5?usp=sharing>

10.7.1 Appendix 7A– EPANET Hydraulic Input and Processed Output Files

Hydraulic Input Files (Baseline + Leak Scenarios)

(Generated using EPANET for leak sizes 10 mm, 50 mm, 100 mm, 150 mm)

1. *input_file_10mm.csv* – Pressure values for baseline + LN1–LN32 leak simulations (10 mm).
2. *input_file_50mm.csv* – Pressure values for baseline + LN1–LN32 leak simulations (50 mm).
3. *input_file_100mm.csv* – Pressure values for baseline + LN1–LN32 leak simulations (100 mm).
4. *input_file_150mm.csv* – Pressure values for baseline + LN1–LN32 leak simulations (150 mm).

Processed Deviation Files (After Stage 2 Filtering)

(Used in sensitivity computation and optimisation stages)

1. *processed_pressure_data_10mm.csv*
2. *processed_pressure_data_50mm.csv*

3. *processed_pressure_data_100mm.csv*
4. *processed_pressure_data_150mm.csv*

Each processed dataset contains:

1. Absolute deviation (ΔP)
2. Percent deviation
3. Threshold flag (1/0 for $|\Delta P| > 0.5$ m)
4. Threshold-filtered ΔP values (Filtered_Delta_LNxx)

Input files: https://drive.google.com/drive/folders/1qIFpk7jyQmTRsXOA5p0OMzmscCBee-V_?usp=sharing

10.7.2 Appendix 7B– Python Scripts Used in Chapter 7 Analysis

Scripts developed for data processing, sensitivity computation, and diagnostics:

1. *process_pressure_data.py* – Generates all Stage-2 filtered deviation matrices.
2. *node_sensitivity_10mm.py* – Computes node-sensitivity values for 10 mm leaks + diagnostic plots.
3. *node_sensitivity_50mm.py* – Computes node-sensitivity values for 50 mm leaks + plots.
4. *node_sensitivity_100mm.py* – Computes node-sensitivity values for 100 mm leaks + plots.
5. *node_sensitivity_150mm.py* – Computes node-sensitivity values for 150 mm leaks + plots.
6. *sensitivity_summary_report_generator.py* – Creates quartile tables, correlation matrices, sensitivity profiles.
7. *sensor_optimisation.py* – Applies D-Optimal, Greedy, Hybrid, Random, Top-K algorithms.
8. Epanet 32 Matrix :

https://docs.google.com/spreadsheets/d/1MZX0WP_EzVJHyAAptdROJ2FdxLHHnDOr/edit?usp=sharing&ouid=105009817931592767625&rtpof=true&sd=true

9. Node Sensitivity Excel:

https://docs.google.com/spreadsheets/d/1MZX0WP_EzVJHyAAptdROJ2FdxLHHnDOr/edit?usp=sharing&ouid=105009817931592767625&rtpof=true&sd=true

10.7.3 Appendix 7C– EPANET Leak Simulation Parameters

Leak scenarios LN1–LN32 were generated using emitter coefficients corresponding to four orifice sizes:

Leak Diameter Emitter Coefficient (C^e)

10 mm	0.000216591
50 mm	5.39226×10^{-3}
100 mm	2.15691×10^{-2}
150 mm	4.85304×10^{-2}

Leak modelling setup:

1. Applied at each of the 32 junction nodes.
2. Steady-state simulation for baseline and leak cases.
3. Pressure vectors stored for all nodes per simulation.
4. Sensitivity calculated from $|\Delta P|$ at each node across all leak nodes.

10.7.4 Appendix 7D– Mathematical Definitions and Threshold Rules

Pressure Deviation

$$\Delta P_{ij} = P_{ij}(\text{leak}) - P_i(\text{baseline})$$

Threshold Filtering ($\tau = 0.5$ m)

$$\Delta P_{ij} = \Delta P_{ij} \text{ if } |\Delta P_{ij}| > 0.5 \text{ m}$$

$$\Delta P_{ij} = 0 \text{ otherwise}$$

Nodal Sensitivity (S_i)

$$S_i = \text{mean}(|\Delta P^*_{ij}| \text{ across all } j = 1 \dots 32)$$

Detection Rate (%)

$$= \% \text{ of leaks where } \geq 1 \text{ sensor registers } |\Delta P| > \tau$$

Stability Index (SI)

SI = fraction of times a sensor set remains unchanged under threshold perturbations.

10.7.5 Appendix 7E– Ranked Sensitivity Tables for All Leak Sizes

10 mm Leak Scenario – Sensitivity Ranking

(Top nodes consistently: 29, 30, 28, 27, 15)

Full table provided in Section 7.3 and appendix files:

1. *node_sensitivity_analysis_10mm.csv*
2. *sensitivity_analysis_report_10mm.txt*

50 mm Leak Scenario – Sensitivity Ranking

Dominant nodes: 27, 16, 20, 14, 15

100 mm Leak Scenario – Sensitivity Ranking

Dominant nodes: 29, 30, 31, 28, 32

150 mm Leak Scenario – Sensitivity Ranking

Dominant nodes: 28, 22, 25, 29, 27

10.7.6 Appendix 7F– Summary Statistics for Top Sensitive Nodes

For each leak size (10, 50, 100, 150 mm), the following were computed:

1. Mean filtered deviation (Sensitivity Value)
2. Non-zero count (how many steps exceeded threshold)
3. Max deviation
4. Mean deviation
5. Std deviation

Tables presented in Chapter 7:

Table 7.2, 7.8, 7.14, 7.20 (per leak size)

10.7.7 Appendix 7G– Correlation Matrices for Top Nodes

Correlation matrices quantify redundancy between highly sensitive nodes.

Leak size correlation matrices:

1. Table 7.5 (10 mm)
2. Table 7.11 (50 mm)
3. Table 7.17 (100 mm)
4. Table 7.23 (150 mm)

Interpretation:

1. 10 mm: almost perfect correlation → high redundancy
2. 50 mm: moderate correlation → some unique nodes
3. 100 mm: mixed → strong + weak clustering
4. 150 mm: diverse correlations → high-value diverse sensor locations

10.7.8 Appendix 7H– Percentage of Significant Filtered Deviations

Percentage of time steps with $|\Delta P| > 0.5$ m for the top five nodes.

Leak sizes:

5. Table 7.6 (10 mm)
6. Table 7.12 (50 mm)

7. Table 7.18 (100 mm)
8. Table 7.24 (150 mm)

All top nodes respond in >96% of events for moderate to large leaks.

10.7.9 Appendix 7I– Comparative Sensitivity Evolution (10 to 150 mm)

Key patterns across leak magnitudes:

- 10 mm:** Highly redundant, weak deviations, top nodes cluster strongly.
- 50 mm:** Mid-range deviations, more diversity appears.
- 100 mm:** Strong deviations, clear separation of dominant nodes.
- 150 mm:** System-wide strong signals; high-variance correlations.

10.7.10 Appendix 7J– Optimisation Algorithm Outputs

Algorithms:

9. D-Optimal
10. Greedy
11. Hybrid (Greedy → D-Optimal)
12. Random (Best-of-n)
13. Top-K (Sensitivity)

Performance tables:

1. Table 7.26: Detection Rate Comparison
2. Table 7.27: Runtime Comparison
3. Table 7.28: Stability Index

Key outcomes:

1. All deterministic algorithms achieve $\geq 96\%$ detection.
2. Top-K is fastest and most stable.
3. Hybrid gives highest detection for large leaks ($\approx 100\%$).
4. Random baseline is worst performer (low stability, higher runtime).

10.7.11 Appendix 7K– Figures Used in Chapter 7

1. Figure 7.1 – OSP Methodological Flowchart
2. Figure 7.2 – Sensitivity patterns (10 mm)
3. Figure 7.3 – Violin and correlation diagnostics (10 mm)
4. Figure 7.4 – Sensitivity patterns (50 mm)
5. Figure 7.5 – Diagnostics (50 mm)
6. Figure 7.6 – Sensitivity patterns (100 mm)

7. Figure 7.7 – Diagnostics (100 mm)
8. Figure 7.8 – Sensitivity patterns (150 mm)
9. Figure 7.9 – Diagnostics (150 mm)
10. Figure 7.10 – Spatial sensor distribution map
11. Figure 7.11 – Detection-rate comparison
12. Figure 7.12 – Runtime comparison
13. Figure 7.13 – Stability index comparison

10.7.12 Appendix 7L– Chapter Specific Notes

- 10.7.12.1 Sensitivity threshold $\tau = 0.5$ m ensures noise-free hydraulic signals.
- 10.7.12.2 High recurrence of Nodes 27, 28, 29, 30 indicates hydraulic dominance along the main transmission corridor.
- 10.7.12.3 Larger leaks show higher variance and lower redundancy → improved diagnostics.
- 10.7.12.4 Hybrid and D-Optimal excel in diverse-response conditions.
- 10.7.12.5 Sensitivity + optimisation integration creates a reproducible, scalable OSP framework.
- 10.7.12.6 Results generalise well for larger WDNs due to consistent hydraulic behaviour patterns.