Doctoral theses at NTNU, 2024:358

Lam Van Nguyen

Integrating Machine Learning and GIS for Sewer Condition Assessment and Visualization

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Engineering
Department of Ocean Operations and Civil
Engineering

NTNU
Norwegian University of
Science and Technology

Lam Van Nguyen

Integrating Machine Learning and GIS for Sewer Condition Assessment and Visualization

Thesis for the Degree of Philosophiae Doctor

Trondheim, September 2024

Norwegian University of Science and Technology
Faculty of Engineering
Department of Ocean Operations and Civil Engineering

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Engineering
Department of Ocean Operations and Civil Engineering

ISBN 978-82-326-8306-2 (printed ver.)
ISBN 978-82-326-8305-5 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2024:358

Printed by NTNU Grafisk senter

Abstract

Sewer network, including wastewater and stormwater pipelines, is among the critical infrastructures that can be seen as one kind of national asset. These asset-related issues can cause serious consequences affecting people and the environment. As with other infrastructures, the sewer network deteriorates over time, and continuous adjustments are required. Consequently, rehabilitation and maintenance activities are needed to ensure its engineered functions work properly, reduce risks, optimize performance, and minimize costs.

One of the effective ways of the predictive maintenance strategy is to estimate the condition of sewer pipelines. In general, the sewer condition assessment model is a tool that provides decision-makers valuable information on not only the current state but also the future state of the sewers and support for prioritization of inspection, reparation, or renewal of sewer pipes. However, the change in sewer condition significantly depends on input factors. Therefore, the quality of sewer condition models is influenced both by the input factors used and the methods employed. Although many different methods and techniques have been proposed and implemented for sewer condition analysis, no agreement has been reached regarding the best method for sewer condition assessment.

In this thesis, a methodology for modelling sewer conditions has been successfully developed and applied for Ålesund city, Norway by employing state-of-the-art machine learning (ML) and Deep Learning (DL). In addition, various feature selection methods have been investigated to assess the importance of ten physical factors (age, diameter, depth, slope, length, pipe type, material, network type, pipe form, and connection type) and ten environmental factors (rainfall, geology, landslide area, population, land use, building area, groundwater, traffic volume, distance to road, and soil type). In this thesis, Geographic Information System (GIS) was used as a main tool to analyze, store, and visualize the results. The primary purpose of the developed methodology is to partially support local water agencies to control and operate the wastewater/stormwater system more effectively and partly optimize predictive maintenance strategies.

The condition of sewer pipes can be defined based on damage score (regression problem) or damage class (classification problem). The performance of sewer condition assessment models using these outputs has not been assessed and compared. This thesis addresses the above statement by developing ML models for sewer condition assessment. The performance of the models was compared using the popular assessment criteria for the regression problem and the classification problem.

Paper I evaluates the potential application of ML algorithms for predicting the damage scores of sewer pipelines. The performance of the developed models was compared using the popular statistical metrics for the regression problem, such as the coefficient of determination (R^2), mean absolute error (MAE), and root means square error (RMSE). The results show that although the damage scores of sewer pipes can be used to assess their conditions, the reliability of the prediction is low due to data divergence (subjectivity in assigning scores for inspected sewers).

By transforming damage scores to damage classes, the prediction performance of ML models has been improved significantly (**Paper II**). The assessment criteria for the classification problem such as geometric mean (GM), accuracy (ACC), F-Score, Matthew’s correlation coefficient (MCC), the area under the Receiver Operating Characteristic curve (AUC-ROC), and the area under the Precision-Recall curve (AUC-PRC) between models are more stable compared to regression models.

In **Paper III**, the efficiency of hybrid ML models in predicting sewer conditions was tested. The hybrid models have better performance, even with a multiclassification problem, compared to the binary classification problem in **Paper II**. This result shows that the potential application of other hybrid ML models should be considered in future research.

The results of these studies show that age and material are the most significant factors affecting the sewer condition in the study area. The hybrid ML models are more sensitive than the normal ML models in predicting sewer conditions. Finally, a sewer condition assessment map was prepared that provides useful information for supporting predictive maintenance strategies.

Paper IV introduces an integrated framework as a combination between GIS, 3D-creation platform, augmented reality (AR) techniques, and ML algorithms for the dynamic visualization of the condition of sewer networks. The positioning accuracy of the application for 2D objects is equivalent to that of well-designed GPS receivers (approximately 1-3 m), depending on the handheld device used.

Acknowledgements

The financial support for this PhD research was provided by the Smart Water Project, financed by Ålesund city and the Norwegian University of Science and Technology. The data analysis and write-up of this thesis were carried out at the Department of Ocean Operations and Civil Engineering, Faculty of Engineering, Norwegian University of Science and Technology. I would like to acknowledge these organizations for their support during my PhD study.

This thesis would not have been completed without the enthusiastic guidance, encouragement, support, and suggestions from the supervisors. First and foremost, I would like to express my profound gratitude to my main supervisor Prof. Razak Seidu for his supervision, patience, and positive encouragement during my PhD work. With his very cheerful attitude and care, all conversations between him and me are always fun, and the pressure during my PhD work has been reduced dramatically. Moreover, his excellent ideas and discussions on technical issues are extensively important to make my work in the correct direction. Similarly, I would like to thank my co-supervisor Prof. Dieu Tien Bui, for his ideas, feedback, and which significantly contributed to my work. Many thanks for his constant and patient advice when I am confused with academic problems.

I would like to say thank you to Bjørn Skulstad, David Chris Mertsching, and Lars Andreas Slyngstad Lågeide at Ålesund city for providing data, constructive contributions, sincere encouragement, and wonderful support. Their ideas and comments at meetings and discussions significantly help me to complete this thesis and integrate academic knowledge with practical experiments.

Furthermore, I am thankful to my colleagues at Smart Water and Environmental Engineering Group, especially Assoc. Prof. Hadi Mohammed and PhD Hoese Michel Tornyeviadzi, for their discussion and interpretation of the study results. I am grateful to Kristian Fjørtoft for the funny discussions and his useful experiences.

Finally, I wish to give special thanks to my wife Thu Thi Do, and my lovely daughter Han Gia Nguyen for their understanding, constant love, emotional support, and inspiration that allowed me to spend most of my time on the research work. I also give special thanks to my family for their support and patience.

Lam Van Nguyen
Ålesund, April 2024

Contents

List of Papers

This thesis is based on research resulting in four journal papers. In the following list of publications, the papers are prioritized in thematic order.

I. **Lam Van Nguyen** and Razak Seidu (2022). Application of Regression-Based Machine Learning Algorithms in Sewer Condition Assessment for Ålesund city, Norway. *Water*, *14*(24), 3993. DOI: https://doi.org/10.3390/w14243993.

II. **Lam Van Nguyen,** Dieu Tien Bui, and Razak Seidu (2022). Comparison of Machine Learning Techniques for Condition Assessment of Sewer Network. *IEEE Access*, *10*, 124238-124258. DOI: https://doi.org/10.1109/ACCESS.2022.3222823.

III. **Lam Van Nguyen** and Razak Seidu (2023). Predicting Sewer Structural Condition Using Hybrid Machine Learning Algorithms. *Urban Water Journal*. DOI: https://doi.org/10.1080/1573062X.2023.2217430.

IV. **Lam Van Nguyen,** Dieu Tien Bui, and Razak Seidu (2023). Utilization of Augmented Reality Technique for Sewer Condition Visualization. *Water*, *15*(24), 4232. DOI: https://doi.org/10.3390/w15244232.

The following papers and tutorials will not be discussed in this thesis. However, they may be considered relevant due to co-authorship contribution or support in teaching students on similar topics during the author's PhD work:

i. **Lam Van Nguyen,** Hoese Michel Tornyeviadzi, Dieu Tien Bui, and Seidu Razak (2022). Predicting Discharges in Sewer Pipes Using an Integrated Long Short-Term Memory and Entropy A-TOPSIS Modeling Framework. *Water*, *14*(3), 300. DOI: https://doi.org/10.3390/w14030300.

ii. **Lam Van Nguyen,** Dieu Tien Bui, and Seidu Razak (2023). A Comparative Flood Susceptibility Assessment in a Norwegian Coastal City using Feature Selection Methods and Machine Learning Algorithms. *Advances in Research on Water Resources and Environmental Systems. International Conference on Geo-Spatial Technologies and Earth Resources*. Environmental Science and Engineering. Springer International Publishing, 591-618, January 2023. DOI: https://doi.org/10.1007/978-3-031-17808-5_36.

iii. **Lam Van Nguyen,** Dieu Tien Bui, and Seidu Razak (2022). Predicting Structural Sewer Condition Using Machine Learning Algorithms. International Water Association IWA World Water Congress & Exhibition, Denmark (**Oral presentation**).

iv. **Lam Van Nguyen,** Dieu Tien Bui, and Seidu Razak (2020). Identification of Sensitive Factors for Placement of Flood Monitoring Sensors in Wastewater/Stormwater Network Using GIS-Based Fuzzy Analytical Hierarchy Process: A Case of Study in Ålesund, Norway. *Proceedings of the International Conference on Innovations for Sustainable and Responsible Mining*. Springer International Publishing, 79-97, October 2020. DOI: https://doi.org/10.1007/978-3-030-60269-7_5.

v. Van-Sang Nguyen, Van-Tuyen Pham, **Lam Van Nguyen**, Ole Baltazar Andersen, Rene Forsberg, Dieu Tien Bui (2020). Marine gravity anomaly mapping for the Gulf of Tonkin area (Vietnam) using Cryosat-2 and Saral/AltiKa satellite altimetry data. *Advances in Space Research*. 66, 505-519. DOI: https://doi.org/10.1016/j.asr.2020.04.051.

vi. Nguyen Van Sang, Khuong Van Long, Tran Tuan Dung, **Lam Van Nguyen**, Bui Cong Que, Do Van Mong, Bui Dang Quang, Ole Baltazar Andersen, Rene Forsberg, Dieu Tien Bui (2023). Seafloor depth mapping of central Vietnam's Sea area and its surrounding using gravity anomaly data and gravity geological method. *Advances in Space Research*. 72(5), 1721-1738. DOI: https://doi.org/10.1016/j.asr.2023.04.033.

vii. **Lam Van Nguyen** and Seidu Razak (2022). Guideline for Using Visualization Platform. *Application guideline* (27 pages). Water and Environmental Engineering Group. Department of Ocean Operations and Civil Engineering. Norwegian University of Science and Technology. Download link: https://www.mediafire.com/file/zgw5xpnw5tsp1km/Application_Guideline.pdf/file.

viii. **Lam Van Nguyen** and Seidu Razak (2022). Modelling and Simulation of Urban Stormwater Collection Systems using Integration of SWMM, GIS, and Python. *Teaching tutorial 1* (62 pages). Water and Environmental Engineering Group. Department of Ocean Operations and Civil Engineering. Norwegian University of Science and Technology. Download link: https://www.mediafire.com/file/b8suckp9hwn5y5a/SWMM_GIS_Stormwater_Tutorial.pdf/file.

ix. **Lam Van Nguyen** and Seidu Razak (2022). Application of MIKE URBAN and GIS for

Sewer Collection Systems Modelling and Simulation. *Teaching tutorial 2* (63 pages). Water and Environmental Engineering Group. Department of Ocean Operations and Civil Engineering. Norwegian University of Science and Technology. Download link: https://www.mediafire.com/file/9e344d81mpvq2mx/MikeUrban_GIS_Wastewater_Tutorial.pdf/file.

x. **Lam Van Nguyen** and Seidu Razak (2022). Integration of GIS and Python for Sewer Condition Assessment. *Teaching tutorial 3* (78 pages). Water and Environmental Engineering Group. Department of Ocean Operations and Civil Engineering. Norwegian University of Science and Technology. Download link: https://www.mediafire.com/file/7xpi5pfbys2jbed/Python_GIS_Condition_Assessment_Tutorial.pdf/file.

Published papers were reprinted with permission from the publishers.

Supervisors

Main Supervisor

Prof. Razak Seidu, Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology, Ålesund, Norway.

Co-supervisor

Prof. Dieu Tien Bui, Department of Business and IT, University of South-Eastern Norway, Gullbringvegen 36, 3800 Bø i Telemark, Norway.

Nomenclature

ACC	Accuracy
ANFIS	Adaptive Neuro-Fuzzy Inference System
ANN	Artificial Neural Networks
AR	Augmented Reality
ARMA	Autoregressive Moving Average
ARIMA	Autoregressive Integrated Moving Average
AUC-PRC	Area Under the Precision-Recall Curve
AUC-ROC	Area Under the ROC Curve
BIM	Building Information Modeling
BG	Bagging
BNB	Bernoulli Naive Bayes
BMA	Bayesian Model Averaging
BPNN	Back-Propagation Neural Network
CARLS	Condition Assessment and Rehabilitation for Large Sewers
CART	Classification and Regression Trees
CCTV	Closed-circuit television
CHAID	Chi-square Automatic Interaction Detector
COAH	Copernicus Open Access Hub
DEM	Digital Elevation Model
DG	Dagging
DL	Deep Learning
ERT	Extremely Randomized Trees
ESRI	Environmental Systems Research Institute, Inc.
ETR	Extra Trees Regression
GIS	Geographic Information System
GNB	Gaussian Naive Bayes
GM	Geometric Mean
GP	Gaussian Process
GPS	Global Positioning System
GTB	Gradient Tree Boosting
HGB	Histogram-Based Gradient Boosting
IDW	Inverse Distance Weighting
IoT	Internet of Things
J48DT	J48 Decision Tree
KNN	K-Nearest Neighbor

LR	Logistic Regression
LSTM	Long Short-Term Memory
NCSC	Norwegian Climate Service Center
NGS	Norwegian Geological Survey
NMA	Norwegian Mapping Authority
NPRA	Norwegian Public Roads Administration
MAE	Mean Absolute Error
MCC	Matthew’s Correlation Coefficient
ML	Machine Learning
MLP	Multi-layer Perceptron
PACP	Pipeline Assessment Certification Program
PCA	Principal Component Analysis
R^2	Coefficient of determination
RBF	Radial Basis Function Neural Network
RC	Ridge Classification
RF	Random Forest
RNN	Recurrent Neural Networks
RMSE	Root Means Square Error
ROC	Receiver Operating Characteristic
RotF	Rotation Forest
RTK	Real-Time Kinematic
SAMME	Stagewise Additive Modeling using a Multi-class Exponential
SCADA	Supervisory Control and Data Acquisition
SFS	Stepwise Feature Selection
SMOTE	Synthetic Minority Oversampling TEchnique
SVM	Support Vector Machine
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
WEKA	Waikato Environment for Knowledge Analysis
WRC	Water Research Center

List of Figures

List of Tables

Chapter 1
Background

1.1. Introduction

Wastewater and stormwater pipelines are critical components of wastewater management infrastructure in many cities (Farkas et al., 2020). They act as the infrastructure nerve for the collection and transport of wastewater/stormwater to either wastewater treatment plants or recipients in centralized wastewater management systems.

In many advanced countries, including Norway, wastewater and stormwater pipelines are rapidly aging and deteriorating; and they cannot perform their engineered functions, thereby leading to significant environmental, public health, and socio-economic impacts (Anand et al., 2022). Managing and maintaining this infrastructure typically requires significant investments. For instance, the total budget spending on the Norwegian drainage system until 2019 was approximately $70.6 billion, and the investment needed for renewing and upgrading this system in Norway by 2040 was estimated at $11.5 billion (Stian et al., 2021).

Condition transformation of sewer pipelines is a multi-step that is influenced by many factors including physical, environmental, and operational at the same time (Hawari et al., 2020). Consequently, the assessment of sewer conditions is more challenging under the interactive relationship of the aforementioned factors. Additionally, sewer pipeline inspection and monitoring have received increased attention from water managers and relevant agencies to address problems and avoid additional failures and collapses (Salihu et al., 2022).

Maintenance management approaches can be generally categorized into reactive, preventive, and predictive maintenance (PdM) (Susto et al., 2015). Recently, more and more enterprises have an awareness of the advantages of PdM in providing cost-effective problems and minimizing downtime and wasted costs (Sezer et al., 2018). In the context of sewage asset management, a PdM strategy cannot be implemented effectively without a deep understanding of the system, and an efficient water management strategy requires a proper condition assessment framework (Chughtai & Zayed, 2007). Internet of Things (IoT) combined with artificial intelligence, such as Machine Learning (ML) and Deep Learning (DL), will be effective tools for fulfilling prognostic and predictive tasks. As a result, they are powerful data-driven models found in much current evolution of PdM solutions (Dalzochio et al., 2020).

1.2. Problem statement

Unlike buildings, road networks, or bridges, which are visible infrastructures, a sewage network is a hidden and underground infrastructure that cannot be monitored or assessed directly using normal ocular or visual measurements. Moreover, monitoring and inspecting all sewer pipelines is nearly impossible due to time, financial, and technology constraints. Therefore, there is a need for prediction models that not only facilitate the systematic monitoring, evaluation, and costing of the maintenance needs of wastewater and stormwater pipelines but also predict the evolution of pipe deterioration; to enable authorities to make short, medium, and long-term investment decisions on their pipe infrastructure.

In general, the performance of sewer infrastructure is a function of a myriad of factors including physical factors (for example, size, age, material, and pipe type), environmental factors (for example, rainfall, land use, and groundwater), and operational factors (for example, high-pressure, temperature, and flow) (Hawari et al., 2020). Over the years, many asset management planning tools have been developed for predicting pipe conditions using the aforementioned factors as the basis for investment decisions (Atambo et al., 2022; Laakso et al., 2018, 2019; Roghani et al., 2019). In this regard, condition assessment models are valuable tools for accurately predicting future sewer pipeline performance and effective maintenance (Caradot et al., 2020; Hawari et al., 2020). Many models for sewer condition assessment, such as *physical*, *statistical*, and *machine learning*, have been employed to assess the status of sewer pipelines (Caradot et al., 2018; Salihu et al., 2022; Tscheikner et al., 2019).

Although physical models perform well in assessing sewer conditions during the initial operational period and the construction phases (Heydarzadeh et al., 2021), the data needed to simulate deterioration mechanisms are normally scarce (Tscheikner et al., 2019). However, sewer degradation is a complex process affected by many factors, and statistical models are likely to have more advantages in calculating speed and straightforward function compared with the physical models (Wei et al., 2020). Nevertheless, some assumptions in these models, such as the distance between consecutive conditions being constant or the sewer status in each condition being a normal distribution, are difficult to satisfy in reality. Additionally, sewer deterioration is a non-linear process, and statistical models are not able to predict the process with high accuracy (Zamanian et al., 2020).

Machine Learning (ML) models can handle the complex non-linear interlinked relationship between input factors and sewer pipe conditions, even if these relationships are unclear or when data is incomplete (Ahmad et al., 2018). With the flexibility of input and output data types, including numeric, nominal, or categorical, the accuracy of ML models can be easily improved by increasing the number of input factors and inspections or using an adequately distributed dataset in assessing sewer conditions (Uddin et al., 2019; Yin, Chen, Bouferguene, & Al-Hussein, 2020). However, the accuracy and performance of different ML models are dissimilar due to the quality of the dataset used, randomness in splitting data, different characteristics of the study area, and used algorithms (Kovacs et al., 2022). Consequently, no ML model is the best for all cases (Tscheikner et al., 2019; Zamanian et al., 2020). Besides, a comprehensive comparison between different types of ML models used for regression and classification problems in modelling the condition of sewer pipelines is still missing. This thesis attempts to partly fill this research gap by exploring and verifying the potential application of ML algorithms to predict the condition of sewer pipes. In addition, the significance of input factors was briefly analyzed using the filter, wrapper, and embedded methods to provide helpful information for decision-makers in prioritizing significant factors during sewer predictive maintenance.

Geographic Information System (GIS) is a powerful tool for processing spatial data, performing spatial analysis, and manipulating spatial outputs. It provides a consistent visualization environment for displaying a model's input data and results (Vairavamoorthy et al., 2007). GIS-based approaches have been applied for (i) designing water, wastewater, and storm-water networks (Shamsi, 2002); (ii) improving the accuracy of pipe failures estimation (Li et al., 2011), and (iii) building GIS-based simulation tools to support asset management and maintenance (Cheng et al., 2020; Sekar et al., 2013). Based on the GIS database, predictive models of the future condition of sewer pipes can be autonomously constructed and updated. This thesis used GIS was used to store and analyze data before feeding it to ML models and visualizing results.

Although the information from field surveys is processed on computers using GIS, the visualization of spatial and environmental data still uses conventional maps and reports (Mirauda et al., 2017). In addition, with the massive development of internet infrastructure, interactive connections between devices are becoming more accessible and faster. Therefore, the development of an architecture that can automatically interact with the GIS database, predict sewer conditions, and visualize the results is essential.

Augmented Reality (AR) is holographic technology used to project 3D virtual models into the physical space and it is being applied for maintenance purposes in many domains such as in production factories (Coscetti et al., 2020; Kostoláni et al., 2019), urban management (Kaddioui et al., 2019), and other fields (Behzadan et al., 2015; Hamacher et al., 2016; Koutitas et al., 2019). A combination of AR and GIS, called ARGIS, is gaining more popularity, and it not only enhances users' experience but also provides powerful tools for large data management (Kamel et al., 2017). Furthermore, the integration of AR technology, GIS, and the Internet of Things (IoT) has proven to be an effective approach to the analysis, visualization, and exploration of spatial data (Carneiro et al., 2018). However, its application for sewer maintenance remains very embryonic. This thesis attempts to partly fill this gap by developing a methodology combining GIS, ML, and AR for sewer network 3D visualization and predictive maintenance purposes.

1.3. Aim and objectives

The overall aim of this thesis is to develop an integrated GIS and Machine Learning-based approach for sewer condition assessment and visualization of wastewater pipelines. The specific objectives of the research are to:

1. Develop a comprehensive sewer asset database from multiple sources for condition assessment using GIS as the main tool to store and process data.
2. Assess sewer conditions employing machine learning algorithms and identify the best-performing machine learning-based sewer condition assessment model for the prediction of the future sewer condition.
3. Provide a comprehensive process for data collection, implementation of ML models, and visualization platform for sewer condition assessment.
4. Develop an interactive geospatial platform for the visualization of the present and future condition of the sewer.

The developed process in this thesis could be applied at an industrial scale to support local water utilities in determining appropriate predictive maintenance strategies. Even though the study focused on Ålesund city in Norway, the approaches presented here are applicable elsewhere.

1.4. Structure of the thesis

This thesis is organized into five chapters. This introductory chapter contains an overview as well as the objectives of this thesis. Chapter 2 briefly introduces an overview of the sewer condition assessment process. Chapter 3 presents the fundamental research methods and data used in this thesis. Chapter 4 summarizes the results from the papers included in this thesis. Discussions of the entire thesis are concluded in Chapter 5. The final chapter provides the major conclusions drawn from this thesis and recommendations for further investigations. The papers selected in this thesis are given in Appendix A. Appendix B shows codes, download sources, and guideline for the application developed in this thesis.

Chapter 2

Condition Assessment Overview

2.1. Condition assessment

Condition assessment is an essential process to distinguish current sewer conditions and establish timely, cost-effective, and appropriate maintenance strategies (Rahman & Vanier, 2004). The workflow in **Figure 2.1** shows the basic elements of asset management based on previous studies (Ana & Bauwens, 2007; Lee et al., 2015; Noshahri et al., 2021).

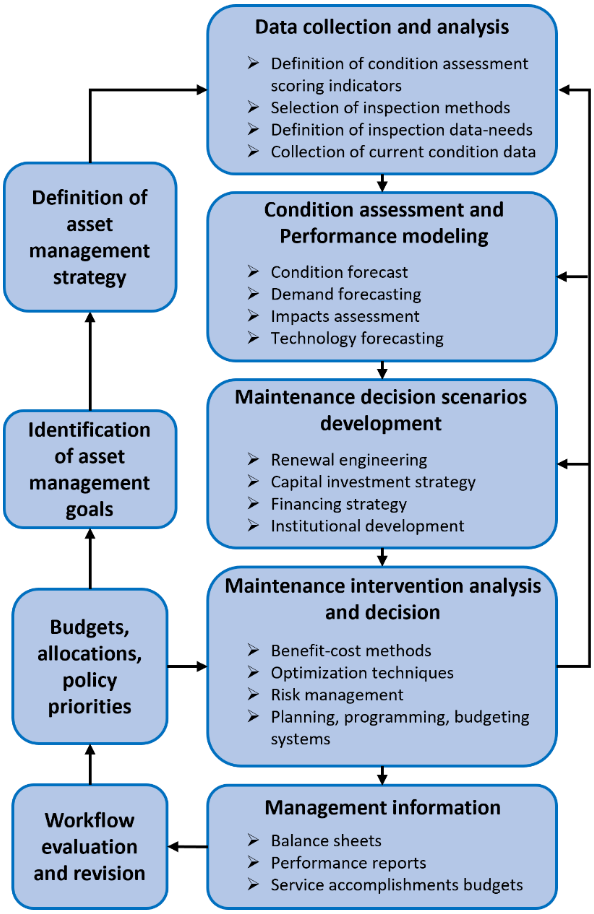

Data collection and analysis
➢ Definition of condition assessment scoring indicators
➢ Selection of inspection methods
➢ Definition of inspection data-needs
➢ Collection of current condition data
Condition assessment and Performance modeling
➢ Condition forecast
➢ Demand forecasting
➢ Impacts assessment
➢ Technology forecasting
Maintenance decision scenarios development
➢ Renewal engineering
➢ Capital investment strategy
➢ Financing strategy
➢ Institutional development
Maintenance intervention analysis and decision
➢ Benefit-cost methods
➢ Optimization techniques
➢ Risk management
➢ Planning, programming, budgeting systems
Management information
➢ Balance sheets
➢ Performance reports
➢ Service accomplishments budgets
Definition of asset management strategy
Identification of asset management goals
Budgets, allocations, policy priorities
Workflow evaluation and revision

Figure 2.1. Workflow for asset management

An effective asset management plan provides comprehensive and cost-effective strategies for utilities and municipalities to undertake timely maintenance, rehabilitation, and replacement of their sewer pipes (Najafi et al., 2021). Sewer condition assessment model is critical for these strategies through the prediction of current and future condition of sewer pipes (Fan et al., 2022).

It is worth noting that there is a difference between a structural and an integrity-oriented condition classification of sewer pipes. While structural condition refers to the capacity of the pipe to fulfil its structural role, integrity-oriented condition focuses more on the overall condition of a sewer pipe (Rahman & Vanier, 2004; Tscheikner et al., 2019). For example, a sewer pipe with a single severe damage (causing a serious structural failure) may require an immediate replacement, but its overall integrity is still good. In contrast, a sewer pipe without serious defect(s), but irreparable deterioration along the whole pipe length requires renovation or replacement. This dissertation does not focus on predicting which condition (in terms of structural or integrity-oriented condition) the sewer pipes belong to; it mainly focuses on the application of ML algorithms for predicting sewer conditions based on given sewer status.

2.2. Pipe condition inspection techniques

Several techniques are used for sewer pipes inspection in order to establish their status. These techniques can be classified into direct and indirect groups based on how information on the sewer condition is obtained. While the direct methods provide pipe distress indicators (measurable defects or flaws such as cracks, corrosion pits, or wire breaks), the indirect methods generate inferential indicators (such as soil type or groundwater fluctuations) (Kleiner & Rajani, 2022; Liu & Kleiner, 2013).

A summary of popular techniques for pipe condition assessment are presented in **Table 2.1**. Detailed information on these techniques as well as their advantages and disadvantages are presented in the study of Liu and Kleiner (2013).

Table 2.1. Summary of methods for pipe condition assessment

Method	Type	Subsets
Visual inspection	Direct method	CCTV, Laser scan
Electromagnetic methods	Direct method	Magnetic flux leakage, Remote field eddy current, Broadband electromagnetic, Pulsed eddy current testing, Ground penetrating radar, Ultra-wideband pulsed radar
Acoustic methods	Direct method	Sonar profiling, Impact echo, SmartBall, Sahara system, Leak detection
Ultrasound methods	Direct method	Guided wave ultrasound, Discrete ultrasound, Phased array technology
Radiographic methods	Direct method	-
Thermography methods	Direct method	-
Linear polarization resistance of soil	Indirect method	-
Soil characterization	Indirect method	-
Pipe to soil potential survey	Indirect method	-

Among the aforementioned methods, CCTV is the most widely used and cost-effective method for the assessment of sewer pipes that have specific characteristics in unsanitary environments, complex surveillance circumstances, and high pressure (Hassan et al., 2019; Koo & Ariaratnam, 2006; Tscheikner et al., 2019).

2.3. General contributing factors of sewer condition assessment

The change of sewer condition is a complex process that is determined by multiple factors. The deterioration of pipes is a continuing process that can be divided into structural deterioration and hydraulic deterioration. While the structural deterioration process is characterized by structural defects (e.g., cracks, deformations, or fractures) that directly reduce the structural integrity (such as the shape and load-bearing capacity of pipes), hydraulic deterioration indicates a reduction of cross-sectional area and an increase in the roughness coefficient (Tran, 2007). Identifying the most significant factors of sewer pipe condition is critical in reducing data collection costs and increasing prediction accuracy and performance of sewer condition assessment models (Kley & Caradot, 2013).

Factors affecting sewer pipe condition can be grouped into three main categories: physical, environmental, and operational factors (Hawari et al., 2020; Hawari et al., 2016; Shi, 2018).

Physical factors are associated with the physical attributes of the pipes, such as diameter, length, or material. Environmental factors are related to the surrounding environment, such as soil type, groundwater, or land cover. Operational factors relate to the function and operation of pipes including flow velocity, maintenance, or sediments in pipes (Ana & Bauwens, 2010). While physical factors are normally stored in the database of agencies and municipalities, information about environmental and operational factors is often unavailable (Mohammadi et al., 2020; Salman, 2010). For example, pipe age, groundwater level, pipe size, or pipe material mainly contribute to the structural deterioration process, the hydraulic deterioration process is significantly affected by tree type, pipe depth, pipe location, or soil type (Tran et al., 2006).

Although many factors are listed in historical studies, only a few of them are collected and used while constructing sewer condition assessment models because of their availability and statistical significance (Tran et al., 2006).

A summary of factors used for sewer condition assessment is presented in **Figure 2.2** (Ana et al., 2009; Hawari et al., 2020; Mohammadi et al., 2020).

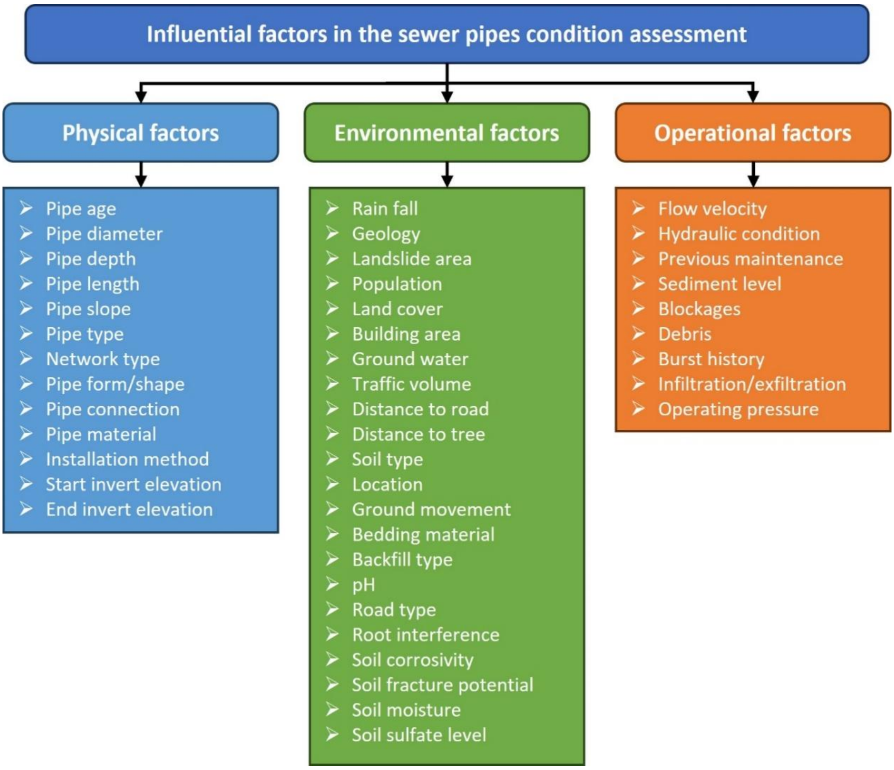


Figure 2.2. Factors affecting sewer condition assessment

2.4. Grading methods for sewer condition

After obtaining the visual inspections of sewer pipes, damaged scores of pipes are assigned based on their observed defects (for example, cracks, broken, collapse, surface damage, and line failure). According to Ahmadi et al. (2014), sewer condition grades are generally identified using two typical methods including subjective grading and distress-based evaluation.

Subjective grading method assesses the condition of sewer pipes based on visual inspection, in-situ measurements, or expert opinion. By using this method, the grading scale of sewer pipes is represented by a score that is consistent with the level of defects. The accuracy of this method depends on an inspector's experience and the reliability of the system used (Ahmadi et al., 2014). The distress-based evaluation method rates the distress degree of sewer pipes based on observations using a predefined protocol developed by experts. The most well-known protocols are the Sewerage Rehabilitation Manual of the Water Research Center (WRC) in the United Kingdom, the Guidelines for Condition Assessment and Rehabilitation for Large Sewers (CARLS) conducted by the National Research Council of Canada, the Manual of Sewer Condition Classification of the North American Association of Pipeline Inspectors, the Norsk Vann Manual reported by the Norwegian organization for the water industry (https://norskvann.no/ (accessed on 06th April 2024)), and the Pipeline Assessment Certification Program (PACP) developed by the National Association of Sewer Service Companies (Rahman & Vanier, 2004).

In general, the coding system is used to assess the CCTV inspection videos of sewer pipes. The defects existing in the pipe are coded according to one of the coding protocols mentioned above, and finally, a score is calculated to represent the overall condition of the sewer pipe (Hawari et al., 2018; Yin, Chen, Bouferguene, Zaman, et al., 2020). For instance, most agencies and utilities use a scale of 1-3 grading system in the wastewater industry (WSAA, 2013), or a 1-5 grading system is normally used for sanitary sewer or stormwater pipes (Haugen & Viak, 2018; Park, 2009). An example of pipe condition grade classified based on the PACP manual is shown in **Table 2.2**.

Table 2.2. Grading codes for sewer pipe condition (Khazraeializadeh, 2012; Yin, Chen, Bouferguene, & Al-Hussein, 2020)

Condition grade	Pipe condition	Description	Time to failure
1	Excellent	Defect is minor	Failure unlikely in the foreseeable future
2	Good	A defect has just started to deteriorate	A pipe unlikely to fail for at least 20 years
3	Fair	A moderate defect that keeps on deteriorating	A pipe may fail in 10 to 20 years
4	Poor	A severe defect will reach its worst situation within an expected period of time	A pipe will probably fail in 5 to 10 years
5	Immediate attention	A defect should be care taken of immediately	A pipe has failed or will likely fail within the next 5 years

2.5. Sewer condition assessment models

A deep understanding of the system is critical for utilities and municipalities to run efficient predictive maintenance strategies (Chughtai & Zayed, 2007). A reliable sewer condition assessment model enhances our understanding of the deterioration process and mechanism, and it is a critical tool for the evaluation of non-inspected pipe conditions and forecast of the future state for rehabilitation strategies (Caradot et al., 2017). Moreover, condition prediction model-based approach is one of the most popular research areas in the past decades because it is based on more objective data and is less expensive compared to expert knowledge and in-situ sensors respectively (Tran & Nguyen, 2010).

Sewer condition assessment models can be generally classified into physical, statistical, and machine learning. In physical models, a clear quantitative relationship between contributing factors and sewer conditions is defined without dealing with the uncertainty of the deterioration process (Hawari et al., 2020). On the contrary, these uncertainties are considered by using probability distribution based on equations in the statistical models (Tscheikner et al., 2019). Machine learning models evaluate the mathematical relationships between sewer conditions and contributing factors by learning the deterioration behavior from inspection data, no assumption about the model structure is required in ML models because these types of models use a data-driven approach (not a model-driven approach) (Ahmad et al., 2018).

Recent models for sewer condition assessment are summarized in **Table 2.3**. Some advantages, limitations, and applications of these models are presented in Hawari et al. (2020).

Table 2.3. Summary of several sewer condition assessment models from the literature

Type	Model	References
Physical model	Power function models	Doleac et al. (1980)
	Linear function models	Randall-Smith et al. (1992)
	UtilNets	Hadzilacos et al. (2000)
	ExtCorr	Hawari et al. (2020)
Statistical model	Regression models	Bakry et al. (2016), Balekelayi and Tesfamariam (2019), Kabir et al. (2018), Sempewo and Kyokaali (2019)
	Markov chains	Sempewo and Kyokaali (2019)
	Cohort survival models	Caradot et al. (2017)
	Multiple discriminant analysis	Vladeanu et al. (2019), Alsaqqar et al. (2017)
	Probabilistic models	Kleiner and Rajani (2001)
	Integrated models	Kabir et al. (2018), Altarabsheh et al. (2018), Hawari et al. (2016)
Machine learning model	Artificial Neural Network	Alsaqqar et al. (2017)
	Decision tree-based models	Laakso et al. (2018), Caradot et al. (2018)
	Support Vector Machine	Hernández et al. (2021)

Chapter 3
Research Methodology

3.1. Conceptual framework

Figure 3.1 presents the conceptual framework for modeling sewer conditions in this thesis. The framework consists of three main interlinked steps (i) data collection and preparation using GIS tools, (ii) sewer condition assessment using ML algorithms, and (iii) visualization platform.

3.1.1. Data collection

In this thesis, sewer condition assessment models were employed using physical and environmental factors, and operational factors were not included because of unavailability at the time of the study. Physical factors were mainly collected from the existing database provided by the Ålesund city (tabular dataset) while environmental factors were obtained from multiple GIS-based sources.

3.1.2. Data preparation

All physical factors were vectorized, and environmental factors were rasterized using GIS tools. Physical factors were identified for each sewer pipe based on its unique name or index in the database. Based on the geographical vertices of each sewer pipe, its centroid point was calculated. After that, rasterized values of environmental factors were assigned to each pipe based on this centroid position and their unique name or index (Fan et al., 2022). Finally, the aggregated GIS database was transferred to the next step to process before feeding them into ML models.

3.1.3. Sewer condition assessment

In this step, all inspected sewer pipes with damage scores or damage classes were selected from the entire dataset or maintenance reports provided by the Ålesund city (Nguyen & Seidu, 2022). All physical and environmental factors of inspected sewer pipes were compared and assigned from the above GIS database. All pipes that did not have any one of the above factors were eliminated from the input dataset.

The input dataset was split into training and validation datasets, the training dataset was used to construct ML models while the validation dataset was used to verify and validate constructed ML models. Feature selection methods were implemented on the training dataset to assess the importance of physical and environmental factors. Consequently, the best ML model was used to predict the condition of sewer pipes in the study area.

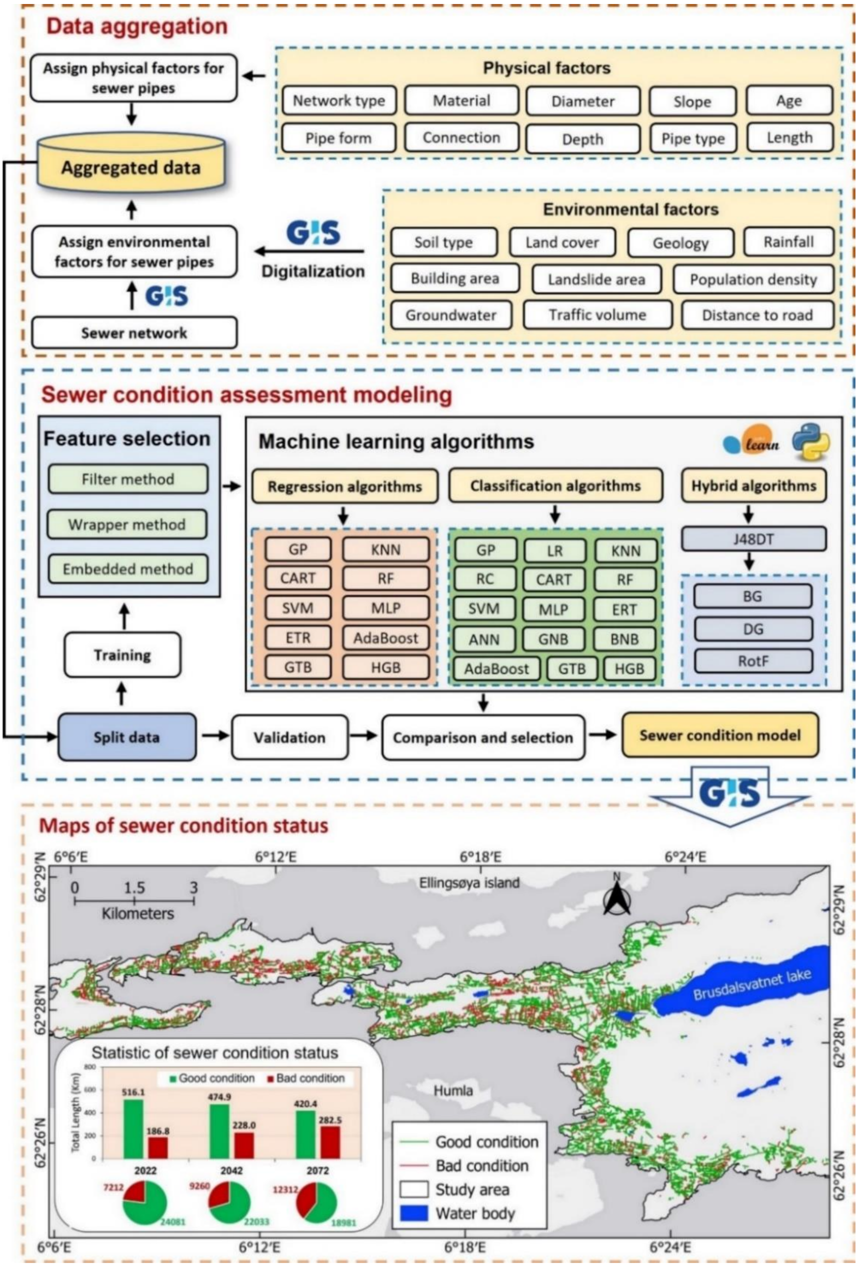


Figure 3.1. Overview of the sewer condition assessment modeling

3.1.4. Sewer condition visualization

The predicted sewer conditions obtained from the previous step were stored and re-assigned into the GIS database using the unique name/index. At this step, sewer conditions were visualized on a desktop using GIS software or on mobile/HoloLens using Unity software.

3.2. Summary of feature selection methods used

Defining significant factors is crucial to improve the models' performance and to reduce computational abundance. It is, therefore, critical in preprocessing step in pattern recognition, dimensionality reduction, and data mining (Kuhn & Johnson, 2013; Tashi et al., 2020).

Feature selection techniques are mainly clustered into the filter, wrapper, and embedded methods (Chandrashekar & Sahin, 2014). In filter methods, the optimal subset of variables is selected mainly based on their statistical properties and relationship with the target variable. The performance of variables is used to select a subset of features by removing and adding the subsets accordingly in the wrapper methods, and the model tuning process is applied in the embedded methods to perform feature selection (Chan et al., 2022).

However, different feature selection methods produce different results using the same input due to a random initialization strategy (Song et al., 2021). Consequently, their practical applications should be further investigated. This thesis applied the filter, wrapper, and embedded feature selection methods in determining the significant factors before constructing ML models.

3.3. Summary of machine learning algorithms used

This section briefly describes the fundamental theories of machine learning algorithms used in this thesis including regression, classification, and hybrid methods. The regression model describes a mapping function that defines the relationship between independent variables/features and a dependent/output variable. The target outcome in the regression models are typically numeric values and the predictive performance of these models is normally assessed using some statistical criteria such as the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE) (Nguyen & Seidu, 2022).

The classification model approximates the mapping function from given input variables to identify discrete output variables, which can be labels or categories. In those models, outputs typically need to be converted into dummy variables if the data contains noise/missing values (Sen et al., 2020). Accuracy, F1-score, or the area under the receiver operating characteristic (AUC-ROC) can be used to estimate the performance of classification machine learning models

(Nguyen et al., 2022).

Hybrid machine learning models combine model-based and data-driven learning systems to potentially capture more characteristics of complex systems to deal with over-fitting and under-fitting (Li et al., 2019). Therefore, recent studies have shown promising results in these models (Khayyam et al., 2015; Phong et al., 2021; Ramezankhani et al., 2019). These models can be used for both classification and regression, therefore, assessment criteria for two types of models are applied to assess the prediction performance of hybrid models (Machado & Karray, 2022; Rawat & Malhan, 2019).

3.3.1. Regression-based machine learning algorithms

a. Gaussian Process for regression

Rasmussen (2004) first introduced Gaussian Process (GP) model for dealing with classification and regression problems. In the GP model, the covariance function is determined using a single or a combination of kernel functions and their hyperparameters (Pedregosa et al., 2011).

For the regression problem, GP is an effective tool for interpolating data points in high-dimensional input space and can be defined as follows (Meng & Zhang, 2020):

$$Y(X) = GP\left(M(X_i), Cov\left(X_i, X_j\right)\right) + \epsilon(X), \qquad i, j = 1, \dots, n \tag{1}$$

where n is the total number of inspected sewer pipes, Y is the damage score, $\epsilon(X)$ is the observation error, $M(X_i)$ and $Cov\left(X_i, X_j\right)$ are the mean and covariance functions, respectively.

b. K-Nearest Neighbor for regression

Lall and Sharma (1996) introduced K-Nearest Neighbor (KNN) to deal with regression problems. For sewer condition prediction, the KNN model approximates the association between physical and environmental factors and the sewer damage score by averaging the observations in the same neighborhood. The grid-search method was used to calculate the number of nearest neighbors, and the distance metric was used to calculate the distance of one test observation from all the observations of the training dataset and find the nearest neighbors.

From the regression perspective, the KNN algorithm has quick computational time, easy interpretability, versatility, and no need for any assumptions (Yao & Ruzzo, 2006). However, this algorithm is sensitive to irrelevant features which can be addressed by feature selection. In addition, this algorithm may be ineffective with large datasets because it calculates and stores the distances from the new test point to all the training data points during implementation.

c. *Classification and Regression Trees*

Breiman et al. (1984) first proposed Classification and Regression Trees (CART) algorithm to solve regression and classification problems based on tree-based structures. In this method, the sewer dataset (also called the root node) was divided into subsets at each node using a series of recursive binary splits based on evaluating every possible predictor (Ebrahimy et al., 2020). Finally, the predicted sewer conditions were defined based on the most commonly occurring class of the node.

For classification problems, Gini Impurity or Entropy index can be used to split root/decision nodes. Otherwise, the "*goodness*" criterion is applied to split root/decision nodes into regression problems. Gini Impurity, Entropy, and the "*goodness*" criteria (normally mean squared error) were presented as follows (Daniel & Chantal, 2014; Rahmati et al., 2022; Yuan et al., 2021):

$$Gini(D,k) = \sum_{j=1}^{m} \frac{\left|D^j\right|}{n}\left(1 - \sum_{i=1}^{n} p_i^2\right) \tag{2}$$

$$Entropy = \sum_{i=1}^{c} -p_i^* log_2 p_i^* \tag{3}$$

$$f(s|t) = 2P_L P_R \sum_{i=1}^{n} |P(i|t_L) - P(i|t_R)| \tag{4}$$

where n is the number of sewer inspections in one dataset D; m is the total number of the subsets; c is the number of classes; D^j is one subset of dataset D classified based on attribute k; $\left|D^j\right|$ is the number of sewer inspections in the subset D^j; p_i is the probability of type i occurring in dataset D; p_i^* is the non-zero probability that the arbitrary rule belongs to class c; $f(s|t)$ is a measure of *"goodness of fit"*; t_L and t_R are the left and right children of a candidate split s at node t, respectively; P_L and P_R are the proportions of records at t_L and t_R, respectively; $P(i|t_L)$ and $P(i|t_R)$ are the proportions of class i at t_L and t_R, respectively.

d. *Random Forest for regression*

Random Forest (RF) regression is an ensemble learning method that uses multiple decision trees as base learning models for regression problems. The bagging (or bootstrap aggregation) algorithm is generally used to create the RF model. In this way, each random subset in the training dataset is selected with replacement to fit the decision trees with the corresponding sewer conditions. After the RF model was trained, damage scores for unseen sewer pipes can be made by averaging the predictions from all the individual regression trees (Kumar & Shaikh, 2017).

e. Multi-Layer Perceptron for regression

A Multi-layer Perceptron (MLP) can be normally used for regression and classification problems which consist of at least three layers (an input layer, an output layer, and one or more hidden layers) and each layer contains different neurons. For sewer condition prediction purposes, the number of neurons in the input layer equals the number of input factors. The number of neurons in the output layer is one (for regression problems) or the number of classes (for classification problems). The number of hidden layers and hidden neurons depends on the complexity of the MLP architecture.

Training an MLP network is normally implemented via followed steps (i) assign each factor for each neuron and add a bias unit into the input layer, (ii) generate a random weight for each neuron, (iii) calculate the sum of each neuron and transfer the results to the hidden layer employing activation function, (iv) calculate a similar process in the hidden layer, and transfer result to the output layer, and (v) recalculate the weights based on the rule minimizing the difference between the predicted damage scores and corresponding actual values (cost function). The above process is repeated until the given epochs are satisfied, or the ideal weights are obtained.

Each neuron j in the hidden layer computes its input signals x_i and produces its output y_j based on the following equation:

$$y_j = f\left(\sum_{j,i=1}^{n} w_{ji}x_i + b_i\right) \tag{5}$$

where n is the number of sewer inspections in the training dataset; f is an activation function; w_{ji} and b are connection weight and bias, respectively. Moreover, many additional hyperparameters such as transfer function, learning rate, number of epochs, dropout, and momentum are required while training an MLP network.

f. Support Vector Machine for regression

Support vector regression is one type of Support Vector Machine (SVM) used for regression problems. In this model, the best-fit hyperplane in n-dimensional space is determined to distinguish data samples in the training dataset in the best way (Trafalis & Ince, 2000). For regression problems, the linear form of the hyperplane can be computed as follows (Wauters & Vanhoucke, 2014):

$$f(x) = w.x + b \tag{6}$$

where $f(x)$ is the predicted value, x is the input vector of the data point, w and b are the slope and intercept. The ideal solution can be defined as follows (Smola & Schölkopf, 2004):

$$\begin{cases} \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b \\ subject\ to \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0, \qquad \alpha_i, \alpha_i^* \in [0, C] \end{cases} \tag{7}$$

where $f(x)$ is the predicted value, n is the number of sewer inspections, x is the input vector of the data point, w and b are the slope and intercept, respectively, α_i, α_i^* are Lagrange multipliers, the constant $C > 0$ is the trade-off between the flatness of the $f(x)$ and the amount up to which deviations larger than the insensitive loss function, $K(x_i, x)$ is the Kernel function, for example, linear function, polynomial function, radial basis function, or sigmoid function.

g. Extra Trees Regression

Extra Trees Regression (ETR) is an ensemble supervised machine learning method that uses decision trees for regression problems. This algorithm uses an entire learning sample (instead of bagging like RF) to split nodes by choosing cut points entirely randomly. This selection contributes to reducing the bias of the model (Geurts et al., 2006).

The relative variance reduction is used as the score measure in the regression problems for the ETR algorithm (Geurts et al., 2006):

$$Score(s, D) = \frac{Var(y|S) - \frac{|S_l|}{|S|} Var(y|S_l) - \frac{|S_r|}{|S|} Var(y|S_r)}{Var(y|S)} \tag{8}$$

where $Var(y|S)$ is the variance of the output y in the sample S, S_l and S_r are two subsets of cases from the sample S corresponding to the two outcomes of a split s, respectively.

h. AdaBoost

AdaBoost (also called Adaptive Boosting) was introduced by Freund and Schapire (1997) for regression and classification problems. This algorithm is an ensemble learning method using an adaptive resampling approach to improve predictive performance for controlling bias and variance from the mistakes of the base algorithm (Hong et al., 2018). In the AdaBoost model, iterations are performed to improve predictive performance by optimizing the target function. In the sewer condition regression problem, the target function is the minimum difference in damage scores between inspected and actual sewer conditions.

In the case of sewer condition prediction, AdaBoost randomly selects subsets from the sewer dataset. After that, these subsets were assigned equal weights to implement a classifier for each iteration. Higher weights will be reassigned for misclassified cases in the previous iteration, and a new iteration process continues after a new normalized training subset is created. The iterative process is terminated if specific stopping criteria are satisfied, and the sewer condition is the result of the weighted sum of all predictions.

The iterative process is ended until it reaches a terminated condition, and the final model is obtained from a weighted sum of all the base models. Although AdaBoost can be built based on various weak base learners, a combination of AdaBoost with the decision tree is often referred to as the best out-of-the-box classifier (Kégl, 2013).

i. Gradient Tree Boosting

Gradient Tree Boosting (GTB), which was introduced by Friedman (2002), is an ML technique used in regression and classification tasks. This technique improves predictive performance by combining weaker learners with strong learners via the iteration approach (Bentéjac et al., 2021).

For sewer condition prediction, a subset of the sewer dataset is randomly generated (without replacement) for each iteration. Each internal node of the tree denotes an attribute, and each leaf node denotes a predictive sewer condition. For each iteration, decision trees have been added repeatedly, and the next decision tree will correct the previous decision tree error (Ayyadevara, 2018). The final sewer condition status is obtained by minimizing the loss of function.

j. Histogram-based Gradient Boosting

Histogram-based Gradient Boosting (HGB), introduced by Guryanov (2019), is a modification of the GTB and can increase the learning process and the model's prediction performance. This method divides the sewer training dataset into bins and constructs a histogram of feature values during the training phase (Aljamaan & Alazba, 2020). The iteration process is stopped when the stopping condition (for example, the limit of tree depth or the number of leaves in the tree) is reached. Then, the sewer conditions are defined using the best-split points based on the feature histograms (Ke et al., 2017).

3.3.2. Classification-based machine learning algorithms

a. Gaussian Process for classification

For the classification problem, the sewer conditions were transformed into $\{-1, +1\}$, a latent function f was used to predict the class membership probability for a new test pipe. The value of the function f was then mapped into the $[0,1]$ interval using the probit function (Rodrigues et al., 2014). The predictive distribution of the sewer conditions can be calculated by getting the weights of all possible predictions by their calculated posterior distribution (Kuss et al., 2005):

$$p(y^* = 1|x^*, \boldsymbol{X}, \boldsymbol{y}) = \int_{f^*} \Phi(f^*) p(f^*|x^*, \boldsymbol{X}, \boldsymbol{y}) df^* \tag{9}$$

where $\boldsymbol{X} = [x_1, \dots, x_n]^T$ and $\boldsymbol{y} = [y_1, \dots, y_n]^T$ are vectors containing factors and sewer condition status, respectively; n is the number of sewer inspections; y^* and x^* are predicted sewer condition status and vector-containing factors of one sewer pipe, respectively; f^* and $\Phi(.)$ are variables corresponding to the test point x^* and the probit function, respectively.

b. K-Nearest Neighbor for classification

KNN can be used as an effective tool for dealing with classification problems (Cover & Hart, 1967). For sewer condition prediction, the distance from the sewer pipe x_i in the test dataset of each sample in the training dataset is computed. The top K points, which have the closest distance to x_i, are stored, and the status probability of sewer x_i is computed as follows (Fan et al., 2022):

$$P(y = D, X = x_i) = \frac{1}{K} \sum_{j \in A} I\left(y_j = D\right) \tag{10}$$

where $I\left(y_j = D\right)$ equals 1 if the instance y_j is in class D, otherwise, it equals 0, A is the dataset that contains K points, and D is the sewer conditions.

c. Random Forest for classification

Random Forest (RF) was developed by Breiman (2001) to significantly improve classification accuracy by creating an ensemble of trees and letting them vote for the most popular class. In the RF model, the sewer input dataset was randomly split into classification trees, and the model was trained through bagging or bootstrap aggregating. The final sewer's condition status was obtained by aggregating the prediction from each tree.

d. Multi-layer Perceptron for classification

Theoretically, training an MLP network in a classification problem is similarly performed as in a regression problem. The main difference is that the output layer in classification problems has more than one neuron. For multi-classification problems, a special activation function such as softmax is used to identify the probability in each class (Malik et al., 2022).

e. Support Vector Machine for classification

Cortes and Vapnik (1995) first proposed SVM for dealing with classification problems. In the case of sewer condition prediction, the sewer conditions were determined by maximizing the distance from the hyperplane to the data points. The hyperplanes can be computed as follows (Zendehboudi et al., 2018):

$$\begin{cases} y_i(\boldsymbol{w}.\emptyset^T(x_i) + \boldsymbol{b}) \geq 1 - \varepsilon_i \\ \varepsilon_i \geq 0, \qquad i = 1,2,\dots,n \end{cases} \tag{11}$$

where n is the number of inspected pipes, $x_i \in \mathbb{R}^n, y_i \in \mathbb{R}$, x, and y are vectors that contain input factors and sewer conditions respectively, $\boldsymbol{w}$ is the coefficient vector, $\boldsymbol{b}$ is and bias of the hyperplane in the feature space, $\emptyset$ is the non-linear mapping function, and ε_i is the positive slack variable. The predicted condition status of the sewer pipe using the SVM is calculated as follows (Cervantes et al., 2020):

$$\begin{cases} f(x) = sign\left(\sum_{i=1}^{n} \alpha_i y_i K(x_i, x) + b\right) \\ \sum_{i=1}^{n} \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \ \forall i = 1,2,\dots,n \end{cases} \tag{12}$$

where auxiliary variables α_i are Lagrange multipliers, C is the regularization parameter, and $\mathrm{K}(x, x_i)$ is the Kernel function.

f. Logistic Regression

Logistic Regression (LR) predicts the probability of the sewer conditions based on their relationship with input factors. The maximum likelihood method is generally used to estimate the intercept and coefficients based on the factors and sewer conditions. This method maximizes the probability of the sewer status given the fitted regression coefficients (Kuk & Chen, 1992).

Although the LR algorithm was originally designed for regression problems, this method was commonly used for classification problems (especially for binary classification) (Dokeroglu et al., 2021; Książek et al., 2021). This thesis investigates the potential application of the LR algorithm for sewer condition prediction in the study area and the results from this method will also be compared with those from other ML models.

g. Extremely Randomized Trees

Extremely Randomized Trees (ERT), proposed by Geurts et al. (2006), is an ensemble supervised machine learning method. In this method, many decision trees are created randomly without replacement (the individual sewer data points are chosen only one time). The most important and unique characteristic of this method is to randomly select a split value instead of calculating locally optimal values using Gini or Entropy to split data (Geurts et al., 2006). By using this approach, generated decision trees are diversified and uncorrelated.

After decision trees are trained, the sewer conditions predicted by single trees are aggregated to yield the final result.

h. Gaussian Naive Bayes

Naive Bayes is a probabilistic ML algorithm based on the Bayes theorem, which assumes that the features have strong independence from each other. Although this algorithm is simple, it is significantly more accurate than sophisticated methods, especially for classification problems (Frank et al., 2000). Moreover, one of the most advantages of this algorithm is that no hyperparameters are required to tune the model (Askari et al., 2020).

Gaussian Naive Bayes (GNB) classifies sewer status based on an assumption of having a Gaussian distribution on input factors (Jahromi & Taheri, 2017). Sewer conditions can be predicted using the Gaussian probability density function by substituting the parameters with the new input values (Cataldi et al., 2021):

$$p(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(x_i-\mu_y)^2}{2\sigma_y^2}} \tag{13}$$

where σ_y and μ_y are the variance and mean of the feature i^{th}, respectively; class y contains sewer conditions.

i. Bernoulli Naive Bayes

The Bernoulli Naive Bayes (BNB) classifies sewer status based on the Bayes theorem using sewer input data that are distributed according to multivariate Bernoulli distributions. Sewer conditions predicted by BNB are made based on the rule as follows (Pedregosa et al., 2011):

$$P(x_i|y) = P(i|y)x_i + \big(1 - P(i|y)\big)(1 - x_i) \tag{14}$$

where $P(x_i|y)$ is the likelihood of the features, x_i is the vector contains the input factor of the feature i^{th}, and y is sewer class.

j. Ridge classification

Ridge Classification (RC) algorithm is developed based on the Ridge regression, which was introduced by Hoerl and Kennard (1970) for solving the multicollinearity problem of covariates in samples. This method assumes that samples from each sewer condition class belong to a linear subspace, and a new test sample can be represented as a linear combination of class-specific training samples (He et al., 2014). This method has been widely applied in diverse applications in chemistry, econometrics, and engineering to deal with multicollinear data because of the small values of its variance and its mean square error (Gruber, 1998).

RC algorithm converts the conditions of sewer pipes into $[-1, +1]$ and solves the problem as a regression task, minimizing the size of the coefficients by imposing a penalty, and the sewer condition class is assigned based on the highest value of the prediction result.

3.3.3. Hybrid machine learning algorithms

The hybrid ML method is a technique that combines the ensembles with the base ML algorithm to make more accurate predictions. This section presents three variants for hybrid ML models used in this thesis: Bagging (BG), Dagging (DG), and Rotation Forest (RotF).

a. Base classification algorithm J48 Decision Tree

Lim et al. (2000) reported that the C4.5 algorithm is the fastest algorithm for building decision trees with reasonable accuracy. J48 Decision Tree (J48DT), a Java version of the C4.5 algorithm in the Waikato Environment for Knowledge Analysis (WEKA), was used in this thesis as the base learner for BG, DG, and RotF ensembles. The reason for choosing this algorithm is that it is an effective technique for classification using ensemble methods (Hong et al., 2018).

b. Bagging hybrid algorithm

The BG algorithm raises the stability of models significantly classification problems by improving accuracy and reducing variance (Breiman, 1996). In this algorithm, new sewer pipe points are created by randomly selecting samples with replacements (specifically, the individual sewer data points can be chosen more than one time) from the original training dataset. These subsets are used to train base learners independently. Finally, the final sewer conditions are defined using a plurality vote of those predictions from the base models.

c. Dagging hybrid algorithm

The DG algorithm creates random training subsets from the original training dataset using the disjoint sampling method (instead of the bootstrap sampling) without replacement (Ting & Witten, 1997). The base learners are trained using the above subset, and the sewer conditions are defined using a plurality vote from the individual predictions.

d. Rotation Forest hybrid algorithm

The RotF algorithm was first introduced by Rodriguez et al. (2006) based on the idea of a random forest algorithm to improve the diversity and accuracy of the base classifier. In this method, a feature extraction technique, namely Principal Component Analysis (PCA), is used to create bootstrap sampling and train the base classifier. Hyperplanes parallel to the feature axes are used to create classification regions while training base learners. The final sewer conditions are computed based on the largest confidence for each status (Kuncheva & Rodríguez, 2007).

3.3.4. Comparison of machine learning models

Various studies have assessed the performance of regression-based ML models for predicting the condition of sewer systems (Salihu et al., 2022). By using an ANN model to predict the rate of sewer pipes deterioration, Najafi and Kulandaivel (2005) concluded that this kind of model exhibited a good learning tendency. Similar conclusions were made by studies applying different regression-based ML models in sewer condition assessment (e.g., BPNN or SVM) (Khan et al., 2010; Sousa et al., 2014). However, these studies showed that outliers/noises affected the performance of the models.

Classification-based ML models also showed good performance in resolving classification problems including both binary and multiple classes in sewer condition assessment. Harvey and McBean (2014) found that RF was an excellent candidate for binary classification in sewer condition assessment. RF model was also applied in the multi-class classification of sewer condition in the study of Vitorino et al. (2014). Other studies have also explored the potential of hybrid ML models for solving classification problems. For example, Salihu et al. (2022) recommended hybrid models need to be developed in order to curtail current models' limitations. Although hybrid models have been recently applied for classification purposes and gained incredible performances, they are mainly considered for studies in other domains such as business (Machado & Karray, 2022), natural disaster (Bui et al., 2022; Masrur Ahmed et al., 2021; Miraki et al., 2019; Ngo et al., 2021; Pham et al., 2020; Wang et al., 2019), or education (Rawat & Malhan, 2019). The application of hybrid ML models in sewer condition assessment is still limited and need to be further explored.

Although many studies have investigated the application of regression-based and classification-based ML algorithms for sewer condition assessment, there is rarely any study considering both approaches to estimate condition of sewer pipes for a specific study area. In this study, a comprehensive assessment of sewer condition is undertaken using regression-based, classification-based and hybrid ML models.

3.4. Visualization platform

Visualization platform, or visual communication, supports human problem-solving and improves user decision-making performance by transforming non-visual objects into visual objects that are accessible to the human mind (Dübel et al., 2014). An effective visualization platform will support managers in having a visual overview and correctly evaluating the system status to inform reasonable maintenance strategies (Beha et al., 2015). An augmented reality (AR) was implemented to visualize pipe conditions in this study.

AR is a technology that integrates digital/virtual information with the user’s environment in real time. By using this technique, the visual perspectives of the physical real-world environments are enhanced by using non-visual properties and means of computing devices (Bottani & Vignali, 2019). With the massive improvements in computing power of computer software and hardware in the Fourth Industrial Revolution era, AR is becoming one of the most promising technologies in the future that supports people to recognize and experience real-world objects in a completely new way (Y. Chen et al., 2019).

Various studies have shown that AR technology is an effective supporting tool in product design, manufacturing, maintenance/inspection, and training activities (Fite-Georgel, 2011). However, the application of AR technique for visualization in the water domain is still challenging and limited (Centeno et al., 2009; Haynes et al., 2018; Mirauda et al., 2017; Schall et al., 2013). The main aim of this work was to develop a mobile application integrated with the AR technique to support sewer management through 2D/3D visualization of sewer conditions. By combining sewer conditions produced from predictive models and AR technology, water engineers/managers can quickly assess a sewer's status in the field and reduce workloads compared to conventional methods such as digging or camera-based inspection.

The integrated AR condition assessment visualization platform developed in this study consists of three main segments, indoor, intermediate, and outdoor segments, presented in **Figure 3.2**.

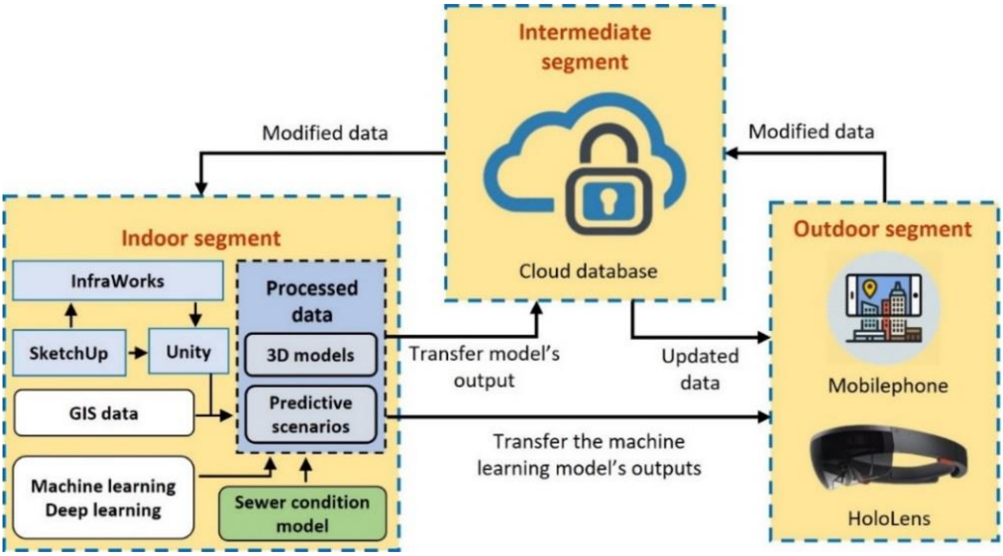


Figure 3.2. Overview of the integrated visualization platform

The indoor segment involves 3D modelling and computation, focusing on the process implemented on personal and supercomputers. In this regard, 3D models of the sewer network (e.g., pipes, manholes, or pumps) were created and visualized using a high-performance computer system associated with the simulation programs BIM or GIS software (Fenais et al., 2018; Han et al., 2019). Sewer conditions were predicted using ML and DL models based on input data, as described in above. The network components' output was coded based on their corresponding indexes for visualization purposes. While creating 3D objects, the names and indexes of original objects were maintained to merge with information obtained from the previous step. This information was reused if any new update was received from the intermediate segment. Finally, the models and auxiliary data were transformed into Unity 3D for simulation on smart devices such as smartphones and HoloLens.

The intermediate segment involves the processing of remote data. This segment employed cloud-based platforms to store data and have remote interactive connections with the computer or handheld devices. In this segment, the database created from the indoor segment was transferred to the host computer system on a cloud-based connection. Real-time data received from sensors in the sewers were updated and stored on the ***StaalCloud*** portal managed by Ålesund municipality. In this segment, the data will undergo initial processing to generate fundamental network details such as pipe index, velocity, and sewer condition status. After that, these generated attributes are assigned to corresponding objects based on their unique indexes, which are generated from the indoor segment. Following this process, the data can be transmitted to the outdoor segment for display on computer or handheld devices. Finally, the intermediate segment is configured to receive updated information from the outdoor segment, store it, and send it back to the indoor segment to update the system.

The outdoor segment involves visualizing and updating data. This segment contains devices that can visualize objects and their attributes in a hands-free manner. In this segment, the processed data obtained from the indoor segment and updated data received from the intermediate segment are used to enhance the user's experience via AR devices. The application directly reads information from computers or the cloud via a Wi-Fi network, matching them with corresponding objects and showing designated information. Any modification can be performed, and modified data in this segment can be transmitted back to the host computer in the indoor segment or the cloud database in the intermediate segment using an application programming interface (API) or equivalent protocols to update the database.

For the 3D visualization development platform, the Unity 3D game engine (version: 2021.3.15f1 Long Term Support), Android Studio (version: 2022.1.1), Java (version: 8), and C# (version: .NET Core 3.0, C# 8.0) programming languages were selected to develop and compile the application on mobile and Microsoft HoloLens devices.

In terms of hardware for running these applications, the Samsung Galaxy A42 5G (Android 10, 4 GB RAM, Qualcomm Snapdragon 690) (Samsung, 2022) and Microsoft HoloLens (Windows 10, 64 GB Flash, 2 GB RAM) (Microsoft, 2021) were used. Except for the shape of objects (in 3D type), other attribute-related data were structured in the comma-separated values (CSV) format that is easily opened and modified by a text editor such as Notepad or Microsoft Excel.

3.5. Implementation of condition assessment models and visualization platform

3.5.1. Description of the study area

Ålesund city is one of the municipalities of Møre og Romsdal County, which is in the northernmost part of Western Norway. The area of the city is approximately 607.3 km^2, and lies between latitudes of 6°05'08" N and 6°40'56" N, and longitudes of 62°25'07" E and 62°30'37" E (**Figure 3.3**).

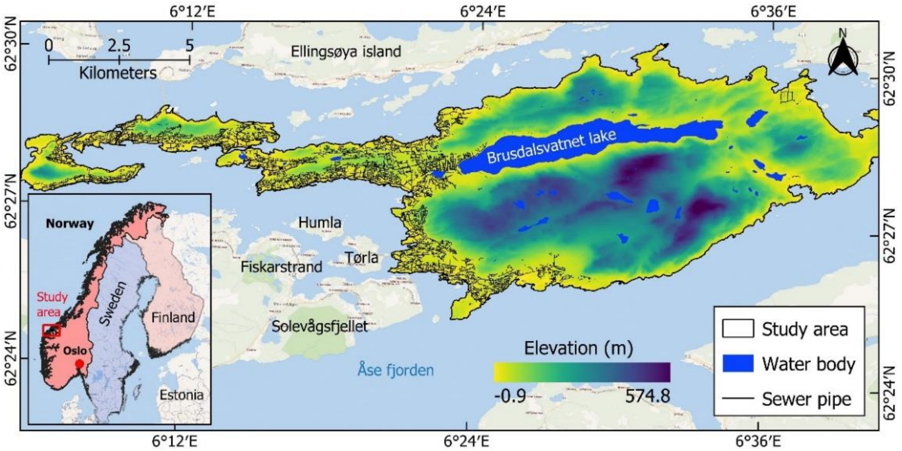


Figure 3.3. Sewer network in Ålesund city

The climate of the city is heavily influenced by ocean currents with cold, rainy winters and cool summers. The city is in high rainfall density region with an average rainfall of 2100 mm per year, and the variation in temperatures throughout the year is 13.6 °C, with average temperatures of the coldest month (February) and the warmest month (August) being -0.6 °C and 13.0 °C, respectively (Climate, 2021). Because it is surrounded by the ocean, the city has high humidity ranging from 77.5 % to 83.7 % in May and August, respectively. Along with the general trend of climate change, the weather in the city is affected by unavoidable fluctuations in temperature, precipitation, and extreme weather events that put pressure on the sewer network (Kvitsjøen et al., 2021).

According to Stian et al. (2021), the total investment fee until 2040 for managing and operating the sewage network is significantly increasing in Møre og Romsdal. Consequently, Ålesund city, as a part of Møre og Romsdal County, needs to develop the establishment of maintenance plans to effectively manage and reduce the investment fee for the wastewater/stormwater network. Therefore, an effective sewer condition assessment model will be a useful tool for local agencies to predict the future status of the sewer network.

3.5.2. Data used

In this thesis, based on the availability of data and related literature (Hawari et al., 2020; Roghani et al., 2019), ten physical factors (for example, age, material, depth, slope, diameter, length, pipe type, network type, pipe form, and connection type) and ten environmental factors (for instance, rainfall, geology, landslide area, building area, population, land cover, groundwater level, traffic volume, distance from the road, and soil type) were used. It should be noted that operational factors (for example, flow rate, blockages, infiltration, and inflow) were not considered because of data unavailability at the time of conducting this work.

a. *Physical factors*

The age of sewer pipes was calculated as the difference between the installation year and the inspection year. This numeric factor was assigned for each sewer pipe and updated for each maintenance time. Other factors, such as material, diameter, length, pipe type, network type, pipe form, and connection type, were obtained from the tabular datasets provided by Ålesund city.

Depth and slope were extracted from a Digital Elevation Model (DEM) with a spatial resolution of 5 m × 5 m. The DEM was generated from the Norwegian Mapping Authority (NMA) (https://www.kartverket.no/en) via the **høydedata** portal (https://hoydedata.no/LaserInnsyn/). Specifically, the depth of individual sewer pipes was calculated as the average value of all manholes' height associated with the specific sewer pipe. Then, these depths were assigned a negative value indicating that the pipe is below the ground surface.

The value of the slope is positive, reflecting that the inverted elevation of the start manhole is higher than the end manhole and vice versa. A summary of the physical factors is shown in **Table 3.1**.

Table 3.1. Summary of the physical variables

Physical variables	Type	Min	Max	Average	Std
Age (year)	Numeric	1.0	104.0	34.4	25.3
Diameter (mm)	Numeric	110.0	1000.0	248.4	98.6
Depth (m)	Numeric	-7.8	-0.1	-1.8	1.2
Length (m)	Numeric	1.0	177.5	38.6	21.3
Slope (º)	Numeric	-17.4	+34.6	+2.7	4.4
Pipe type	Categorical	-	-	-	-
Network type	Categorical	-	-	-	-
Pipe form	Categorical	-	-	-	-
Connection	Categorical	-	-	-	-
Material	Categorical	-	-	-	-

b. *Environmental factors*

Environmental factors used in this thesis are mainly extrinsic elements that relate to relative geo-location with sewers. The data was collected from many sources (**Table 3.2**) with different formats and spatial resolutions. Therefore, they were processed to transfer into the same coordinate system, format, and spatial resolution. In this thesis, post-processed data were re-sampled to the aforementioned spatial resolution (5 m × 5 m) and transformed into a grid spatial database. Spatial maps of the environmental factors are shown in **Figure 3.4**.

Table 3.2. Summary of the environmental factors

Data	Data source	Accessed link	Accessed date
Rainfall	NCSC	https://klimaservicesenter.no	14.02.2020
Geology	NMA	https://www.kartverket.no/en	10.01.2020
Landslide area	NMA	https://www.kartverket.no/en	12.01.2020
Population	NMA	https://www.kartverket.no/en	08.02.2020
Land cover	COAH	https://scihub.copernicus.eu	17.01.2020
Building area	NMA	https://www.kartverket.no/en	10.01.2020
Groundwater	NGS	https://www.ngu.no	24.04.2020
Traffic volume	NPRA	https://www.vegvesen.no/en	13.02.2020
Distance to road	NMA	https://www.kartverket.no/en	01.03.2020
Soil type	NMA	https://www.kartverket.no/en	10.02.2020

Rainfall data were obtained from annual average rainfall over many years at nine weather stations near the study area. The Inverse Distance Weighting (IDW) method, which is the most common spatial interpolation method (Ajaj et al., 2018), was used to construct the rainfall map for the study area.

Land cover map was derived from the Sentinel-2 images Level 1C downloaded from the website of Copernicus Open Access Hub (COAH). Detailed information on satellite images is presented in **Table 3.3** and **Table 3.4**. A Google background satellite image was put on the Sentinel-2 image to get the land cover classifications such as forest areas, roads, and residential areas. The samples of different land covers were taken and assigned specific values, and different bands of the Sentinel-2 image overlapped. Finally, object-based classification was applied to cluster areas in the image into different objects based on given land covers (Sánchez & Schröder, 2019).

Table 3.3. Satellite images used in the study area

Image name	Band	Color	Spatial resolution
T32VLQ_20191108T111251_B02.jp2	02	Blue	10 m × 10 m
T32VLQ_20191108T111251_B03.jp2	03	Green	10 m × 10 m
T32VLQ_20191108T111251_B04.jp2	04	Red	10 m × 10 m

Table 3.4. Auxiliary information of satellite images

Parameter	Explanation
Satellite name	Sentinel-2
Satellite number	A
Acquisition date	08.11.2019
Processing level	Level-1C
Cloud cover (%)	4.798
Degraded ancillary data (%)	0.0
Format correctness	PASS
Geometric quality	PASS
General quality	PASS
Orbit number (start)	22871
Pass direction	DESCENDING
Radiometric quality	PASS
Relative orbit (start)	137
Sensor quality	PASS
Tile identifier	32VLQ
Tile Identifier horizontal order	VQ32L
Instrument abbreviation	MSI
Instrument mode	INS-NOBS
Instrument name	Multi-Spectral Instrument
NSSDC identifier	2015-028A

The groundwater map was processed from 31 drill data around the study area using the IDW method. Different road classes were derived from the road network. We consider a 5m-range road distance for the first road class; larger distances can be accepted for classifying further pipes into different road classes. In this work, five ordinal road classes were used based on the road's buffers of 0-5 m, 5-10 m, 10-20 m, 20-50 m, and >50 m (Laakso et al., 2018).

Other maps such as geology, landslide area, population, land use, building area, traffic volumes, and soil type, were rasterized based on their different layers from the original properties. Finally, all GIS database was converted to raster format with a grid size of 5 m × 5 m in the WGS84-UTM32N (EPSG:32632) coordinate system.

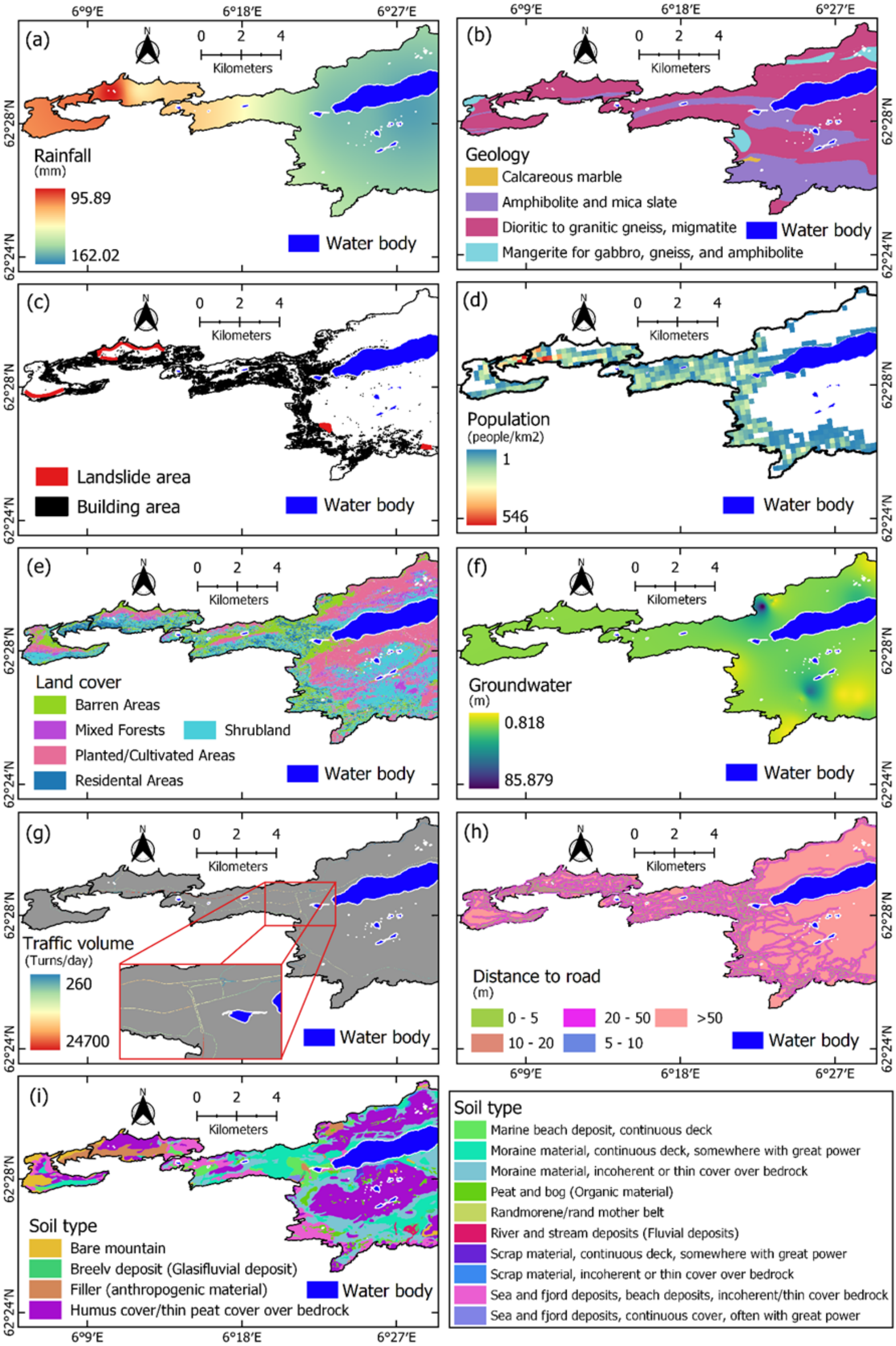

6°9′E
6°18′E
6°27′E
62°28′N
62°24′N
(a)
N
0 2 4
Kilometers
Rainfall
(mm)
95.89
162.02
Water body
(b)
Geology
Calcareous marble
Amphibolite and mica slate
Dioritic to granitic gneiss, migmatite
Mangerite for gabbro, gneiss, and amphibolite
Water body
(c)
Landslide area
Building area
Water body
(d)
Population
(people/km2)
1
546
Water body
(e)
Land cover
Barren Areas
Mixed Forests
Shrubland
Planted/Cultivated Areas
Residental Areas
Water body
(f)
Groundwater
(m)
0.818
85.879
Water body
(g)
Traffic volume
(Turns/day)
260
24700
Water body
(h)
Distance to road
(m)
0 - 5
20 - 50
>50
10 - 20
5 - 10
Water body
(i)
Soil type
Bare mountain
Breelv deposit (Glasifluvial deposit)
Filler (anthropogenic material)
Humus cover/thin peat cover over bedrock
Water body
Soil type
Marine beach deposit, continuous deck
Moraine material, continuous deck, somewhere with great power
Moraine material, incoherent or thin cover over bedrock
Peat and bog (Organic material)
Randmorene/rand mother belt
River and stream deposits (Fluvial deposits)
Scrap material, continuous deck, somewhere with great power
Scrap material, incoherent or thin cover over bedrock
Sea and fjord deposits, beach deposits, incoherent/thin cover bedrock
Sea and fjord deposits, continuous cover, often with great power

Figure 3.4. Maps of the environmental factors

c. Sewer condition

Sewer condition data was obtained from the closed-circuit television (CCTV) dataset and the Gemini VA, which is the market-leading solution in Norway for the management and documentation of the water and sewage network (https://www.volue.com/product/gemini-va). The conditions of the sewer pipe are assigned via damage numerical scores (for example, from 0 to 5 for class 1, from 6 to 10 for class 2) using the CCTV data. Next, these damage scores are coded into damage classes representing the sewer conditions. According to Haugen and Viak (2018), sewer conditions in Norway are classified into five-grade scales based on their damage scores (**Table 3.5**). In this thesis, a total of 1449 individual pipelines were used to model the sewer conditions in the study area. All damage classes were processed and integrated into a GIS database.

Table 3.5. Sewer condition dataset

Damage class	Damage score	Sewer condition
Class 1	0 - 5	Very good status
Class 2	6 - 10	Good status
Class 3	11 - 20	Questionable status
Class 4	21 - 50	Bad status
Class 5	>50	Very bad status

3.5.3. Machine learning implementation

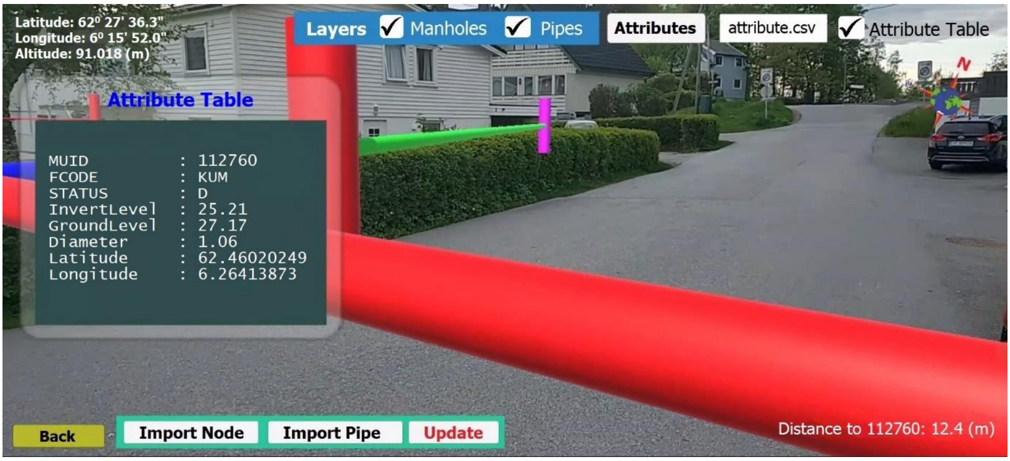
The ML models developed in this thesis were mainly implemented using the Keras and Scikit-learn libraries. Moreover, several extra libraries, such as pandas and NumPy, were used to process and organize the structure of the data. The Keras is an open-source software library that is developed for Artificial Neural Networks (ANN) and Deep Learning (DL) (Zhang et al., 2018). This library provides a Python interface and supports multiple backends, including TensorFlow, Theano, and Microsoft Cognitive Toolkit (Conlin et al., 2021). Keras with TensorFlow backend, an open-source software developed by Google for DL (Abadi et al., 2016; Ketkar, 2017), was used in this thesis.

The Scikit-learn library is a Python module integrating a wide range of state-of-the-art ML algorithms. This library provides a user-friendly and consistent interface that supports implementing ML algorithms more efficiently and productively (Raschka & Mirjalili, 2019).

Before constructing ML models, the data were divided into a training dataset and a validation or testing dataset with a ratio of 80 % and 20 %, respectively. Before being trained, the input factors were scaled in the range [0, 1] due to normalizing the data, while training generally speeds up learning and leads to faster convergence (Lee et al., 2020). For regression models, the damage scores of sewer pipes were normalized in the range [0, 1]. In contrast, the damage classes were coded using the one-hot encoding technique in the classification method. After the models were trained, the final results were rescaled to get the real values.

Hybrid ML models were implemented using Waikato Environment for Knowledge Analysis (WEKA) which is an open-source data mining software toolbox developed at the University of Waikato, New Zealand. WEKA is a collection of ML algorithms containing tools for data preparation, classification, regression, clustering, association rules mining, and visualization (Witten et al., 1999). This software is used worldwide because of its flexibility and effectiveness (Alonso & Bugarín, 2019).

3.5.4. 3D visualization of pipe conditions



Latitude: 62⁰ 27' 36.3"
Longitude: 6⁰ 15' 52.0"
Altitude: 91.018 (m)

Attribute Table

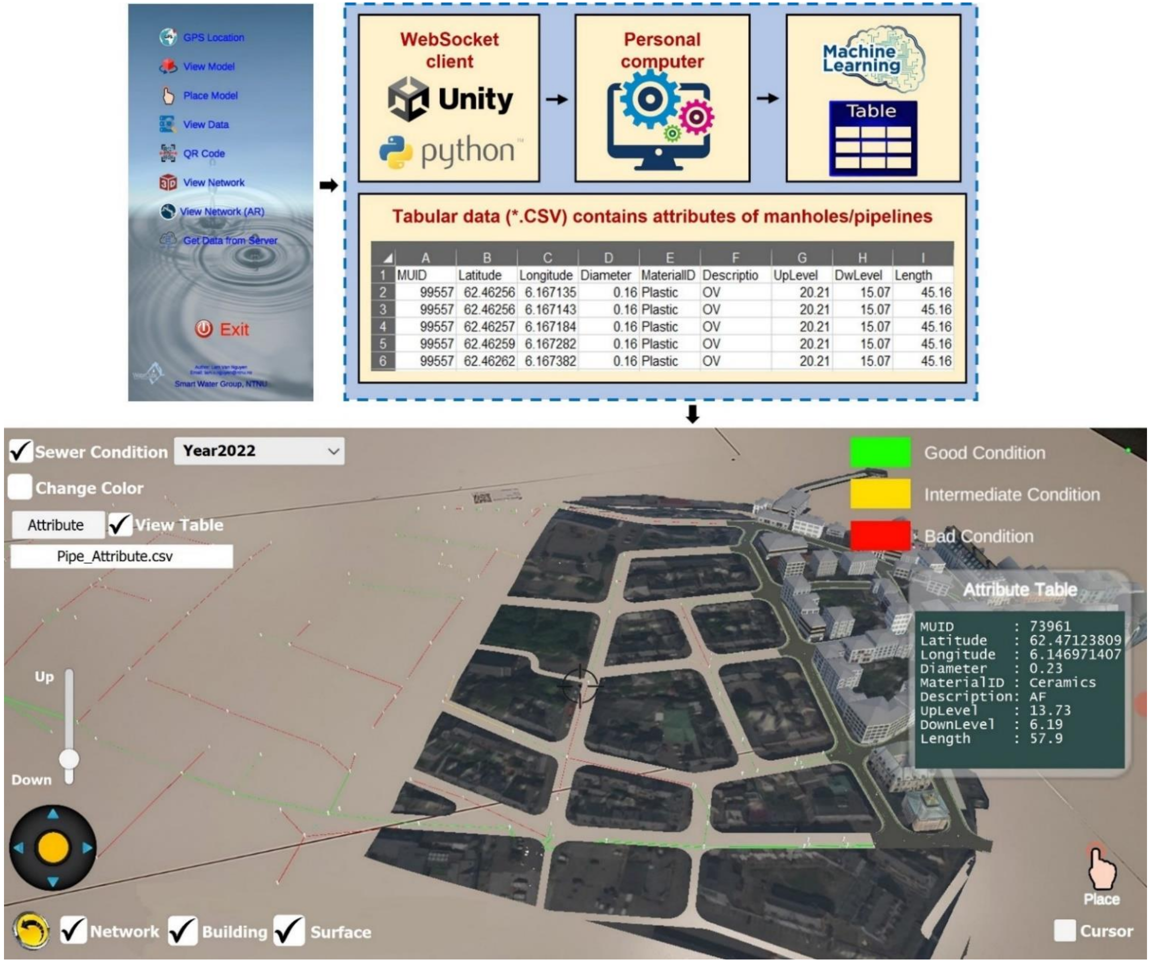
```
MUID          : 112760
FCODE         : KUM
STATUS        : D
InvertLevel   : 25.21
GroundLevel   : 27.17
Diameter      : 1.06
Latitude      : 62.46020249
Longitude     : 6.26413873
```

Distance to 112760: 12.4 (m)

Figure 3.5. An augmented reality visualization

To build AR visualization applications, the Unity cross-platform was used in this thesis. Unity is a professional-quality game engine targeting a variety of platforms, including mobile platforms (Android, iOS, or tvOS), desktop platforms (Windows, Mac, or Linux), web platforms (WebGL), console platforms (PlayStation or Xbox), and virtual or extended reality platforms (Windows Mixed Reality, Oculus, or PlayStation Virtual Reality) (Juliani et al., 2018). This is one of the most accessible and free modern tools for game developers (Hocking & Schell, 2022). **Figure 3.5** illustrates an AR function for network visualization developed in this thesis.

Moreover, Trimble SketchUp and Autodesk InfraWorks were used to create 3D models. SketchUp owned by Trimble, Inc. (Lewis & Hampton, 2015) has been widely applied for visualizing and analyzing many real-world applications (Burner et al., 2018; Jusuf et al., 2017). In this thesis, Trimble SketchUp was mainly used for designing and initiating 3D models of the water network, such as manholes and other objects (for example, buildings or terrain), before importing them into Autodesk InfraWorks software. InfraWorks developed by Autodesk, Inc. is one of the most widely used software in building information modeling (BIM) environments for the planning and designing of infrastructure projects such as sewer networks. Integration of the above software, sewer condition assessment model, and wireless interaction is presented in **Figure 3.6**.



Tabular data (*.CSV) contains attributes of manholes/pipelines

	A	B	C	D	E	F	G	H	I
1	MUID	Latitude	Longitude	Diameter	MaterialID	Descriptio	UpLevel	DwLevel	Length
2	99557	62.46256	6.167135	0.16	Plastic	OV	20.21	15.07	45.16
3	99557	62.46256	6.167143	0.16	Plastic	OV	20.21	15.07	45.16
4	99557	62.46257	6.167184	0.16	Plastic	OV	20.21	15.07	45.16
5	99557	62.46259	6.167282	0.16	Plastic	OV	20.21	15.07	45.16
6	99557	62.46262	6.167382	0.16	Plastic	OV	20.21	15.07	45.16

☑ Sewer Condition Year2022

□ Change Color

Attribute ☑ View Table

Pipe_Attribute.csv

Attribute Table

```
MUID        : 73961
Latitude    : 62.47123809
Longitude   : 6.146971407
Diameter    : 0.23
MaterialID  : Ceramics
Description: AF
UpLevel     : 13.73
DownLevel   : 6.19
Length      : 57.9
```

Figure 3.6. Integrated platform for visualization

Chapter 4
Summary of Results

4.1. Paper I. Application of Regression-based Machine Learning Algorithms in Sewer Condition Assessment for Ålesund City, Norway

Ten state-of-the-art ML algorithms for predicting the damage score of sewer networks were successfully developed for the study area using ten physical factors (for example, age, diameter, depth, slope, length, pipe type, material, network type, pipe form, and connection type) and ten environmental factors (for example, rainfall, geology, landslide area, population, land use, building area, groundwater, traffic volume, distance to road, and soil type). The highest prediction capability was provided by the Extra Trees Regression model (R^2 = 0.90), followed by the Gaussian Process Regression model (R^2 = 0.86). The prediction capability of the Multi-layer Perceptron model (R^2 = 0.10) and the K-Nearest Neighbor model (R^2 = 0.07) was the worst among the developed ML models.

This study also shows that age and material are the most sensitive factors affecting the sewer conditions in the study area. Moreover, resampling techniques such as adaptive and gradient boosting techniques were unsuitable approaches for developing ML models in the study area. Interestingly, using damage scores of sewer pipes to predict their status can lead to fluctuation of reliability of developed ML models due to the skewness of damage score data. The study also recommended that the classification-based approach be considered to predict sewer conditions in the future.

4.2. Paper II. Comparison of Machine Learning Techniques for Condition Assessment of Sewer Network

This study is a continuation of the work presented in **Paper I**. In this work, nine physical factors, such as age, diameter, depth, slope, length, pipe type, material, pipe form, and connection type, and ten environmental factors, including rainfall, geology, landslide area, population, land use, building area, groundwater, traffic volume, distance to road, and soil type, were used to construct seventeen ML models. The performance of developed ML models was compared using the ROC curve, area under the ROC curve (AUC-ROC), and accuracy (ACC). The result shows that the Random Forest (AUC-ROC = 77.6 % and ACC = 78.3 %) is a sensitive model for predicting the condition of sewer pipes in the study area.

Based on the Random Forest model, the sewer status in the present time (2022) and two different future scenarios (2042 and 2072) were predicted in this study. In these scenarios, the changes in the most popularly dynamic factors, such as rainfall, groundwater, and population in the study area, were considered. The study also shows that aggregating sewer damage scores into classes can improve accuracy in predicting sewer status compared to using damage scores. It is recommended for the local water utilities that sewer classes should be considered instead of damage scores when assessing sewer conditions in the study area.

4.3. Paper III. Predicting Sewer Structural Condition Using Hybrid Machine Learning Algorithms

This work presents the potential application of hybrid ML models in predicting sewer conditions. In this work, three hybrid ML models, which are the combination of Bagging (BG), Dagging (DG), and Rotation Forest (RotF) ensembles with a J48 Decision Tree (J48DT) based classifier, were used to predict multiple sewer pipe conditions (for example, bad, intermediate, and good conditions).

The results show that the RotF-J48DT model provides the highest prediction capability with the weighted average AUC-ROC (WA-AUC-ROC) being 83.3 %, followed by the BG-J48DT model (WA-AUC-ROC = 81.6 %), the DG-J48DT model (WA-AUC-ROC = 79.5 %), and the base classifier J48DT model (WA-AUC-ROC = 74.0 %). The results of this study show that the hybrid models are better than the single classifier. It is worth noticing that this study used three sewer classes to construct the hybrid ML models, using a binary classification approach (good and bad classes), such as used in the study in **Paper II** can improve the prediction capability.

4.4. Paper IV. Utilization of Augmented Reality Technique for Sewer Condition Visualization

This work is conducted as a visualization part for studies. A framework for integrating AR techniques, GIS, 3D-creation platform, and sewer’s conditions produced from the above machine learning algorithms is developed and implemented on an Android OS and Microsoft HoloLens. By combining sewer conditions produced from predictive models and AR technology, water engineers/managers can quickly estimate a sewer's status in the field and reduce workloads compared to conventional methods such as digging or camera-based inspection.

The outputs of this work reveal the potential integration of mobile devices, GIS, and AR techniques in the management of water infrastructure. The water engineers/managers can collect data on pipe and environmental attributes, processing, condition assessment, and visualization of pipe status as well as their attributes on a mobile device. The real-time visualization of dynamic data (e.g., water flow or water temperature) of the pipes through integration of geo-pipe locations and sensor data can be implemented. The users' awareness of water infrastructure and the surrounding environment is enhanced in a way that allows the exploration of the relations between them. A limitation of the application is the accuracy of the visualized and existing pipe infrastructure. Integrating handheld devices with external equipment such as orientation sensors or real-time kinematic technology will significantly enhance the positional accuracy of the system.

Chapter 5
General Discussion

Effectively wastewater and stormwater network management significantly impact critical human lives, for example, the environment, water use or discharge, food production, cleaning, irrigation, wastewater treatment, and energy balance (Robles et al., 2014). Sewer condition assessment models are useful tools to assess the status of sewer pipelines based on the interlinked relationships between affecting factors and sewer condition. The analysis of the importance of factors on sewer condition will provide helpful information on model performance improvement, and support the rehabilitation, repair, and maintenance strategies (Atambo et al., 2022; Li et al., 2020; Nethra Betgeri et al., 2023; Nguyen et al., 2022).

The results in **Paper I, Paper II**, and **Paper III** also indicated that material and age are the most significant factors affecting the sewer conditions in the study area. More specifically, polyvinyl chloride (PVC) pipes were more vulnerable to the deterioration process, followed by polypropylene and concrete materials in **Paper II**. Laakso et al. (2018) concluded that reinforced concrete pipes are more significant than other materials in predicting sewer conditions. The influence of aging on the sewer process is inevitable, but local water managers can partly reduce the negative effect of this process by selecting consistent materials. However, the relative importance of the other factors is slightly different between the feature selection methods; this can be explained by the random feature selection of each method in splitting and combining subsets to optimize model performance (Kuhn & Johnson, 2013).

The results in **Paper I** show that the prediction performance of regression-based ML models significantly fluctuates. This can be attributed to the skewness of damage score data, which affects the predictive performance of the models due to the large variability (Bellon-Maurel et al., 2010; Krawczyk, 2016). To address this problem, sewer damage scores were aggregated into classes and the regression problem is converted into a classification problem (**Paper II**). It is therefore recommended that future studies consider the classification-based approach to ML models for sewer condition assessment.

The findings in **Paper I** and **Paper II** demonstrate that the constructed MLP models (including single and multiple hidden layers) have lower prediction capability than others in modelling sewer conditions. This conclusion agrees with previous studies in which the neural network-based model is not a perfect solution for predicting pipes-related problems such as failures or corrosion (Fan et al., 2022; Zounemat-Kermani et al., 2020). The reason for the low performance of MLP models may be due to the limited quantity of input data to train the neural network models (Hawari et al., 2020).

The results in **Paper III** show that the hybrid ML models have higher performance than the base classifier. This conclusion is in line with previous studies' findings (W. Chen et al., 2019; Miraki et al., 2019; Phong et al., 2021). Different sample strategies of each hybrid model can be explained for the higher performance of these models. More specifically, the integration of the bootstrap sample and Principal Component Analysis (PCA) methods in the Rotation Forest (RotF) ensemble (Kuncheva & Rodríguez, 2007) is better than the bootstrap aggregating and disjoint aggregating methods in the Bagging (Breiman, 1996) and Dagging (Ting & Witten, 1997) ensembles, respectively. This statement is only concluded for the dataset in the study area.

When analyzing **Paper III**, we realized that resampling approaches are likely inefficient for dealing with imbalanced datasets while developing ML prediction models in the study area. This conclusion agrees with previous studies. For instance, Fan et al. (2022) found that the oversampling method does not increase the prediction capability of ML models and significantly increases computational time. Therefore, merging minor classes in the dataset should be considered before building the sewer condition prediction ML models. Moreover, more inspections and other factors should be considered to improve the prediction performance of the condition models.

Due to a lack of inspection data, operational factors were not considered in this thesis. Therefore, the influence of these elements on sewer conditions in the study area was not assessed specifically. From the literature review, the role of operational factors in sewer deterioration was a controversial problem. For example, Malek Mohammadi (2019) showed that flow was an insignificant factor, while other studies indicated that this factor significantly effects on sewer condition process (Koo & Ariaratnam, 2006; Laakso et al., 2018; Mohammadi et al., 2020). Moreover, the role of these factors in the assessment of the sewer condition process was still limited and needed to be investigated in further research (Atambo et al., 2022).

Based on the developed ML model in this thesis, the maps of the sewer pipes conditions for the years 2022, 2042, and 2072 in Ålesund city were created. Some assumptions were made, such as future population density having a linear trend or changing groundwater only depending on the change in rainfall. These maps can be used as reference materials or documents in developing future maintenance strategies in this study area.

It is easy to re-implement and re-distribute the ML models in this thesis because they are constructed and run using all open-source packages and software. Moreover, the input datasets used for model implementation are structured in table-based formats. Observations/attributes are easily inserted or eliminated using the popular spreadsheet (Microsoft Excel) or text editor program (TextEdit or Notepad).

Paper IV stressed the application possibility of handheld devices such as mobile and HoloLens in the water domain. These devices can be continuously used in other stages of the water system control and operation such as product design, maintenance, and staff training (Park et al., 2021). The users' awareness of water infrastructure and the surrounding environment is enhanced in a way that allows the exploration of the relations between them. Minimizing the delay in field data acquisition is one of the advantages of using an AR-based application; users can quickly and easily view the objects as well as their attributes. Compared to the traditional method, this will significantly reduce time wasted in the field without requiring manually checking real-world objects with corresponding ones from the record (an easily confused process).

Due to the GPS error on smartphones and environmental objects (Fenais et al., 2019), the positioning accuracy might not meet the desired expectations. Therefore, integrating handheld devices with external equipment such as orientation sensors or real-time kinematic (RTK) receivers should be considered to improve positional accuracy (Schall et al., 2013). Moreover, differences in positioning accuracy on different mobile devices will be significant because of dissimilar hardware structures and positioning sensors (Blum et al., 2013; Li et al., 2023).

Using this application in fieldwork requires the users to pay attention to some things during the operation. For example, the function *"GPS Location"* for pinpointing geographical locations requires mobile devices to continuously receive a GPS signal that consumes a lot of power, and their batteries are quickly drained. With the long period of investigations, the users need to prepare more batteries or external powers.

It is worth noting that in the dynamic visualization for the water flow in pipes, other dynamic input data (for example, temperature or contamination degree) can be simulated in the same way if they have the same data structure.

The experimental functions in this application were built and conducted based on the real sewer network in a specific study area in Ålesund city. The steps will be executed in the same process as other input databases for other areas. This paper presented the fundamental steps for sewer network 3D visualization on handheld devices such as smartphones and HoloLens, and a circle interaction between this application and a computer or server to transfer and process data for sewer management purposes was introduced. The technical and non-technical operators can consult this application as an assistant tool for collecting and visualizing data in the water sector.

Chapter 6
Conclusion and Recommendation

6.1. Main conclusions

Geospatial analysis techniques, machine learning algorithms, and state-of-the-art visualization applications supporting predictive maintenance purposes in selected wastewater and stormwater systems have been utilized and assessed in this thesis. Consequently, a GIS-based database has been established from many spatial and non-spatial data sources and used for the sewer network condition assessment of Ålesund city.

Various physical and environmental factors affecting the sewer condition at Ålesund city were investigated and their importance to the sewer condition was assessed by applying the different filters, wrappers, and embedded methods. The analysis results show that material and age are the most important factors while network type is the least factor affecting the sewer condition in the study area.

The results of this thesis identified that the ML-based binary classification models are generally outperformed and more stable than regression models in predicting sewer conditions in the study area. Hybrid ML models worked better than classification models even with imbalanced datasets and multiclass problems. The Extra Tree Regression is an effective algorithm for the regression problem whilst the Random Forest and a combination of the Rotation Forest technique and J48 Decision Tree are efficient algorithms for binary classification and multiclass hybrid models, respectively.

For the classification problem in this thesis, resampling methods (including under-sampling, over-sampling, and Synthetic Minority Oversampling Technique (SMOTE)) did not work well in dealing with an imbalanced dataset in the study area. In general, the predictive accuracy of almost ML models was not improved significantly using resampling methods. Especially, it took longer in the computational time when over-sampling methods were applied.

A platform for 3D visualization applying Augmented Reality techniques was developed in this thesis to visualize predictive obtained from ML models on handheld devices (for example, Android mobile and Microsoft HoloLens).

The outputs of this thesis can contribute to modify the theoretical basis and overall implications in term of sewer condition assessment. The main contributions of this thesis are as follows:

- The thesis redefines how to structure and manage data for sewer condition process. Instead of storing data of sewer networks, the utility engineers/managers need to collect only properties that are used to build ML model. Moreover, the process for sewer condition assessment introduced in the thesis can help them reduce their workload. For example, they do not need to undertake full inspections, prepare assessment reports from videos, and then classify sewer pipe status based on current standards or personal experiences. With appropriate data from history, status of sewer pipes in entire network can be easily estimated.
- The proposed comprehensive sewer asset database and related ML models developed can help managers to build their own database and ML-based sewer condition assessment models. After that, they can estimate the status of sewer pipes in real-time or near real-time irrespective of the study area. When status of a sewer pipe needs to be estimated, its condition can be quickly predicted using the proposed ML models from the physical and environmental characteristics. By utilizing the integrated GIS framework, outputs received from condition assessment models can be easily visualized and updated continuously.
- With the developed AR based visualization platform, water engineers and managers can quickly monitor sewer's status in real time on the fields instead of direct inspections such as visual inspection or sound-based methods that are normally time consuming and require highly professional qualifications (e.g., ability to use specialized equipment or having depth understanding on sewer condition).

6.2. Limitations and recommendations for further work

This thesis has mainly focused on sewer condition assessment supporting the predictive maintenance of the water collection network using ML models. The results of this thesis show the potential capabilities of data-driven methods in the water domain. Some limitations and issues can be recommended as subjects for further investigation are listed as follows:

- This thesis only accounted for some physical and environmental factors for sewer condition assessment. Operational factors (for example, flow rate, blockages, infiltration, and inflow) were not considered due to their unavailability at the time this thesis was constructed.

- Different factors (including physical, environmental, and operational factors) and inspection data should be considered in the future to increase the prediction capability of sewer condition assessment models.
- This thesis demonstrated the use of ML models in assessing sewer status and mapping future sewer conditions. However, these predictions were deployed based on the assumption that rainfall, groundwater, and population factors linearly change and other environmental factors are unchanged. Due to projected changes in precipitation in the Møre og Romsdal county of Norway, rainfall and relevant factors (such as discharge or flow rate) in the sewer network could shift in the future. To improve prediction capability for future scenarios, the dynamic characteristics of the aforementioned factors should be considered. Climate projection scenarios should be involved in calculating these abovementioned factors.
- The accuracy of GPS-based functions in the visualization application developed in this thesis is still limited due to the used devices' hardware limitations itself. Integration of handheld devices with location-supported devices (such as GPS receivers) to improve location accuracy can be considered in the future.

The developed applications on mobile and HoloLens devices should be considered in the future to integrate more needed functions based on the visualization platform developed in this thesis. In addition, model optimization (in terms of capacity and detailed level) for running on hardware-limited devices (for example, mobile or HoloLens) needs to be improved for further investigations.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., & Isard, M. (2016). Tensorflow: a system for large-scale machine learning. Osdi, https://doi.org/10.48550/arXiv.1605.08695

Ahmad, T., Chen, H., Huang, R., Yabin, G., Wang, J., Shair, J., Azeem Akram, H. M., Hassnain Mohsan, S. A., & Kazim, M. (2018). Supervised based machine learning models for short, medium and long-term energy prediction in distinct building environment. *Energy*, *158*, 17-32. https://doi.org/10.1016/j.energy.2018.05.169

Ahmadi, M., Cherqui, F., de Massiac, J.-C., Werey, C., Lagoutte, S., & Le Gauffre, P. (2014). Condition grading for dysfunction indicators in sewer asset management. *Structure and Infrastructure Engineering*, *10*(3), 346-358. https://doi.org/10.1080/15732479.2012.756916

Ajaj, Q. M., Shareef, M. A., Hassan, N. D., Hasan, S. F., & Noori, A. M. (2018). GIS Based Spatial Modeling to Mapping and Estimation Relative Risk of Different Diseases Using Inverse Distance Weighting (IDW) Interpolation Algorithm and Evidential Belief Function (EBF) (Case study: Minor Part of Kirkuk City, Iraq). *Int J Eng Technol*, *7*(4.37), 185-191. https://doi.org/10.14419/ijet.v7i4.37.24098

Aljamaan, H., & Alazba, A. (2020). *Software defect prediction using tree-based ensembles* Proceedings of the 16th ACM International Conference on Predictive Models and Data Analytics in Software Engineering, Virtual, USA. https://doi.org/10.1145/3416508.3417114.

Alonso, J. M., & Bugarín, A. (2019, 23-26 June 2019). ExpliClas: Automatic Generation of Explanations in Natural Language for Weka Classifiers. 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), https://doi.org/10.1109/FUZZ-IEEE.2019.8859018

Alsaqqar, A. S., Khudair, B. H., & Jbbar, R. K. (2017). Rigid Trunk Sewer Deterioration Prediction Models using Multiple Discriminant and Neural Network Models in Baghdad City, Iraq. *Journal of Engineering*, *23*(8), 70-83. https://doi.org/10.31026/j.eng.2017.08.06

Altarabsheh, A., Ventresca, M., & Kandil, A. (2018). New Approach for Critical Pipe Prioritization in Wastewater Asset Management Planning. *Journal of Computing in Civil Engineering*, *32*(5), 18. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000784

Ana, E., & Bauwens, W. (2007). Sewer network asset management decision-support tools: a review. International Symposium on New Directions in Urban Water Management, https://www.semanticscholar.org/paper/SEWER-NETWORK-ASSET-MANAGEMENT-DECISION-SUPPORT-A-Ana-Bauwens/4b78b611297ea17456095bb4c855bcb4345fb4b9

Ana, E., Bauwens, W., Pessemier, M., Thoeye, C., Smolders, S., Boonen, I., & De Gueldre, G. (2009). An investigation of the factors influencing sewer structural deterioration. *Urban Water Journal*, *6*(4), 303-312. https://doi.org/10.1080/15730620902810902

Ana, E. V., & Bauwens, W. (2010). Modeling the structural deterioration of urban drainage pipes: the state-of-the-art in statistical methods. *Urban Water Journal*, *7*(1), 47-59.

https://doi.org/10.1080/15730620903447597

Anand, U., Li, X., Sunita, K., Lokhandwala, S., Gautam, P., Suresh, S., Sarma, H., Vellingiri, B., Dey, A., Bontempi, E., & Jiang, G. (2022). SARS-CoV-2 and other pathogens in municipal wastewater, landfill leachate, and solid waste: A review about virus surveillance, infectivity, and inactivation. *Environmental Research*, *203*, 111839. https://doi.org/10.1016/j.envres.2021.111839

Askari, A., d'Aspremont, A., & Ghaoui, L. E. (2020). *Naive Feature Selection: Sparsity in Naive Bayes* Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, Proceedings of Machine Learning Research. https://proceedings.mlr.press/v108/askari20a.html.

Atambo, D. O., Najafi, M., & Kaushal, V. (2022). Development and Comparison of Prediction Models for Sanitary Sewer Pipes Condition Assessment Using Multinomial Logistic Regression and Artificial Neural Network. *Sustainability*, *14*(9), 5549. https://doi.org/10.3390/su14095549

Ayyadevara, V. K. (2018). Gradient Boosting Machine. In *Pro Machine Learning Algorithms: A Hands-On Approach to Implementing Algorithms in Python and R* (pp. 117-134). Apress. https://doi.org/10.1007/978-1-4842-3564-5_6

Bakry, I., Alzraiee, H., Kaddoura, K., Masry, M. E., & Zayed, T. (2016). Condition Prediction for Chemical Grouting Rehabilitation of Sewer Networks. *Journal of Performance of Constructed Facilities*, *30*(6), 11. https://doi.org/10.1061/(ASCE)CF.1943-5509.000089

Balekelayi, N., & Tesfamariam, S. (2019). Statistical inference of sewer pipe deterioration using Bayesian geoadditive regression model. *Journal of Infrastructure Systems*, *25*(3), 14. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000500

Beha, F., Göritz, A., & Schildhauer, T. (2015). Business model innovation: The role of different types of visualizations. ISPIM Conference Proceedings, Germany. https://www.hiig.de/wp-content/uploads/2015/06/671925834_Paper.pdf

Behzadan, A. H., Dong, S., & Kamat, V. R. (2015). Augmented reality visualization: A review of civil infrastructure system applications. *Advanced Engineering Informatics*, *29*(2), 252-267. https://doi.org/10.1016/j.aei.2015.03.005

Bellon-Maurel, V., Fernandez-Ahumada, E., Palagos, B., Roger, J.-M., & McBratney, A. (2010). Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *TrAC Trends in Analytical Chemistry*, *29*(9), 1073-1081. https://doi.org/10.1016/j.trac.2010.05.006

Bentéjac, C., Csörgő, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, *54*(3), 1937-1967. https://doi.org/10.1007/s10462-020-09896-5

Blum, J. R., Greencorn, D. G., & Cooperstock, J. R. (2013). Smartphone Sensor Reliability for Augmented Reality Applications. In K. Zheng, M. Li, & H. Jiang, *Mobile and Ubiquitous Systems: Computing, Networking, and Services* Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-40238-8_11

Bottani, E., & Vignali, G. (2019). Augmented reality technology in the manufacturing industry: A review of the last decade. *IISE Transactions*, *51*(3), 284-310. https://doi.org/10.1080/24725854.2018.1493244

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random Forests. *Machine learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and Regression Trees* (1st Edition ed.). Routledge. https://doi.org/10.1201/9781315139470

Bui, K.-T. T., Torres, J. F., Gutiérrez-Avilés, D., Nhu, V.-H., Bui, D. T., & Martínez-Álvarez, F. (2022). Deformation forecasting of a hydropower dam by hybridizing a long short-term memory deep learning network with the coronavirus optimization algorithm. *Computer-Aided Civil and Infrastructure Engineering*. https://doi.org/10.1111/mice.12810

Burner, D. M., Ashworth, A. J., Laughlin, K. F., & Boyer, M. E. (2018). Using SketchUp to Simulate Tree Row Azimuth Effects on Alley Shading. *Agronomy Journal*, *110*(1), 425-430. https://doi.org/10.2134/agronj2017.04.0224

Caradot, N., Riechel, M., Fesneau, M., Hernandez, N., Torres, A., Sonnenberg, H., Eckert, E., Lengemann, N., Waschnewski, J., & Rouault, P. (2018). Practical benchmarking of statistical and machine learning models for predicting the condition of sewer pipes in Berlin, Germany. *Journal of Hydroinformatics*, *20*(5), 1131-1147. https://doi.org/10.2166/HYDRO.2018.217

Caradot, N., Riechel, M., Rouault, P., Caradot, A., Lengemann, N., Eckert, E., Ringe, A., Clemens, F., & Cherqui, F. (2020). The influence of condition assessment uncertainties on sewer deterioration modelling. *Structure and Infrastructure Engineering*, *16*(2), 287-296. https://doi.org/10.1080/15732479.2019.1653938

Caradot, N., Sonnenberg, H., Kropp, I., Ringe, A., Denhez, S., Hartmann, A., & Rouault, P. (2017). The relevance of sewer deterioration modelling to support asset management strategies. *http://dx.doi.org/10.1080/1573062X.2017.1325497*, *14*(10), 1007-1015. https://doi.org/10.1080/1573062X.2017.1325497

Carneiro, J., Rossetti, R. J. F., Silva, D. C., & Oliveira, E. C. (2018, 16-19 Sept. 2018). BIM, GIS, IoT, and AR/VR Integration for Smart Maintenance and Management of Road Networks: a Review. 2018 IEEE International Smart Cities Conference (ISC2), https://doi.org/10.1109/ISC2.2018.8656978

Cataldi, L., Tiberi, L., & Costa, G. (2021). Estimation of MCS intensity for Italy from high quality accelerometric data, using GMICEs and Gaussian Naïve Bayes Classifiers. *Bulletin of Earthquake Engineering*, *19*(6), 2325-2342. https://doi.org/10.1007/s10518-021-01064-6

Centeno, J. A. S., Kishi, R. T., & Mitishita, E. A. (2009, July 21-24, 2009). Three-dimensional Data Visualization in Water Quality Studies using Augmented Reality. 6th International Symposium on Mobile Mapping Technology, Brazil. http://www2.fct.unesp.br/docentes/carto/JoaoFernando/Artigos_MMT_2009/050_Centeno_MMT09.pdf

Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, *408*, 189-215. https://doi.org/10.1016/j.neucom.2019.10.118

Chan, J. Y., Leow, S. M., Bea, K. T., Cheng, W. K., Phoong, S. W., Hong, Z.-W., Lin, J.-M., & Chen, Y.-L. (2022). A Correlation-Embedded Attention Module to Mitigate Multicollinearity: An Algorithmic Trading Application. *Mathematics*, *10*(8). https://doi.org/10.3390/math10081231

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers &*

Electrical Engineering, *40*(1), 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chen, W., Hong, H., Li, S., Shahabi, H., Wang, Y., Wang, X., & Ahmad, B. B. (2019). Flood susceptibility modelling using novel hybrid approach of reduced-error pruning trees with bagging and random subspace ensembles. *Journal of Hydrology*, *575*, 864-873. https://doi.org/10.1016/j.jhydrol.2019.05.089

Chen, Y., Wang, Q., Chen, H., Song, X., Tang, H., & Tian, M. (2019). An overview of augmented reality technology. *Journal of Physics: Conference Series*, *1237*(2), 6. https://doi.org/10.1088/1742-6596/1237/2/022082

Cheng, J. C. P., Chen, W., Chen, K., & Wang, Q. (2020). Data-driven predictive maintenance planning framework for MEP components based on BIM and IoT using machine learning algorithms. *Automation in Construction*, *112*, 103087. https://doi.org/10.1016/j.autcon.2020.103087

Chughtai, F., & Zayed, T. (2007). Sewer Pipeline Operational Condition Prediction Using Multiple Regression. In *Pipelines 2007* (pp. 1-11). https://doi.org/10.1061/40934(252)18

Climate, D. (2021). *Ålesund Climate: Average Temperature, Weather by Month, Ålesund Water Temperature - Climate-Data.org*. Available online: https://en.climate-data.org/europe/norway/m%C3%B8re-og-romsdal/alesund-9937/ (accessed on April 20th, 2020)

Conlin, R., Erickson, K., Abbate, J., & Kolemen, E. (2021). Keras2c: A library for converting Keras neural networks to real-time compatible C. *Engineering Applications of Artificial Intelligence*, *100*, 104182. https://doi.org/10.1016/j.engappai.2021.104182

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Coscetti, S., Moroni, D., Pieri, G., & Tampucci, M. (2020). *Factory Maintenance Application Using Augmented Reality* Proceedings of the 3rd International Conference on Applications of Intelligent Systems, Las Palmas de Gran Canaria, Spain. https://doi.org/10.1145/3378184.3378218.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE transactions on information theory*, *13*(1), 21-27. https://doi.org/10.1109/TIT.1967.1053964

Dalzochio, J., Kunst, R., Pignaton, E., Binotto, A., Sanyal, S., Favilla, J., & Barbosa, J. (2020). Machine learning and reasoning for predictive maintenance in Industry 4.0: Current status and challenges. *Computers in Industry*, *123*, 103298. https://doi.org/10.1016/j.compind.2020.103298

Daniel, T. L., & Chantal, D. L. (2014). *Discovering Knowledge in Data: An Introduction to Data Mining* (Vol. 4). John Wiley & Sons. https://doi.org/10.1002/9781118874059

Dokeroglu, T., Deniz, A., & Kiziloz, H. E. (2021). A robust multiobjective Harris' Hawks Optimization algorithm for the binary classification problem. *Knowledge-Based Systems*, *227*, 107219. https://doi.org/10.1016/j.knosys.2021.107219

Doleac, M., Lackey, S., & Bratton, G. (1980). Prediction of time-to failure for buried cast iron pipe. Proceedings of American water works association annual conference, Denver, Colorado.

Dübel, S., Röhlig, M., Schumann, H., & Trapp, M. (2014, 9-9 Nov. 2014). 2D and 3D presentation of spatial data: A systematic review. 2014 IEEE VIS International Workshop on 3DVis (3DVis), https://doi.org/10.1109/3DVis.2014.7160094

Ebrahimy, H., Feizizadeh, B., Salmani, S., & Azadi, H. (2020). A comparative study of land subsidence susceptibility mapping of Tasuj plane, Iran, using boosted regression tree, random forest and classification and regression tree methods. *Environmental Earth Sciences*, *79*(10), 223. https://doi.org/10.1007/s12665-020-08953-0

Fan, X., Wang, X., Zhang, X., & Asce Xiong Yu, P. E. F. (2022). Machine learning based water pipe failure prediction: The effects of engineering, geology, climate and socio-economic factors. *Reliability Engineering & System Safety*, *219*, 108185. https://doi.org/10.1016/j.ress.2021.108185

Farkas, K., Hillary, L. S., Malham, S. K., McDonald, J. E., & Jones, D. L. (2020). Wastewater and public health: the potential of wastewater surveillance for monitoring COVID-19. *Current Opinion in Environmental Science & Health*, *17*, 14-20. https://doi.org/10.1016/j.coesh.2020.06.001

Fenais, A., Ariaratnam, S. T., Ayer, S. K., & Smilovsky, N. (2019). Integrating Geographic Information Systems and Augmented Reality for Mapping Underground Utilities. *Infrastructures*, *4*(4), 60. https://doi.org/10.3390/infrastructures4040060

Fite-Georgel, P. (2011, 26-29 Oct. 2011). Is there a reality in Industrial Augmented Reality? 2011 10th IEEE International Symposium on Mixed and Augmented Reality, https://doi.org/10.1109/ISMAR.2011.6092387

Frank, E., Trigg, L., Holmes, G., & Witten, I. H. (2000). Technical Note: Naive Bayes for Regression. *Machine learning*, *41*(1), 5-25. https://doi.org/10.1023/A:1007670802811

Freund, Y., & Schapire, R. E. (1997). A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of computer and system sciences*, *55*(1), 119-139. https://doi.org/10.1006/jcss.1997.1504

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367-378. https://doi.org/10.1016/S0167-9473(01)00065-2

Geurts, P., Ernst, D., & Wehenkel, L. (2006). Extremely randomized trees. *Machine learning*, *63*(1), 3-42. https://doi.org/10.1007/s10994-006-6226-1

Gruber, M. H. (1998). *Improving Efficiency by Shrinkage: The James--Stein and Ridge Regression Estimators* (1st Edition ed.). Routledge. https://doi.org/10.1201/9780203751220

Guryanov, A. (2019, 2019). Histogram-Based Algorithm for Building Gradient Boosting Ensembles of Piecewise Linear Decision Trees. Analysis of Images, Social Networks and Texts, Cham. https://doi.org/10.1007/978-3-030-37334-4_4

Hadzilacos, T., Kalles, D., Preston, N., Melbourne, P., Camarinopoulos, L., Eimermacher, M., Kallidromitis, V., Frondistou-Yannas, S., & Saegrov, S. (2000). UtilNets: a water mains rehabilitation decision-support system. *Computers, Environment and Urban Systems*, *24*(3), 215-232. https://doi.org/10.1016/S0198-9715(99)00058-7

Hamacher, A., Kim, S. J., Cho, S. T., Pardeshi, S., Lee, S. H., Eun, S.-J., & Whangbo, T. K. (2016). Application of Virtual, Augmented, and Mixed Reality to Urology. *International neurourology journal*, *20*(3), 172-181. https://doi.org/10.5213/inj.1632714.357

Harvey, R. R., & McBean, E. A. (2014). Predicting the structural condition of individual sanitary sewer pipes with random forests. *Canadian Journal of Civil Engineering*, *41*(4), 294-303. https://doi.org/10.1139/cjce-2013-0431

Hassan, S. I., Dang, L. M., Mehmood, I., Im, S., Choi, C., Kang, J., Park, Y.-S., & Moon, H.

(2019). Underground sewer pipe condition assessment based on convolutional neural networks. *Automation in Construction*, *106*, 102849. https://doi.org/10.1016/j.autcon.2019.102849

Haugen, H. J., & Viak, A. (2018). *Datafl yt – Klassifi sering av avløpsledninger*. https://docplayer.me/211256711-Norsk-vann-rapport-dataflyt-klassifisering-av-avlopsledninger.html

Hawari, A., Alamin, M., Alkadour, F., Elmasry, M., & Zayed, T. (2018). Automated defect detection tool for closed circuit television (cctv) inspected sewer pipelines. *Automation in Construction*, *89*, 99-109. https://doi.org/10.1016/j.autcon.2018.01.004

Hawari, A., Alkadour, F., Elmasry, M., & Zayed, T. (2020). A state of the art review on condition assessment models developed for sewer pipelines. *Engineering Applications of Artificial Intelligence*, *93*, 103721. https://doi.org/10.1016/j.engappai.2020.103721

Hawari, A., Firas, A., Elmasry, Mohamed, & Zayed, T. (2016). Simulation-Based Condition Assessment Model for Sewer Pipelines. *Journal of Performance of Constructed Facilities*, *31*(1), 15. https://doi.org/10.1061/(ASCE)CF.1943-5509.0000914

Haynes, P., Hehl-Lange, S., & Lange, E. (2018). Mobile Augmented Reality for Flood Visualisation. *Environmental Modelling & Software*, *109*, 380-389. https://doi.org/10.1016/j.envsoft.2018.05.012

He, J., Ding, L., Jiang, L., & Ma, L. (2014, 6-11 July 2014). Kernel ridge regression classification. 2014 International Joint Conference on Neural Networks (IJCNN), https://doi.org/10.1109/IJCNN.2014.6889396

Hernández, N., Caradot, N., Sonnenberg, H., Rouault, P., & Torres, A. (2021). Optimizing SVM models as predicting tools for sewer pipes conditions in the two main cities in Colombia for different sewer asset management purposes. *Structure and Infrastructure Engineering*, *17*(2), 156-169. https://doi.org/10.1080/15732479.2020.1733029

Heydarzadeh, R., Tabesh, M., & Scholz, M. (2021). Dissolved oxygen determination in sewers using flow hydraulic parameters as part of a physical-biological simulation model. *Journal of Hydroinformatics*, *24*(1), 1-15. https://doi.org/10.2166/hydro.2021.051

Hocking, J., & Schell, J. (2022). *Unity in action : multiplatform game development in C#* (Third edition ed.). Manning Publications Co., Shelter Island. https://manning-content.s3.amazonaws.com/download/c/1efc2a6-266f-4b10-a25b-f39b37d605ea/SampleChapter-01.pdf

Hoerl, A. E., & Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, *12*(1), 55-67. https://doi.org/10.1080/00401706.1970.10488634

Hong, H., Liu, J., Bui, D. T., Pradhan, B., Acharya, T. D., Pham, B. T., Zhu, A. X., Chen, W., & Ahmad, B. B. (2018). Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China). *CATENA*, *163*, 399-413. https://doi.org/10.1016/j.catena.2018.01.005

Jahromi, A. H., & Taheri, M. (2017, 25-27 Oct. 2017). A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features. 2017 Artificial Intelligence and Signal Processing Conference (AISP), https://doi.org/10.1109/AISP.2017.8324083

Juliani, A., Berges, V.-P., Teng, E., Cohen, A., Harper, J., Elion, C., Goy, C., Gao, Y., Henry, H., & Mattar, M. (2018). Unity: A General Platform for Intelligent Agents. *arXiv Artificial Intelligence*. https://doi.org/10.48550/arXiv.1809.02627

Jusuf, S. K., Ignatius, M., Wong, N. H., & Tan, E. (2017). STEVE Tool Plug-in for SketchUp: A User-Friendly Microclimatic Mapping Tool for Estate Development. In T. H. Karyono, R. Vale, & B. Vale (Eds.), *Sustainable Building and Built Environments to Mitigate Climate Change in the Tropics: Conceptual and Practical Approaches* (pp. 113-130). Springer International Publishing. https://doi.org/10.1007/978-3-319-49601-6_9

Kabir, G., Balekelayi, N., Celestin Balek, B., & Tesfamariam, S. (2018). Sewer Structural Condition Prediction Integrating Bayesian Model Averaging with Logistic Regression. *Journal of Performance of Constructed Facilities*, *32*(3), 04018019-04018019. https://doi.org/10.1061/(ASCE)CF.1943-5509.0001162

Kaddioui, A., Shahrour, I., & El Oirrak, A. (2019). Uses of Augmented reality for urban utilities management. *MATEC Web Conf.*, *295*, 02009. https://doi.org/10.1051/matecconf/201929502009

Kamel, B. M. N., Lu, Z., Guerrero, P., Jennett, C., & Steed, A. (2017). From urban planning and emergency training to Pokémon Go: applications of virtual reality GIS (VRGIS) and augmented reality GIS (ARGIS) in personal, public and environmental health. *International Journal of Health Geographics*, *16*(1), 7. https://doi.org/10.1186/s12942-017-0081-0

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). *Lightgbm: A highly efficient gradient boosting decision tree* NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, https://hal.science/hal-03953007.

Kégl, B. (2013). The return of AdaBoost. MH: multi-class Hamming trees. *arXiv preprint arXiv:1312.6086*. https://doi.org/10.48550/arXiv.1312.6086

Ketkar, N. (2017). Introduction to Keras. In *Deep Learning with Python: A Hands-on Introduction* (pp. 97-111). Apress. https://doi.org/10.1007/978-1-4842-2766-4_7

Khan, Z., Zayed, T., & Moselhi, O. (2010). Structural Condition Assessment of Sewer Pipelines. *Journal of Performance of Constructed Facilities*, *24*(2), 170-179. https://doi.org/10.1061/(ASCE)CF.1943-5509.0000081

Khayyam, H., Naebe, M., Zabihi, O., Zamani, R., Atkiss, S., & Fox, B. (2015). Dynamic Prediction Models and Optimization of Polyacrylonitrile (PAN) Stabilization Processes for Production of Carbon Fiber. *IEEE Transactions on Industrial Informatics*, *11*(4), 887-896. https://doi.org/10.1109/TII.2015.2434329

Khazraeializadeh, S. (2012). *A Comparative Analysis on Sewer Structural Condition Grading Systems Using Four Sewer Condition Assessment Protocols* [Master of Science, University of Alberta]. Edmonton, Alberta. https://era.library.ualberta.ca/items/05fdad1c-b34e-4b32-a933-27542ccf8f88/view/a306a266-ac96-42ba-895f-08894d9ad71a/Soroush_Khazraeializadeh_Thesis.pdf

Kleiner, Y., & Rajani, B. (2001). Comprehensive review of structural deterioration of water mains: statistical models. *Urban Water*, *3*(3), 131-150. https://doi.org/10.1016/S1462-0758(01)00033-4

Kleiner, Y., & Rajani, B. (2022). Economics of Inspection and Condition Assessment of High-Consequence Water Pipeline and Assessing Its Remaining Life: Theoretical Framework. *Journal of Pipeline Systems Engineering and Practice*, *13*(4), 13. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000681

Kley, G., & Caradot, N. (2013). *Review of sewer deterioration models* (KWB project SEMA, Report, Issue. https://publications.kompetenz-wasser.de/en/publication/663/

Koo, D.-H., & Ariaratnam, S. T. (2006). Innovative method for assessment of underground sewer pipe condition. *Automation in Construction*, *15*(4), 479-488. https://doi.org/10.1016/j.autcon.2005.06.007

Kostoláni, M., Murín, J., & Š, K. (2019, 11-14 June 2019). Intelligent predictive maintenance control using augmented reality. 2019 22nd International Conference on Process Control (PC19), https://doi.org/10.1109/PC.2019.8815042

Koutitas, G., Smith, K. S., Lawrence, G., Metsis, V., Stamper, C., Trahan, M., & Lehr, T. (2019). *A virtual and augmented reality platform for the training of first responders of the ambulance bus* Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments, Rhodes, Greece. https://doi.org/10.1145/3316782.3321542.

Kovacs, D. J., Li, Z., Baetz, B. W., Hong, Y., Donnaz, S., Zhao, X., Zhou, P., Ding, H., & Dong, Q. (2022). Membrane fouling prediction and uncertainty analysis using machine learning: A wastewater treatment plant case study. *Journal of Membrane Science*, *660*, 120817. https://doi.org/10.1016/j.memsci.2022.120817

Krawczyk, B. (2016). Learning from imbalanced data: Open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221-232. https://doi.org/10.1007/s13748-016-0094-0

Książek, W., Gandor, M., & Pławiak, P. (2021). Comparison of various approaches to combine logistic regression with genetic algorithms in survival prediction of hepatocellular carcinoma. *Computers in Biology and Medicine*, *134*, 104431. https://doi.org/10.1016/j.compbiomed.2021.104431

Kuhn, M., & Johnson, K. (2013). An Introduction to Feature Selection. In *Applied Predictive Modeling* (pp. 487-519). Springer New York. https://doi.org/10.1007/978-1-4614-6849-3_19

Kuk, A. Y. C., & Chen, C.-H. (1992). A mixture model combining logistic regression with proportional hazards regression. *Biometrika*, *79*(3), 531-541. https://doi.org/10.1093/biomet/79.3.531

Kumar, S. S., & Shaikh, T. (2017, 6-7 Sept. 2017). Empirical Evaluation of the Performance of Feature Selection Approaches on Random Forest. 2017 International Conference on Computer and Applications (ICCA), https://doi.org/10.1109/COMAPP.2017.8079769

Kuncheva, L. I., & Rodríguez, J. J. (2007). *An Experimental Study on Rotation Forest Ensembles* International workshop on multiple classifier systems, https://doi.org/10.1007/978-3-540-72523-7_46.

Kuss, M., Rasmussen, C. E., & Herbrich, R. (2005). Assessing Approximate Inference for Binary Gaussian Process Classification. *Journal of machine learning research*, *6*(10). https://www.jmlr.org/papers/volume6/kuss05a/kuss05a.pdf

Kvitsjøen, J., Paus, K. H., Bjerkholt, J. T., Fergus, T., & Lindholm, O. (2021). Intensifying rehabilitation of combined sewer systems using trenchless technology in combination with low impact development and green infrastructure. *Water Science and Technology*, *83*(12), 2947-2962. https://doi.org/10.2166/wst.2021.198

Laakso, T., Kokkonen, T., Mellin, I., & Vahala, R. (2018). Sewer Condition Prediction and Analysis of Explanatory Factors. *Water 2018, Vol. 10, Page 1239*, *10*(9), 1239-1239. https://doi.org/10.3390/W10091239

Laakso, T., Kokkonen, T., Mellin, I., & Vahala, R. (2019). Sewer Life Span Prediction: Comparison of Methods and Assessment of the Sample Impact on the Results. *Water*, *11*(12), 2657. https://doi.org/10.3390/w11122657

Lall, U., & Sharma, A. (1996). A Nearest Neighbor Bootstrap For Resampling Hydrologic Time Series. *Water Resources Research*, *32*(3), 679-693. https://doi.org/10.1029/95WR02966

Lee, S.-H., Park, S., & Kim, J. M. (2015). Suggestion for a Framework for a Sustainable Infrastructure Asset Management Manual in Korea. *Sustainability*, *7*(11), 15003-15028. https://doi.org/10.3390/su71115003

Lee, T., Shin, J.-Y., Kim, J.-S., & Singh, V. P. (2020). Stochastic simulation on reproducing long-term memory of hydroclimatological variables using deep learning model. *Journal of Hydrology*, *582*, 124540. https://doi.org/10.1016/j.jhydrol.2019.124540

Lewis, G. M., & Hampton, S. J. (2015). Visualizing volcanic processes in SketchUp: An integrated geo-education tool. *Computers & Geosciences*, *81*, 93-100. https://doi.org/10.1016/j.cageo.2015.05.003

Li, M., Feng, X., Han, Y., & Liu, X. (2023). Mobile augmented reality-based visualization framework for lifecycle O&M support of urban underground pipe networks. *Tunnelling and Underground Space Technology*, *136*, 21. https://doi.org/10.1016/j.tust.2023.105069

Li, W., Ling, W., Liu, S., Zhao, J., Liu, R., Chen, Q., Qiang, Z., & Qu, J. (2011). Development of systems for detection, early warning, and control of pipeline leakage in drinking water distribution: A case study. *Journal of Environmental Sciences*, *23*(11), 1816-1822. https://doi.org/10.1016/S1001-0742(10)60577-3

Li, X., Chen, W., Zhang, Q., & Wu, L. (2020). Building Auto-Encoder Intrusion Detection System based on random forest feature selection. *Computers & Security*, *95*, 101851. https://doi.org/10.1016/j.cose.2020.101851

Li, Y., Liu, K., Foley, A. M., Zülke, A., Berecibar, M., Nanini-Maury, E., Van Mierlo, J., & Hoster, H. E. (2019). Data-driven health estimation and lifetime prediction of lithium-ion batteries: A review. *Renewable and Sustainable Energy Reviews*, *113*, 109254. https://doi.org/10.1016/j.rser.2019.109254

Lim, T.-S., Loh, W.-Y., & Shih, Y.-S. (2000). A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms. *Machine learning*, *40*(3), 203-228. https://doi.org/10.1023/A:1007608224229

Liu, Z., & Kleiner, Y. (2013). State of the art review of inspection technologies for condition assessment of water pipes. *Measurement*, *46*(1), 1-15. https://doi.org/10.1016/j.measurement.2012.05.032

Machado, M. R., & Karray, S. (2022). Assessing credit risk of commercial customers using hybrid machine learning algorithms. *Expert Systems with Applications*, *200*, 116889. https://doi.org/10.1016/j.eswa.2022.116889

Malek Mohammadi, M. (2019). *Development of condition prediction models for sanitary sewer pipes* The University of Texas at Arlington]. http://hdl.handle.net/10106/28665

Malik, H., Bashir, U., & Ahmad, A. (2022). Multi-classification neural network model for detection of abnormal heartbeat audio signals. *Biomedical Engineering Advances*, *4*, 100048. https://doi.org/10.1016/j.bea.2022.100048

Masrur Ahmed, A. A., Deo, R. C., Feng, Q., Ghahramani, A., Raj, N., Yin, Z., & Yang, L. (2021). Deep learning hybrid model with Boruta-Random forest optimiser algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity. *Journal of Hydrology*, *599*, 126350. https://doi.org/10.1016/j.jhydrol.2021.126350

Meng, L., & Zhang, J. (2020). Process Design of Laser Powder Bed Fusion of Stainless Steel Using a Gaussian Process-Based Machine Learning Model. *JOM*, *72*(1), 420-428. https://doi.org/10.1007/s11837-019-03792-2

Microsoft. (2021). *HoloLens (1st gen) hardware*. Available online: https://learn.microsoft.com/en-us/hololens/hololens1-hardware (accessed on 01 November 2022, 2022)

Miraki, S., Zanganeh, S. H., Chapi, K., Singh, V. P., Shirzadi, A., Shahabi, H., & Pham, B. T. (2019). Mapping Groundwater Potential Using a Novel Hybrid Intelligence Approach. *Water Resources Management*, *33*(1), 281-302. https://doi.org/10.1007/s11269-018-2102-6

Mirauda, D., Erra, U., Agatiello, R., & Cerverizzo, M. (2017). Applications of Mobile Augmented Reality to Water Resources Management. *Water*, *9*(9). https://doi.org/10.3390/w9090699

Mohammadi, M. M., Najafi, M., Kermanshachi, S., Kaushal, V., & Serajiantehrani, R. (2020). Factors Influencing the Condition of Sewer Pipes: State-of-the-Art Review. *Journal of Pipeline Systems Engineering and Practice*, *11*(4), 03120002. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000483

Najafi, M., Gokhale, S., Calderón, D. R., & Ma, B. (2021). *Trenchless technology: Pipeline and utility design, construction, and renewal*. McGraw-Hill Education. https://www.accessengineeringlibrary.com/content/book/9780071422666

Najafi, M., & Kulandaivel, G. (2005). Pipeline Condition Prediction Using Neural Network Models. In *Pipelines 2005* (pp. 767-781). https://doi.org/10.1061/40800(180)61

Nethra Betgeri, S., Reddy Vadyala, S., Matthews, J. C., Madadi, M., & Vladeanu, G. (2023). Wastewater pipe condition rating model using K- Nearest Neighbors. *Tunnelling and Underground Space Technology*, *132*, 104921. https://doi.org/10.1016/j.tust.2022.104921

Ngo, P.-T. T., Pham, T. D., Nhu, V.-H., Le, T. T., Tran, D. A., Phan, D. C., Hoa, P. V., Amaro-Mellado, J. L., & Bui, D. T. (2021). A novel hybrid quantum-PSO and credal decision tree ensemble for tropical cyclone induced flash flood susceptibility mapping with geospatial data. *Journal of Hydrology*, *596*, 125682. https://doi.org/10.1016/j.jhydrol.2020.125682

Nguyen, L. V., Bui, D. T., & Seidu, R. (2022). Comparison of Machine Learning Techniques for Condition Assessment of Sewer Network. *IEEE Access*, *10*, 124238-124258. https://doi.org/10.1109/ACCESS.2022.3222823

Nguyen, L. V., & Seidu, R. (2022). Application of Regression-Based Machine Learning Algorithms in Sewer Condition Assessment for Ålesund City, Norway. *Water*, *14*(24), 3993. https://www.mdpi.com/2073-4441/14/24/3993

Noshahri, H., olde Scholtenhuis, L. L., Doree, A. G., & Dertien, E. C. (2021). Linking sewer condition assessment methods to asset managers' data-needs. *Automation in Construction*, *131*, 103878. https://doi.org/10.1016/j.autcon.2021.103878

Park, S., Bokijonov, S., & Choi, Y. (2021). Review of Microsoft HoloLens Applications over the Past Five Years. *Applied Sciences*, *11*(16), 7259. https://doi.org/10.3390/app11167259

Park, T. (2009). *A comprehensive asset management system for sewer infrastructures*. The Pennsylvania State University.

https://www.proquest.com/openview/9a4a13d47f153dd80328ecd7adcd788f/1?pq-origsite=gscholar&cbl=18750

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830. https://scikit-learn.org/stable/

Pham, B. T., Avand, M., Janizadeh, S., Phong, T. V., Al-Ansari, N., Ho, L. S., Das, S., Le, H. V., Amini, A., Bozchaloei, S. K., Jafari, F., & Prakash, I. (2020). GIS Based Hybrid Computational Approaches for Flash Flood Susceptibility Assessment. *Water*, *12*(3), 683. https://doi.org/10.3390/w12030683

Phong, T. V., Pham, B. T., Trinh, P. T., Ly, H.-B., Vu, Q. H., Ho, L. S., Le, H. V., Phong, L. H., Avand, M., & Prakash, I. (2021). Groundwater Potential Mapping Using GIS-Based Hybrid Artificial Intelligence Methods. *Groundwater*, *59*(5), 745-760. https://doi.org/10.1111/gwat.13094

Rahman, & Vanier. (2004). *An evaluation of condition assessment protocols for sewer management* (Client Report (National Research Council of Canada. Institute for Research in Construction); no. B-5123.6, Issue. https://nrc-publications.canada.ca/eng/view/object/?id=93ed3e91-be5e-452d-b79d-c20d4ac77002

Rahmati, O., Avand, M., Yariyan, P., Tiefenbacher, J. P., Azareh, A., & Bui, D. T. (2022). Assessment of Gini-, entropy- and ratio-based classification trees for groundwater potential modelling and prediction. *Geocarto International*, *37*(12), 3397-3415. https://doi.org/10.1080/10106049.2020.1861664

Ramezankhani, M., Crawford, B., Khayyam, H., Naebe, M., Seethaler, R., & Milani, A. S. (2019). A multi-objective Gaussian process approach for optimization and prediction of carbonization process in carbon fiber production under uncertainty. *Advanced Composites and Hybrid Materials*, *2*, 444-455. https://doi.org/10.1007/s42114-019-00107-6

Randall-Smith, M., Oliphant, R., & Russell, A. (1992). *Guidance manual for the structural condition assessment of trunk mains*. Water Research Centre. https://books.google.no/books/about/Guidance_Manual_for_the_Structural_Condi.html?id=Xb0WAAAACAAJ&redir_esc=y

Raschka, S., & Mirjalili, V. (2019). *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2, 3rd Edition* (3rd Edition ed.). Packt Publishing Ltd. https://falksangdata.no/wp-content/uploads/2022/07/python-machine-learning-and-deep-learning-with-python-scikit-learn-and-tensorflow-2.pdf

Rasmussen, C. E. (2004). Gaussian Processes in Machine Learning. In O. Bousquet, U. von Luxburg, & G. Rätsch (Eds.), *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures* (pp. 63-71). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-540-28650-9_4

Rawat, K. S., & Malhan, I. V. (2019). A Hybrid Classification Method Based on Machine Learning Classifiers to Predict Performance in Educational Data Mining. In C. R. Krishna, M. Dutta, & R. Kumar, *Proceedings of 2nd International Conference on Communication, Computing and Networking* Singapore. https://doi.org/10.1007/978-981-13-1217-5_67

Robles, T., Alcarria, R., Martín, D., Morales, A., Navarro, M., Calero, R., Iglesias, S., & López, M. (2014, 13-16 May 2014). An Internet of Things-Based Model for Smart Water

Management. 2014 28th International Conference on Advanced Information Networking and Applications Workshops, https://doi.org/10.1109/WAINA.2014.129

Rodrigues, F., Pereira, F., & Ribeiro, B. (2014). Gaussian Process Classification and Active Learning with Multiple Annotators. Proceedings of the 31st International Conference on Machine Learning, https://fprodrigues.com/publications/gaussian-process-classification-and-active-learning-with-multiple-annotators/

Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation Forest: A New Classifier Ensemble Method. *IEEE transactions on pattern analysis and machine intelligence*, *28*(10), 1619-1630. https://doi.org/10.1109/TPAMI.2006.211

Roghani, B., Cherqui, F., Ahmadi, M., Le Gauffre, P., & Tabesh, M. (2019). Dealing with uncertainty in sewer condition assessment: Impact on inspection programs. *Automation in Construction*, *103*, 117-126. https://doi.org/10.1016/j.autcon.2019.03.012

Salihu, C., Hussein, M., Mohandes, S. R., & Zayed, T. (2022). Towards a comprehensive review of the deterioration factors and modeling for sewer pipelines: A hybrid of bibliometric, scientometric, and meta-analysis approach. *Journal of Cleaner Production*, *351*, 131460. https://doi.org/10.1016/j.jclepro.2022.131460

Salman, B. (2010). *Infrastructure management and deterioration risk assessment of wastewater collection systems* University of Cincinnati]. https://www.proquest.com/openview/da6d944a3e88b5c865d49f11c6750ca8/1?pq-origsite=gscholar&cbl=18750

Samsung. (2022). *Galaxy A42 5G*. Available online: https://www.samsung.com/us/smartphones/galaxy-a42-5g/ (accessed on 31 October 2022,

Sánchez, E. A., & Schröder, C. (2019). Land use and land cover mapping in wetlands one step closer to the ground: Sentinel-2 versus landsat 8. *Journal of Environmental Management*, *247*, 484-498. https://doi.org/10.1016/j.jenvman.2019.06.084

Schall, G., Zollmann, S., & Reitmayr, G. (2013). Smart Vidente: advances in mobile augmented reality for interactive visualization of underground infrastructure. *Personal and Ubiquitous Computing*, *17*(7), 1533-1549. https://doi.org/10.1007/s00779-012-0599-x

Sekar, V. R., Sinha, S. K., & Welling, S. M. (2013). Web-Based and Geospatially Enabled Risk Screening Tool for Water and Wastewater Pipeline Infrastructure Systems. *Journal of Pipeline Systems Engineering and Practice*, *4*(4), 04013003. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000139

Sempewo, J. I., & Kyokaali, L. (2019). Comparative performance of regression and the Markov based approach in the prediction of the future condition of a water distribution pipe network amidst data scarce situations: a case study of Kampala water, Uganda. *Water Practice and Technology*, *14*(4), 946-958. https://doi.org/10.2166/WPT.2019.075

Sen, P. C., Hajra, M., & Ghosh, M. (2020). Supervised Classification Algorithms in Machine Learning: A Survey and Review. In J. K. Mandal & D. Bhattacharya, *Emerging Technology in Modelling and Graphics* Singapore. https://doi.org/10.1007/978-981-13-7403-6_11

Sezer, E., Romero, D., Guedea, F., Macchi, M., & Emmanouilidis, C. (2018). An industry 4.0-enabled low cost predictive maintenance approach for smes. 2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC), https://doi.org/10.1109/ICE.2018.8436307

Shamsi, U. M. (2002). *GIS Tools for Water, Wastewater, and Stormwater Systems*. https://doi.org/10.1061/9780784405734

Shi, F. (2018). *Data-driven predictive analytics for water infrastructure condition assessment and management* [Text, https://open.library.ubc.ca/collections/24/items/1.0372323

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and computing*, *14*(3), 199-222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

Song, X. F., Zhang, Y., Gong, D.-w., & Sun, X.-y. (2021). Feature selection using bare-bones particle swarm optimization with mutual information. *Pattern Recognition*, *112*, 107804. https://doi.org/10.1016/j.patcog.2020.107804

Sousa, V., Matos, J. P., & Matias, N. (2014). Evaluation of artificial intelligence tool performance and uncertainty for predicting sewer structural condition. *Automation in Construction*, *44*, 84-91. https://doi.org/10.1016/j.autcon.2014.04.004

Stian, B., Mareike, A. B., Håkon, R., & Tom, B.-M. (2021). *Investment needs for water and wastewater 2021 – 2040*. N. V. BA. https://295965-www.web.tornado-node.net/wp-content/uploads/Rapport259.pdf

Susto, G. A., Schirru, A., Pampuri, S., McLoone, S., & Beghi, A. (2015). Machine Learning for Predictive Maintenance: A Multiple Classifier Approach. *IEEE Transactions on Industrial Informatics*, *11*(3), 812-820. https://doi.org/10.1109/TII.2014.2349359

Tashi, Q. A., Abdulkadir, S. J., Rais, H. M., Mirjalili, S., & Alhussian, H. (2020). Approaches to Multi-Objective Feature Selection: A Systematic Literature Review. *IEEE Access*, *8*, 125076-125096. https://doi.org/10.1109/ACCESS.2020.3007291

Ting, K. M., & Witten, I. H. (1997). *Stacking Bagged and Dagged Models* [Working Paper](1170-487X). https://hdl.handle.net/10289/1072

Trafalis, T. B., & Ince, H. (2000, 27-27 July 2000). Support vector machine for regression and applications to financial forecasting. Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, https://doi.org/10.1109/IJCNN.2000.859420

Tran, D. H., Nguyen, A. W. M., Perera, B. J. C., Burn, S., & Davis, P. (2006). Application of probabilistic neural networks in modelling structural deterioration of stormwater pipes. *Urban Water Journal*, *3*(3), 175-184. https://doi.org/10.1080/15730620600961684

Tran, H. D. (2007). *Investigation of deterioration models for stormwater pipe systems* [PhD thesis, Victoria University]. https://vuir.vu.edu.au/1456/

Tran, H. D., & Nguyen, A. W. M. (2010). Classifying Structural Condition of Deteriorating Stormwater Pipes Using Support Vector Machine. *Pipelines 2010: Climbing New Peaks to Infrastructure Reliability - Renew, Rehab, and Reinvest - Proc. of the Pipelines 2010 Conference*, *386*, 857-866. https://doi.org/10.1061/41138(386)82

Tscheikner, G. F., Caradot, N., Cherqui, F., Leitão, J. P., Ahmadi, M., Langeveld, J. G., Le Gat, Y., Scholten, L., Roghani, B., Rodríguez, J. P., Lepot, M., Stegeman, B., Heinrichsen, A., Kropp, I., Kerres, K., Almeida, M. d. C., Bach, P. M., Moy de Vitry, M., Sá Marques, A., . . . Clemens, F. (2019). Sewer asset management – state of the art and research needs. *Urban Water Journal*, *16*(9), 662-675. https://doi.org/10.1080/1573062X.2020.1713382

Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A. (2019). Comparing different supervised machine learning algorithms for disease prediction. *BMC Medical Informatics and Decision Making*, *19*(1), 281. https://doi.org/10.1186/s12911-019-1004-8

Vairavamoorthy, K., Yan, J., Galgale, H. M., & Gorantiwar, S. D. (2007). IRA-WDS: A GIS-based risk analysis tool for water distribution systems. *Environmental Modelling &*

Software, *22*(7), 951-965. https://doi.org/10.1016/j.envsoft.2006.05.027

Vitorino, D., Coelho, S. T., Santos, P., Sheets, S., Jurkovac, B., & Amado, C. (2014). A Random Forest Algorithm Applied to Condition-based Wastewater Deterioration Modeling and Forecasting. *Procedia Engineering*, *89*, 401-410. https://doi.org/10.1016/j.proeng.2014.11.205

Vladeanu, G., Matthews, J., & Asce, M. (2019). Wastewater Pipe Condition Rating Model Using Multicriteria Decision Analysis. *Journal of Water Resources Planning and Management*, *145*(12), 10. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001134

Wang, Y., Hong, H., Chen, W., Li, S., Pamučar, D., Gigović, L., Drobnjak, S., Bui, D. T., & Duan, H. (2019). A Hybrid GIS Multi-Criteria Decision-Making Method for Flood Susceptibility Mapping at Shangyou, China. *Remote Sensing*, *11*(1), 62. https://doi.org/10.3390/rs11010062

Wauters, M., & Vanhoucke, M. (2014). Support Vector Machine Regression for project control forecasting. *Automation in Construction*, *47*, 92-106. https://doi.org/10.1016/j.autcon.2014.07.014

Wei, B., Chen, L., Li, H., Yuan, D., & Wang, G. (2020). Optimized prediction model for concrete dam displacement based on signal residual amendment. *Applied Mathematical Modelling*, *78*, 20-36. https://doi.org/10.1016/j.apm.2019.09.046

Witten, I. H., Frank, E., Trigg, L. E., Hall, M. A., Holmes, G., & Cunningham, S. J. (1999). *Weka: Practical machine learning tools and techniques with Java implementations* [Working Paper](Computer Science Working Papers, Issue. https://hdl.handle.net/10289/1040

WSAA. (2013). *Conduit Inspection Reporting Code of Australia*. https://www.wsaa.asn.au/sites/default/files/products/extracts/WSA%2005%20Version%203_1_Code%20Contents%20Extract_3.pdf

Yao, Z., & Ruzzo, W. L. (2006). A Regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. *BMC Bioinformatics*, *7*(1), S11. https://doi.org/10.1186/1471-2105-7-S1-S11

Yin, X., Chen, Y., Bouferguene, A., & Al-Hussein, M. (2020). Data-driven bi-level sewer pipe deterioration model: Design and analysis. *Automation in Construction*, *116*, 103181. https://doi.org/10.1016/j.autcon.2020.103181

Yin, X., Chen, Y., Bouferguene, A., Zaman, H., Al-Hussein, M., & Kurach, L. (2020). A deep learning-based framework for an automated defect detection system for sewer pipes. *Automation in Construction*, *109*, 17. https://doi.org/10.1016/j.autcon.2019.102967

Yuan, Y., Wu, L., & Zhang, X. (2021). Gini-Impurity Index Analysis. *IEEE Transactions on Information Forensics and Security*, *16*, 3154-3169. https://doi.org/10.1109/TIFS.2021.3076932

Zamanian, S., Hur, J., & Shafieezadeh, A. (2020). A high-fidelity computational investigation of buried concrete sewer pipes exposed to truckloads and corrosion deterioration. *Engineering Structures*, *221*, 111043. https://doi.org/10.1016/j.engstruct.2020.111043

Zendehboudi, A., Baseer, M. A., & Saidur, R. (2018). Application of support vector machine models for forecasting solar and wind energy resources: A review. *Journal of Cleaner Production*, *199*, 272-285. https://doi.org/10.1016/j.jclepro.2018.07.164

Zhang, D., Lindholm, G., & Ratnaweera, H. (2018). Use long short-term memory to enhance

Internet of Things for combined sewer overflow monitoring. *Journal of Hydrology*, *556*, 409-418. https://doi.org/10.1016/j.jhydrol.2017.11.018

Zounemat-Kermani, M., Stephan, D., Barjenbruch, M., & Hinkelmann, R. (2020). Ensemble data mining modeling in corrosion of concrete sewer: A comparative study of network-based (MLPNN & RBFNN) and tree-based (RF, CHAID, & CART) models. *Advanced Engineering Informatics*, *43*, 101030. https://doi.org/10.1016/j.aei.2019.101030

Appendix A

Appended Papers

Paper I

Lam Van Nguyen and Seidu Razak (2022). Application of Regression-Based Machine Learning Algorithms in Sewer Condition Assessment for Ålesund city, Norway. *Water*, *14*(24), 3993. https://doi.org/10.3390/w14243993.

water

IMPACT FACTOR 3.530

CITESCORE 4.8

Article

Application of Regression-Based Machine Learning Algorithms in Sewer Condition Assessment for Ålesund City, Norway

Lam Van Nguyen and Razak Seidu

Special Issue
Urban Sewer System Management
Edited by
Dr. Keshab Sharma and Dr. Oriol Gutierrez

MDPI

https://doi.org/10.3390/w14243993

water

MDPI

Article

Application of Regression-Based Machine Learning Algorithms in Sewer Condition Assessment for Ålesund City, Norway

Lam Van Nguyen [1,2,*] and Razak Seidu [1]

[1] Smart Water and Environmental Engineering Group, Department of Ocean Operations and Civil Engineering, Faculty of Engineering, Norwegian University of Science and Technology, N-6025 Ålesund, Norway
[2] Department of Geodesy, Faculty of Geomatics and Land Administration, Hanoi University of Mining and Geology, 18 Pho Vien, Duc Thang, Bac Tu Liem, Hanoi 100000, Vietnam
* Correspondence: lam.v.nguyen@ntnu.no

Abstract: Predicting the condition of sewer pipes plays a vital role in the formulation of predictive maintenance strategies to ensure the efficient renewal of sewer pipes. This study explores the potential application of ten machine learning (ML) algorithms to predict sewer pipe conditions in Ålesund, Norway. Ten physical factors (age, diameter, depth, slope, length, pipe type, material, network type, pipe form, and connection type) and ten environmental factors (rainfall, geology, landslide area, population, land use, building area, groundwater, traffic volume, distance to road, and soil type) were used to develop the ML models. The filter, wrapper, and embedded methods were used to assess the significance of the input factors. A dataset consisting of 1159 inspected sewer pipes was used to construct the sewer condition models, and 290 remaining inspections were used to verify the models. The results showed that sewer material and age are the most significant factors, otherwise the network type is the least contributor affecting the sewer conditions in the study area. Among the considered ML models, the Extra Trees Regression ($R^2 = 0.90$, MAE = 11.37, and RMSE = 40.75) outperformed the other ML models and it is recommended for predicting sewer conditions for the study area. The results of this study can support utilities and relevant agencies in planning predictive maintenance strategies for their sewer networks.

Keywords: sewer network; condition assessment; machine learning; GIS

check for updates

Citation: Nguyen, L.V.; Seidu, R. Application of Regression-Based Machine Learning Algorithms in Sewer Condition Assessment for Ålesund City, Norway. *Water* **2022**, *14*, 3993. https://doi.org/10.3390/w14243993

Academic Editors: Keshab Sharma and Oriol Gutierrez

Received: 25 October 2022
Accepted: 25 November 2022
Published: 7 December 2022

1. Introduction

A sewer network is one of the most important components of the urban water infrastructure [1]. This network plays a vital role in the collection and transport of wastewater and stormwater from the urban landscape to reduce the incidence of flooding, mitigate environmental pollution and protect public health [2,3]. However, sewer networks in operation are subjected to different intrinsic and extrinsic factors that contribute to their deterioration and failures [4], thereby preventing the network from realizing its intended objectives. Failures in the sewer system often result in debilitating impacts on infrastructure, the environment, and public health with a significant economic burden on society [5]. Therefore, investments in maintenance programs that reduce the incidence of sewer pipe failure are a priority in many countries [6,7].

Maintenance management approaches can be generally categorized into *Reactive Maintenance* (RaM), *Preventive Maintenance* (PvM), and *Predictive Maintenance* (PdM) [8]. The RaM, or run-to-failure, is the simplest approach that is only implemented when break(s) in sewer pipes occur. This reactive maintenance approach is also the least effective one. The PvM, or proactive maintenance, is implemented based on predetermined intervals (usually time or event-based triggers). This approach is more effective than the RaM method because many failures can be prevented. However, several unnecessary corrective actions are usually implemented [8]. The PdM approach mainly focuses on assessing sewer pipes based on condition assessment. In this way historical data are combined with analytic and prediction tools to predict the condition of sewer pipes, and maintenance strategies are scheduled.

A predictive maintenance strategy cannot be implemented effectively without a deep understanding of the system, and an efficient water management strategy requires a proper condition assessment framework [9]. Many condition assessment models have been developed in literature and they can be divided into three main groups: *physical*, *statistical*, and *machine learning* models [10]. The physical models assess the deterioration process based on the influence of the physical properties and the mechanical processes in the sewer pipes [11,12]. However, these types of models are suitable for the construction period and initial operation, and data for the simulation of the deterioration mechanism are not always available [13]. The statistical models (e.g., linear regression, cohort survival model, or Markov chains) can produce good accuracy but they are limited in revealing the physical relationship between limited physical factors and the target [14]. In recent times, machine learning (ML) algorithms have been widely used to model sewer pipe deterioration because they are capable of handling the complex non-linear interlinked processes involved in the deterioration of sewer pipes [15]. However, a large number of input factors and observations are needed to improve the accuracy of these models [6].

The output of a mathematical model in general, and an ML model in particular, significantly depends on the quality of input data. Factors considered for building condition assessment models can be divided into three groups: *physical*, *operational*, and *environmental* factors [16]. In general, physical data on most sewer networks are readily available. The same can be said about data on environmental factors. However, when it comes to operational data, it is most often scarce [9]. Therefore, considering the quality of input data plays a vital role in improving the ML models' predictive performance. The importance of the input data should be assessed to prioritize inputs while collecting and preparing data before building condition assessment models. Therefore, defining significant factors for building condition assessment models is a key task to improve the efficiency of the predictive models. This task is accomplished via feature selection methods that can be grouped into *filter*, *wrapper*, and *embedded* methods [17]. The filter methods assess the importance of input variables, mainly based on their statistical properties and relationship with the output variable. The wrapper methods select a sensitive subset of features by adding and removing subsets based on the performance of the model. In the embedded methods, the effectiveness of input variables is assessed by tuning predictive models [17]. The important degree of different feature selection methods may be incompatible due to randomness in selecting and combining subsets [18]. In this study, all three types of feature selection are investigated, and the insignificant features are eliminated based on a consensus of the three methods used.

Closed-circuit television (CCTV) is the most widely used method for assessing the condition of the sewer network because it can directly provide sewer pipes statements with very high accuracy [19]. To inspect the sewer pipes' status, a camera is put inside pipelines or drains without needing to conduct more invasive methods like digging, removing walls, or flooring to gain access to plumbing. Based on the recovered CCTV videos, the trained inspectors can monitor the status of sewers in real-time (while controlling the camera inside pipes) or offline (after finishing the inspection). Depending on the status (e.g., roots, sediments, cracks, deformations), the local and global damaged score can be assigned for the particular sewer pipe and rehabilitation schedules and be prioritized [20]. However, this method is time-consuming and expensive because workers need to inspect sewer pipes individually. As a result, only a small fraction of all sewer pipes, depending on their role and importance, are inspected during a specific period [21]. This data can be used to construct sewer condition models using ML algorithms, and derived ML models can be used to predict the sewer's status for the entire network.

Although ML models used for regression problems have been successfully applied in many fields [22,23], their application in sewer status prediction is still limited. Moreover, no ML model is the best in all cases for modeling sewer deterioration and a comprehensive comparison of prediction performance between these models needs to be investigated. In the literature the influence of factors on sewer condition is still controversial, and the determination of the significance of these factors is valuable for local water utilities to prioritize their maintenance and rehabilitation activities. This work is an attempt to partly fill these gaps by developing ML models for sewer condition status prediction and assessing the importance of factors affecting sewer condition. In this study, ten state-of-the-art ML algorithms are explored to predict the damage score of the sewer network in Ålesund city, Norway. Ten physical factors (i.e., age, diameter, depth, slope, length, pipe type, material, network type, pipe form, and connection type) and ten environmental factors (i.e., rainfall, geology, landslide area, population, land use, building area, groundwater level, traffic volume, distance to road, and soil type) were used for training ML models. The best model is selected to predict the sewer's damage score and it can help water engineers/workers predict sewer status on the large scale in a short time. Consequently, the model effectively supports water network management and maintenance. The final condition assessment model can help local water utilities/managers to have an overview of the status of the sewer network and support maintenance strategies in the future.

The rest of the paper is organized as follows: Section 2 describes the study area and data used. The overview of the feature selection techniques, the basic theory of used algorithms, the criteria for evaluating the developed models, and the framework for modeling the condition of sewer pipes are also discussed in this section. Section 3 presents and discusses the results. Finally, Section 4 presents the conclusions of the work.

2. Materials and Methods

2.1. Sewer Network

The sewer network of Ålesund, which is a coastal city in the eastern part of Norway, was used as a case study (Figure 1). The city is located between longitudes of 62°25′07″ E and 62°30′37″ E and between latitudes of 6°05′08″ N and 6°40′56″ N, with an area of 607.3 km^2 and a population of 66,600 in 2021 [24]. With the characteristics of an ocean climate, the average annual rainfall in Ålesund city is 2100 mm with an average temperature of 8.1 °C [25].

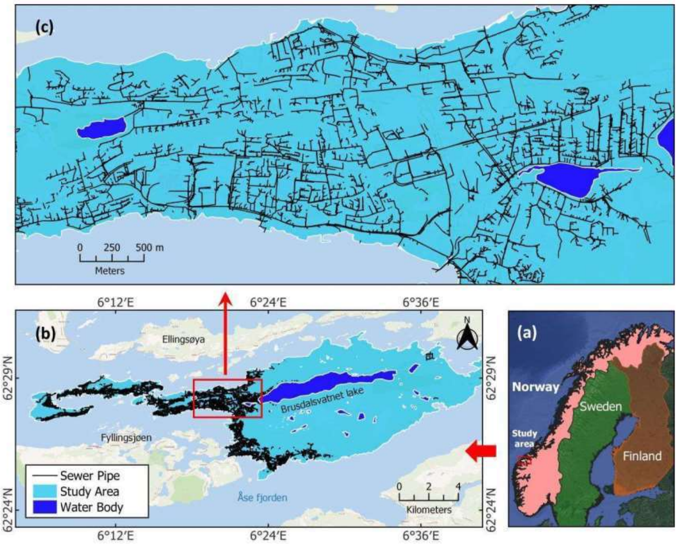


Figure 1. Location of the study area and sewer network ((**a**) Location of Ålesund in Norway, (**b**) entire sewer network in Ålesund city, and (**c**) sewer network in a selected area of Ålesund).

The sewer network is located in the central area of the city where the elevation fluctuates from 0 to 100 m. The western and eastern parts of the city are hilly and mountainous areas with altitudes of 300 m and 500 m, respectively. As a coastal city, Ålesund has been affected by consequences of the climate change such as extreme weather events and sea level rise [26,27]. Figure 1a shows an overview of the study area in Norway, while Figure 1b shows the entire sewer network in Ålesund city, and Figure 1c captures the sewer network in a specific area.

2.2. Proposed Framework for Sewer Condition Assessment

The framework for the sewer condition assessment is shown in Figure 2. The main steps for constructing the models include: (1) collecting and processing physical and environmental data; (2) digitalizing data using Geographic Information System (GIS) tools; (3) splitting the data into training and testing datasets; (4) determining feature importance and removing redundant features; (5) constructing the ML models; and (6) validating and selecting ML model for sewer condition assessment.

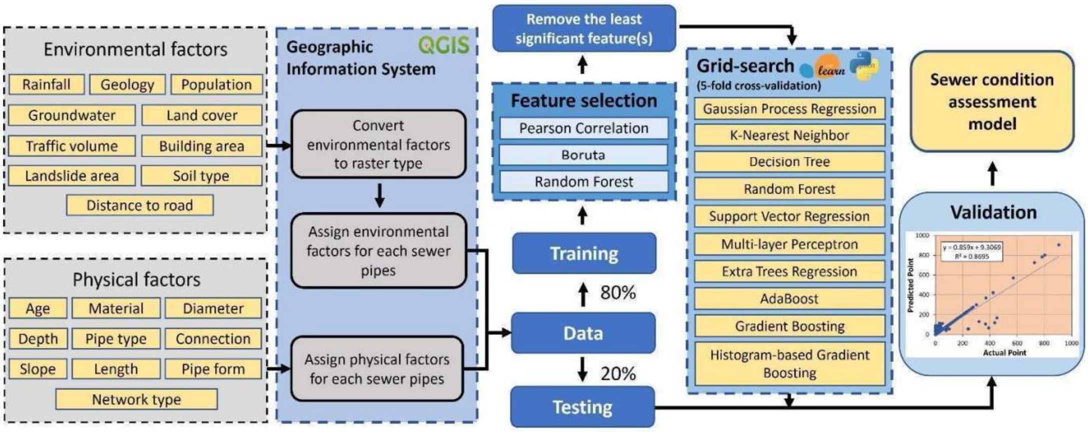


Figure 2. The flowchart for sewer condition assessment modeling.

2.3. Data Used

2.3.1. Physical Factors

In this study, ten physical factors comprising age, diameter, depth, slope, length, pipe type, material, network type, pipe form, and connection type of sewer network were used. Seven of these factors including diameter, length, pipe type, material, network type, pipe form, and connection type were obtained from the database of Ålesund Municipality.

The remaining factors (i.e., age, depth, and slope) were obtained by doing additional computations. Specifically, the age of sewer pipes was calculated as the difference between the installation and the inspection years [28]. The depth of pipes was computed as the distance from the ground surface to the centroid of the pipe. The elevation of the ground surface was obtained from a Digital Elevation Model (DEM) with pixels of 1 m × 1 m received from the Norwegian Mapping Authority (NMA) (https://www.kartverket.no/ (accessed on 20 March 2020)). The slope of sewer pipes was computed from the elevation of the start point and the endpoint of each pipe. All the above computations were implemented using the open-source QGIS software.

The processed data in the entire city consisted of 31,293 pipes with a total length of 703.0 km (Figure 1). In this dataset, the length and the number of wastewater, stormwater, and combined pipes were 339.0 km, 276.6 km, 87.4 km and 15,409, 12,722, 3163, respectively. After comparing the pipe index in this dataset with the corresponding pipes index in the inspected dataset, a total of 1449 sewer pipes were used to train and validate the condition assessment models. The detail of the physical factors is summarized in Table 1.

Table 1. Summary of physical factors in this study.

Factor	Type	Min	Max	Average	Std
Age (year)	Numeric	1.0	104.0	34.4	25.3
Diameter (mm)	Numeric	110.0	1000.0	248.4	98.6
Depth (m)	Numeric	−0.1	−7.8	−1.8	1.2
Slope (°)	Numeric	−17.4	+34.6	+2.7	4.4
Length (m)	Numeric	1.0	177.5	38.6	21.3
Pipe type	Categorical				
Material	Categorical				
Network type	Categorical				
Pipe form	Categorical				
Connection type	Categorical				

2.3.2. Environmental Factors

The environmental factors used in this study are presented in Table 2. The selection of these environmental factors was based on the study of Hawari, Alkadour, Elmasry and Zayed [10].

Table 2. Environmental factors in this analysis.

Factor	Spatial Resolution	GIS Type	Source
Rainfall		Point	NCSC
Geology	1:50,000	Polygon	NMA
Landslide area	1:5000	Polygon	NMA
Population	250 m × 250 m	Grid	NMA
Land use	10 m × 10 m	Grid	Sentinel-2 Images
Building area	1:5000	Polygon	NMA
Groundwater		Point	NGS
Traffic volume	5 m × 5 m	Grid	NPRA
Distance to road	5 m × 5 m	Grid	NMA
Soil type	1:50,000	Polygon	NMA

The rainfall map was interpolated from the annual average rainfall provided by nine weather stations (Table A1) inside and outside of the study area using the Inverse Distance Weighting (IDW) method [29]. Data on annual average rainfall at the weather stations were obtained from the Norwegian Climate Service Center (NCSC) (https://klimaservicesenter.no/ (accessed on 14 February 2020)).

The land use map was obtained from the Sentinel-2 images Level 1C downloaded from the website of Copernicus Open Access Hub (https://scihub.copernicus.eu/ (accessed on 17 January 2020)). A Google background satellite image was superimposed on the Sentinel-2 image to get the land use classifications (e.g., forest areas, roads, residential areas, etc.). The samples of different land uses were taken and assigned specific values, and different bands of the Sentinel-2 image overlapped. Finally, object-based classification was applied to cluster areas in the image into different objects based on given land uses [30].

The groundwater map for the study area was interpolated from 31 drills received from the Norwegian Geological Survey (NGS) (https://www.ngu.no/ (accessed on 24 April 2020)) using the IDW method. The map of traffic volume was received from the Norwegian Public Roads Administration (NPRA) (https://www.vegvesen.no/en/ (accessed on 13 February 2020)). Finally, the environmental factors maps were resampled to a spatial resolution of 5 m × 5 m and transformed into a grid database using the QGIS before developing ML models. Maps of the environmental factors are shown in Figure A1.

Categorical factors such as pipe type, material, network type, pipe form, connection type, land use, road class, geology, building area, landslide area, and soil type were coded by integer values before constructing ML algorithms and feature selection. Furthermore, concrete, other, polypropylene, and polyvinyl chloride (PVC) pipes were coded by values 0, 1, 2, 3, and 4, respectively, for analysis in this study.

2.3.3. Sewer Damage Score

The damaged scores are obtained from CCTV datasets, and for this study damage scores based on the CCTV dataset of 1449 pipes (55.8 km) provided by the Ålesund Municipality were used for the condition assessment model. All damage score data were processed and integrated into a GIS database.

2.4. Description of Feature Selection Methods

2.4.1. Pearson Correlation Method

Pearson correlation is a filter feature selection method that defines the linear relationship between independent variables and the output target (e.g., a higher correlation value reflects a stronger relationship between input and output) [31]. Pearson's correlation coefficient (PR) falls between −1 and +1 to indicate the extent to which two variables are linearly related. A value closer to 0 implies a weaker correlation, and a value closer to +1 (or −1) implies a stronger positive (or negative) correlation. In other words, the variables that have PR closer to +1 (or −1) are more important than the variables closer to 0 [32].

The Pearson correlation is computed as follows [17]:

$$PR_i = \frac{Cov(x_i, y)}{\sqrt{Var(x_i) \times Var(y)}}, \tag{1}$$

where x_i is the i^{th} variable, y is the output, $Cov()$ and $Var()$ are the covariance and variance, respectively.

2.4.2. Boruta Method

Boruta works as a wrapper algorithm around Random Forest [33]. This method is a suitable candidate for reducing the dimensionality of the data [34]. The Boruta algorithm uses the Out-of-Bag (OOB) error to define the important score of the input features [33]. Steps for implementing the Boruta algorithm are shown in Figure 3.

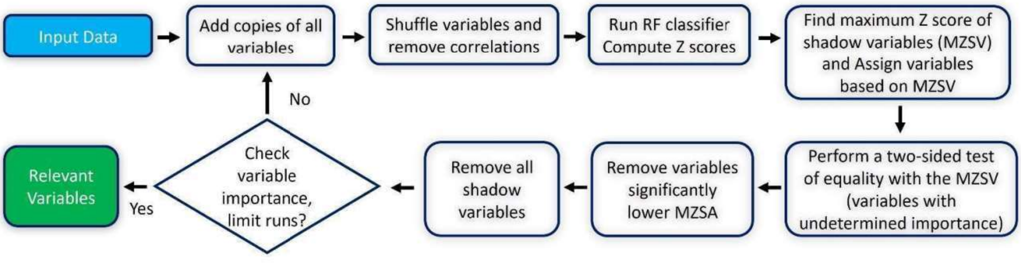

Input Data
Add copies of all variables
Shuffle variables and remove correlations
Run RF classifier Compute Z scores
Find maximum Z score of shadow variables (MZSV) and Assign variables based on MZSV
Perform a two-sided test of equality with the MZSV (variables with undetermined importance)
Remove variables significantly lower MZSA
Remove all shadow variables
Check variable importance, limit runs?
No
Yes
Relevant Variables

Figure 3. Overview of the Boruta feature selection method.

The Z-Score in the Boruta algorithm is computed by the following equation [33]:

$$Z - Score = \frac{1}{n \times \sigma} \sum_{i=1}^{m} \frac{\sum_{i \in E_{OOB}} F(y_i = f(x_i)) - \sum_{i \in E_{OOB}} F\left(y_i = f\left(x_i^j\right)\right)}{|E_{OOB}|}, \tag{2}$$

where n is the number of decision trees, $F(\cdot)$ is the indicator function, $y_i = f(x_i)$ and $y_i = f\left(x_i^j\right)$ are predicted values before and after permuting, E_{OOB} is the prediction error of each of the training samples based on bootstrap aggregation, and σ is the standard deviation of accuracy losses.

2.4.3. Random Forest Method

Random Forest is one of the most popular embedded feature selection methods [35]. For regression problems, the final value of the importance of variable i (I_i) can be computed as follows [36]:

$$\begin{cases} I_i = \frac{\overline{\delta_{bi}}}{\sigma_{\delta_{bi}}/\sqrt{B}} \\ \overline{\delta_{bi}} = \frac{1}{B}\sum\limits_{b=1}^{B}\left(MSE_{before} - MSE_{after}\right) = \frac{1}{B}\sum\limits_{b=1}^{B}\delta_{bi} \end{cases}, \tag{3}$$

where $\overline{\delta_{bi}}$ is the average importance of variable i ($\overline{I_i}$) for each tree b, B is an average overall tree, $\sigma_{\delta_{bi}}$ is the standard deviation of the δ_{bi}, and MSE_{before} and MSE_{after} are mean squared error before and after permuting and root mean squared error (RMSE).

Feature selection methods are implemented using the related packages in R software, and the ML library Scikit-Learn is used to construct ML models. GIS is used to collect, preprocess, and aggregate data before constructing condition assessment models. In this paper, the libraries "corrplot", "Boruta", and "randomforest" in R were used to implement the Pearson correlation, Boruta, and Random Forest feature selection methods, respectively.

2.5. Regression-Based Machine Learning Algorithms

2.5.1. Gaussian Process Regression

A Gaussian Process Regression (GPR) is a subset of Gaussian Processes used for dealing with regression problems [37]. The GPR is an effective tool for interpolating data points in high-dimensional input space and can be defined as follows [38]:

$$Y(X) = GP\left(M(X_i), Cov\left(X_i, X_j\right)\right) + \epsilon(X),\ \ i,j = 1,\ldots,n, \tag{4}$$

where n is the total number of inspected sewer pipes, Y is the damage score, $\epsilon(X)$ is the observation error, $M(X_i)$ and $Cov\left(X_i, X_j\right)$ are the mean and covariance functions, respectively.

In Gaussian Process, the covariance function is determined using a single or a combination of kernel functions (i.e., Radial-basis function, Dot-Product, Matérn, Rational Quadratic, Exp-Sine-Squared, and White kernels) and their hyperparameters (i.e., noise level, length-scale, scale mixture, or periodicity) [39]. This method has been applied for assessing sewer deterioration in previous studies but only for specific purposes such as sediment-related blockage or corrosion [40,41].

2.5.2. K-Nearest Neighbor

A K-Nearest Neighbor (KNN) regression introduced by Lall and Sharma [42] is a non-parametric method that approximates the association between factors (e.g., physical and environmental factors) and the sewer damage score by averaging the observations in the same neighborhood. The KNN model predicts the status of new sewer pipes based on using the similarity of K neighbor pipes in the training dataset. Therefore, the main advantages of KNN are the quick computational time, easy interpretability, versatility, and no need for any assumptions [43]. However, this algorithm is sensitive to irrelevant features which can be addressed by feature selection. Moreover, because it stores the distances from the new test point to all the training data points during implementation, this algorithm can be costly in the case of large datasets.

The basic steps for KNN implementation are as follows [44]:

- *Step 1:* Loading sewer inspection (training) dataset for constructing the KNN model;
- *Step 2:* Choose the value of K neighbors to define the nearest data points;
- *Step 3:* For each new sewer data point: (1) calculated distance between a new sewer data point to the training data points; (2) sort calculated distances in ascending order; (3) select the top K features from the sorted array; and (4) assign the sewer dam-

age score to the new data point using weights calculated from distances neighbors' data points;

- *Step 4:* Repeat steps 2 and 3 until all new sewer data points have been assigned new values.

2.5.3. Classification and Regression Trees

A Decision Tree (DT) regression creates regression models in the form of a tree structure in which the sewer training dataset is split into smaller and smaller subsets while at the same time an associated decision tree is developed. The decision tree consists of four basic components: root, internal nodes, leaf nodes, and branches. The root node contains all the factors, an internal node can contain two or more branches that are associated with a decision function, and the leaf node indicates the sewer damage score. A decision tree can be constructed via several steps [45]: (1) assigning all observations in the root node; (2) splitting the root node into branches based on the predicted sewer damage score using the decision function; (3) distributing observations on the higher node to the lower nodes; and (4) repeating the process until all sewer pipes have been processed.

A Classification and Regression Trees (CART) algorithm was used in this study. The decision trees created by CART have two branches for each decision node. Difference from the decision tree for classification, which uses Gini Impurity or Entropy values as criteria for splitting root/decision nodes, the "goodness" criterion is applied in the CART algorithm to split root/decision nodes and is computed as follows [46]:

$$f(s|t) = 2P_L P_R \sum_{i=1}^{n} |P(i|t_L) - P(i|t_R)|, \quad (5)$$

where n is the number of sewer inspections, $f(s|t)$ is a measure of "goodness of fit", t_L and t_R are the left and right children of a candidate split s at node t, respectively, P_L and P_R are the proportions of records at t_L and t_R, respectively, $P(i|t_L)$ and $P(i|t_R)$ are the proportions of class i at t_L and t_R, respectively.

2.5.4. Random Forest

Random Forest (RF) regression is an ensemble learning method that uses multiple decision trees as base learning models for regression problems. The bagging (or bootstrap aggregation) algorithm is generally used to create the RF model. In this way, each decision tree in the RF is created from different samples at each node and produces an individual prediction. This model generates hundreds or thousands of regression decision trees and the average sewer status predicted from the individual trees is calculated for the final result [47]. As a result, the RF regression model generally has higher performance compared to the DT because it can avoid the correlation of different trees and the final results are obtained from the diversity of the trees [48].

2.5.5. Support Vector Regression

A Support Vector Regression (SVR) is one type of Support Vector Machine used for regression problems. This algorithm creates and finds the best-fit hyperplane in n-dimensional space that is close to as many of the data points as possible [49]. For regression problems, the linear form of the hyperplane can be computed as follows [50]:

$$f(x) = wx + b, \quad (6)$$

where $f(x)$ is the predicted value, x is the input vector of the data point, w and b are the slope and intercept.

The goal function of the SVR model can be defined as follows [51]:

$$\begin{cases} \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) K(x_i, x) + b \\ subject\ to \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0,\ \ \alpha_i, \alpha_i^* \in [0, C] \end{cases}, \tag{7}$$

where $f(x)$ is the predicted value, n is the number of sewer inspections, x is the input vector of the data point, w and b are the slope and intercept, respectively, α_i, α_i^* are Lagrange multipliers, the constant $C > 0$ is the trade-off between the flatness of the $f(x)$ and the amount up to which deviations larger than the insensitive loss function, $K(x_i, x)$ is kernel function (e.g., linear function, polynomial function, radial basis function, or sigmoid function).

2.5.6. Multi-Layer Perceptron Neural Network

A Multi-layer Perceptron Neural Network (MLP) is a fully connected class of feed-forward artificial neural networks. This architecture normally consists of three or more layers (i.e., an input layer, an output layer, and one or more hidden layers) and each layer contains different neurons. In general, the number of neurons in the input layer is equal to the number of input factors, the number of neurons in the output layer is equal to one for the regression problem, and the number of hidden layers and hidden neurons fluctuates depending on the complexity of the MLP architecture. Determining the number of hidden layers and hidden neurons is generally implemented using a trial-and-error approach [52].

Each neuron j in the hidden layer computes its input signals x_i and produces its output y_j based on the following equation:

$$y_j = f\left(\sum_{j,i=1}^{n} w_{ji} x_i + b_i\right), \tag{8}$$

where n is the number of sewer inspections in the training dataset; f is an activation function; w_{ji} and b are connection weight and bias, respectively.

In this study, a single-layer MLP architecture was used. The number of hidden neurons, various activation functions in the hidden layer, and several optimization solvers were tuned using the Scikit-learn ML library. The early-stopping technique was used to avoid overfitting while training the model.

2.5.7. Extra Trees Regression

An Extra Trees Regression (ETR) is a tree-based structure ensemble learning algorithm used for regression problems. This algorithm uses an entire learning sample (instead of a bootstrap replica) to split nodes by choosing cut points entirely randomly. In the regression problem, the result is obtained by averaging predictions from decision trees. The relative variance reduction is used as the score measure in the regression problems for the ETR algorithm [53]:

$$Score(s, D) = \frac{Var(y|S) - \frac{|S_l|}{|S|} Var(y|S_l) - \frac{|S_r|}{|S|} Var(y|S_r)}{Var(y|S)}, \tag{9}$$

where $Var(y|S)$ is the variance of the output y in the sample S, S_l and S_r are two subsets of cases from the sample S corresponding to the two outcomes of a split s, respectively.

2.5.8. AdaBoost

An AdaBoost regression (AdaBoost) is an ensemble learning method that uses an adaptive resampling approach to improve predictive performance from the mistakes of the base algorithm [54]. The basic idea of the AdaBoost algorithm is to build models via iterations in which models in the next iterations are built to rectify the errors present in the previous model. This process is ended when it reaches a terminated condition, and the final model is obtained from a weighted sum of all the base models.

Although AdaBoost can be used to combine various weak base learners, a combination of AdaBoost with the decision tree is often referred to as the best out-of-the-box classifier [55] and is used in this study.

2.5.9. Gradient Boosting

A Gradient Boosting Regression (GBR) improves predictive performance by combining weaker learners with strong learners via the iteration approach [56]. The decision tree is one of the most popular weak learners in the GBR [39]. The gradient boosting algorithm suffers from over-fitting if the iteration process is not regularized properly [57].

The decision tree solves the problems by transforming the data into a tree representation. Each internal node of the tree denotes an attribute, and each leaf node denotes a prediction. In the gradient boosting approach, decision trees have been added repeatedly and the next decision tree will correct the previous decision tree error [58].

2.5.10. Histogram-Based Gradient Boosting

A Histogram-based Gradient Boosting regression (HGB) is a modified version of the GBR to significantly increase the speed of decision tree split. During the training period of the HGB, factors were divided into bins and a histogram of factors was constructed [59]. The number of histogram bins is significantly less than the number of sewer inspections in the training set. The sewer damage score was predicted using the best split points based on the feature histograms [60].

2.6. Model Implementation

An ML regression model tries to fit the data by drawing rule(s) that minimize the difference between the actual value and the predicted value. The smaller the differences are, the better the model behaved for the point. Different ML models effectively fit with different hyperparameters to produce the optimal prediction. Therefore, the ML models need to be tuned to find the sensitive hyperparameters for the specific dataset. In this analysis, the Grid-Search (GS) method with a 5-fold cross-validation approach, which is integrated into the Scikit-learn ML library, was used to tune parameters for developed ML models.

The obtained GIS database was split into training and validation datasets to construct and validate ML models. In general, there is no consensus on the ratio of training and testing datasets when building an ML model. Choosing training and testing ratios mainly depends on the particular study and the subjective opinion of researchers. For example, a ratio of 80/20 was selected to build structural condition models in some studies [61]. In contrast, a ratio of 70/30 was used to predict the sewer's status [28]. In this study, the ratio of 80/20 was used to split the dataset. Accordingly, a total of 1159 sewer pipes were used to construct the ML models and 290 sewer pipes were used for model validation.

Feature selection methods were used to assess the significance of each factor. After that, the least significant factors influencing the sewer damage score were eliminated from the training data. The final data were used in the sewer condition models.

The constructed ML models were compared to select the best model for the sewer condition prediction. In this study, the predictive performance of the ten ML regression models was assessed using the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error ($RMSE$) expressed as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i^{act} - y_i^{pred}\right)^2}{\sum_{i=1}^{n}\left(y_i^{act} - \overline{y}\right)^2}, \tag{10}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|y_i^{act} - y_i^{pred}\right|, \tag{11}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i^{act} - y_i^{pred}\right)^2}, \tag{12}$$

where n is the total number of measurements; $\overline{y}$ is the mean value of the actual measurements; y_i^{act} and y_i^{pred} are the i^{th} actual and predicted measurements.

The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) was applied to rank the ML models [62]. This method is a common approach for ranking ML algorithms, using multiple criteria on a single dataset by choosing the alternatives that have the shortest distance to the positive-ideal solution and the longest distance to the negative-ideal solution [63]. These distances relate to the alternative weights that are used to compute the overall performance score [64]. Interested readers can find more detailed information on the TOPSIS in Behzadian, Khanmohammadi Otaghsara, Yazdani and Ignatius [63]. In this study, the package "TOPSIS" in R was used to implement the TOPSIS method [65].

3. Results and Discussions

3.1. Feature Selection

Results of feature selection using the filter, wrapper, and embedded methods are shown in Figure A2. The results revealed slight differences in determining the most significant factors by each feature selection method. For instance, while the RF feature selection method identifies slope as the most significant factor, followed by length, both the Pearson correlation and the Boruta methods identify age (PR = 0.30) and material (PR = −0.25) as the most significant factors. The Boruta method also identifies age (Z-score = 12) and material (Z-score = 10) as the most significant factors for sewer damage. In Figure A2a, the negative values represent the inverse relationships between physical/environmental variables and a sewer's damaged scores, and vice versa. A positive correlation between a continuous input variable and the output shows that when the values of the input increase, the value of the output increase as well [66]. For example, the sewer's age has a positive correlation (PR = 0.30) with the sewer's damaged score, showing that when the age of the sewer pipe increases (old pipes) the damaged score of the sewer pipe will rises (worse condition). Material has a negative correlation (PR = −0.25) with the damaged score indicating that sewer pipes in concrete material are more durable than sewer pipes in polypropylene and PVC materials.

In contrast, there is a less significant difference between the feature selection methods in terms of the least important determinations of sewer condition. For example, the Pearson correlation coefficients revealed network type and groundwater do not affect the sewer pipe condition (PRs = 0.00). Two factors associated with distance to road (PR = 0.01), land use and depth (PRs = −0.03), and diameter (PRs = 0.03) are the six lowest significant factors (Figure A2a). For the Boruta method, landslide area, building area, pipe form, network type, depth, and diameter were assessed as insignificant factors (Figure A2b). Network type, pipe form, landslide area, geology, pipe type, and connection were identified as the least significant factors in the RF feature selection method (Figure A2c).

Table 3 summarizes the importance of the factors from each feature selection method, where the number represents the important degree (1: the highest importance, 20: the lowest importance). The same important factors, which have similar PR values, are denoted by the slash. For example, the rainfall factor and connection factor have the same importance. In conclusion, all feature selection methods show that network type is the least significant factor. Therefore, this factor was eliminated from the dataset before building the condition assessment models.

3.2. Model Comparison

The optimal hyperparameters used for tuning the ML models are shown in Table A2. The performance of ten ML models was compared based on the training and validation phases as presented in Table 4.

Table 3. Summary of feature selection.

Factor	Pearson's Correlation	Boruta	Random Forest	Selection
Material	2	1	9	✔
Age	1	2	3	✔
Rainfall	8/9	3	4	✔
Traffic volume	12	4	8	✔
Population	7	5	6	✔
Connection	8/9	6	15	✔
Soil type	10	7	11	✔
Pipe type	4	8	16	✔
Groundwater	19/20	9	7	✔
Geology	13/14	10	17	✔
Length	11	11	2	✔
Slope	5/6	12	1	✔
Distance to road	18	13	13	✔
Land use	15/16/17	14	14	✔
Diameter	15/16/17	15	10	✔
Depth	15/16/17	16	5	✔
Network type	19/20	17	20	×
Pipe form	13/14	18	19	✔
Building area	3	19	12	✔
Landslide area	5/6	20	18	✔

Table 4. Performance of the machine learning models in this analysis.

Model	Training Dataset			Validation Dataset		
	R^2	MAE	RMSE	R^2	MAE	RMSE
GPR	1.00	0.00	0.00	0.86	14.36	46.95
KNN	0.09	88.17	196.24	0.07	76.07	122.96
DT	1.00	0.00	0.00	0.76	15.19	69.96
RF	0.95	31.03	75.61	0.81	30.59	58.19
SVR	0.33	47.18	182.85	0.38	36.82	108.07
MLP	0.18	87.23	185.86	0.10	73.93	126.07
ETR	1.00	0.00	0.03	0.90	11.37	40.75
AdaBoost	0.15	75.23	189.93	0.20	58.86	113.15
GB	0.42	76.18	163.71	0.20	63.20	113.45
HGB	0.16	80.24	188.13	0.17	64.18	117.10

The results in Table 4 show that the GPR, DT, and ETR models fit very well with the training dataset (the values of R^2 and errors are equal to 1.0 and 0.0, respectively). In contrast, the KNN model performed poorly in predicting the sewers' damage scores (R^2 is almost equal to zero) indicating the worst ML model. The predictive capability of the ML models on the testing dataset is presented in Figure A3.

In general, predictive models have been assessed as effective tools if they can effectively predict unseen data that are not used for model construction. Therefore, the validation data are used to assess the constructed ML models. In the validation phase, the ETR has the best performance (R^2 = 0.90, MAE = 11.37, RMSE = 40.75), followed by the GPR (R^2 = 0.86, MAE = 14.36, RMSE = 46.95) and the RF model (R^2 = 0.81, MAE = 30.59, RMSE = 58.19). The KNN (R^2 = 0.07, MAE = 76.07, RMSE = 122.96) and MLP (R^2 = 0.10, MAE = 73.93, RMSE = 126.07) performed poorly in predicting the condition status of the sewer pipes.

Even though all ensembles ETR, RF, HGB, AdaBoost, and GB use the DT as the base learner, their predictive performance is significantly different. For instance, the ETR and RF remarkably improve the predictive performance of the original DT algorithm (R^2 = 0.76, MAE = 15.19, RMSE = 69.96). In contrast, the HGB (R^2 = 0.17, MAE = 64.18, RMSE = 117.10), AdaBoost (R^2 = 0.20, MAE = 58.86, RMSE = 113.15), and GB (R^2 = 0.20, MAE = 63.20, and RMSE = 113.45) significantly reduce the predictive capability of the DT algorithm. These results show that the adaptive boosting and gradient boosting techniques are unsuitable approaches for the dataset in the study area; in contrast, the randomly generated threshold method (in the ETR algorithm) or the bootstrap aggregation method (in the RF algorithm) is a more suitable option.

The prediction performance of the KNN model mainly depends on the number of neighbors that were obtained based on similar characteristics [44]; limited data in the study area may not provide enough information for the KNN algorithm to effectively distinguish clusters resulting in the low prediction performance. The GPR model with high interpolating ability can deal with high-dimensional input for the complex process of sewer deterioration in the study area [38].

In this study, the MLP algorithm has a low prediction capability in modeling sewer pipes' damage scores. This agrees with a previous study in which the neural network-based models have lower performance in regression problems [67]. Similarly, the prediction capabilities of KNN, SVM, AdaBoost, GB, and HGB models were low in both training and validation datasets. The reason is that there are several sewer pipes that have excessive damage score values (over 1000); meanwhile, the majority of sewer pipes (approximately 90%) have damage score values below 1000. To test the prediction ability of ML algorithms in distinguishing these values, we prioritized using the original dataset. The results showed that the overmentioned models did not effectively distinguish the excessive values of sewer pipes, indicating they are unsuitable for the study area. In conclusion, among the constructed models, the ETR is the most suitable ML algorithm for modeling the sewer conditions in the study area.

The constructed ML algorithms have been ranked using the TOPSIS method and the results are shown in Table 5. According to these results, the ETR is the most suitable ML algorithm and the KNN is the worst ML algorithm for modeling the sewer's condition in Ålesund city.

Table 5. The rank of the machine learning algorithm.

Model	Score	Rank
ETR	1.0000	1
GPR	0.9476	2
DT	0.8202	3
RFR	0.7961	4
SVR	0.4336	5
AdaBoost	0.1993	6
GB	0.1710	7
HGB	0.1432	8
MLP	0.0318	9
KNN	0.0147	10

Although sewer damage scores can be used to predict sewer status using regression-based ML models they present varied levels of accuracy (Table 4). This can be attributed to the skewness of damage score data, which affects the predictive performance of the models due to the large variability [68,69]. To address this problem, sewer damage scores are aggregated into classes and the regression problem is converted into a classification problem. It is therefore recommended that future studies consider the classification-based approach to ML models for sewer condition assessment.

4. Conclusions

This paper investigated the potential application of ten state-of-the-art regression-based machine learning algorithms to model sewer conditions in the city of Ålesund, Norway. A dataset consisting of 1449 CCTV inspections and 20 physical and environmental factors was considered to construct and verify the sewer condition assessment models. Three feature-selection methods were applied to assess the importance of variables. The study revealed that:

- Age and material are the most sensitive factors affecting the sewer condition, while network type is the least contributor. Water utilities can refer to the age and material of sewer pipes as the priority factors when building predictive maintenance strategies;
- The performance of the ML models used was affected by the skewness and variability of the damage score data. Damage scores should be clustered into fewer condition classes to make the predictive results more convergent and improve prediction performance;
- The ETR model outperformed other ML models and this algorithm should be considered for modeling the sewer pipe condition in the study area;
- The results from this study can be critical for local water managers or engineers in assessing the condition of the entire sewer network. Based on the framework developed in this work, future sewer conditions can be predicted if the input factors are quantified. For instance, if rainfall/groundwater or population factors in the future are computed based on climate projection or annual population increase, respectively, the condition status of the sewer pipes in the corresponding time can be obtained. This is very important because local water organizations not only assess the current status of sewer pipes, but they also monitor the changes in the entire sewer network over time. It is a very useful tool supporting rehabilitation and maintenance strategies;
- Another advantage of our work is that all software and packages used in this work are open and free. Water engineers can easily add available observations and factors, reimplement, and reproduce the results under different scenarios;
- A limitation of this study is the lack of operational factors which were not available when undertaking the work. In the future, this limitation could be addressed by using operational factors and more sewer pipe inspections.

Author Contributions: Conceptualization, L.V.N. and R.S.; methodology, L.V.N. and R.S.; software, L.V.N.; validation, L.V.N.; investigation, L.V.N.; writing—original draft preparation, L.V.N.; writing—review and editing, L.V.N. and R.S.; supervision, R.S.; project administration, R.S.; funding acquisition, R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Smart Water Project, the grant number is 90392200 financed by Ålesund Municipality and Norwegian University of Science and Technology (NTNU), Norway.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: The authors would like to thank the Norwegian Mapping Authority, the Norwegian Climate Service Center, the Norwegian Water Resources and Energy Directorate, the Weather Atlas, the Mapping Authority, and the Norwegian Geological Survey for providing data for this research.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

Abbreviation	**Meaning**
RaM	Reactive Maintenance
PvM	Preventive Maintenance
PdM	Predictive Maintenance
CCTV	Closed-Circuit Television
ML	Machine Learning
GIS	Geographic Information System
DEM	Digital Elevation Model
NMA	the Norwegian Mapping Authority
NCSC	the Norwegian Climate Service Center
NGS	the Norwegian Geological Survey
NPRA	the Norwegian Public Roads Administration
IDW	Inverse Distance Weighting
PVC	Polyvinyl Chloride
PR	Pearson's Correlation Coefficient
OOB	Out-of-Bag
RMSE	Root Mean Squared Error
GPR	Gaussian Process Regression
KNN	K-Nearest Neighbor
DT	Decision Tree
CART	Classification and Regression Trees
RF	Random Forest
SVR	Support Vector Regression
MLP	Multi-layer Perceptron Neural Network
ETR	Extra Trees Regression
GB	Gradient Boosting
HGB	Histogram-based Gradient Boosting
R^2	coefficient of determination
MAE	Mean Absolute Error
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
l_{GPR}	length-scale
n_{GPR}	smoothness function
$K_{neighbor}$	the number of neighbors
ball_tree	ball tree nearest neighbors search algorithm
squared_error	mean squared error
$n_{feature}$	the number of features to choose the best subset
n_{tree}	the number of trees
rbf	Radial basis function
C	regularization parameter
γ	kernel width
relu	Rectified Linear Unit
lbfgs	Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm
n_{neuron}	the number of neurons in the hidden layer
$n_{iteration}$	the maximum number of iterations of the boosting process
n_{depth}	the maximum depth of each tree
$n_{boosting}$	the number of boosting iterations

Appendix A

Table A1. The weather stations used in this analysis.

Weather Station	Latitude	Longitude	Average Rainfall (mm)	Period
Brusdalsvann	62°27′59.8″	6°27′45.4″	157.0	01.1907–12.1972
Brusdalsvann II	62°27′55.4″	6°24′04.7″	152.1	01.1973–12.2014
Skodje	62°30′00.0″	6°42′01.4″	139.8	01.1961–12.1979
Ålesund	62°28′31.1″	6°09′04.0″	105.8	01.1895–12.1930
Ålesund II	62°28′25.3″	6°10′22.4″	95.5	01.1908–12.1954
Ålesund III	62°28′31.4″	6°12′06.1″	125.9	01.1955–12.2004
Ørskog	62°28′39.0″	6°49′00.1″	130.7	01.1896–12.2019
Hildre	62°36′05.8″	6°19′07.0″	125.5	01.1970–12.2018
Vigra	62°33′40.7″	6°06′40.7″	113.7	01.1959–12.2019

Table A2. Tuned hyperparameters for machine learning models.

Model	Range of Hyperparameters	Tuned Hyperparameters
GPR	- Kernel Function: *RationalQuadratic, RBF(), DotProduct(), Matern(), WhiteKernel(), ExpSineSquared()*	- Kernel Function: - *RationalQuadratic* - $l_{GPR} = 1.0$ - $\alpha_{GPR} = 1.0$
KNN	- $K_{neighbor} = 1, 2, \ldots, 99, 100$ - Weight function: *ball_tree, kd_tree, brute* - Metric: *manhattan, minkowski, euclidean* - Search *algorithm: uniform, distance*	- $K_{neighbor} = 90$ - Weight function: *ball_tree* - Metric: *manhattan* - Search algorithm: *uniform*
DTR	- Split criteria: *squared_error, friedman_mse, absolute_error* - $n_{feature} = 1, 2, \ldots, 19, 20$	- Split criteria: *squared_error* - $n_{feature} = 4$
RFR	- $n_{feature} = 1, 2, \ldots, 19, 20$ - $n_{tree} = 1, 2, \ldots, 99, 100$	- $n_{feature} = 2$ - $n_{tree} = 96$
SVR	- Kernel function: *rbf, linear, poly, sigmoid* - $C = 2^{-5}, 2^{-4}, \ldots, 2^{14}, 2^{15}$ - $\gamma = 2^{-15}, 2^{-14}, \ldots, 2^{4}, 2^{5}$	- Kernel function: *rbf* - $C = 2^{7}$ - $\gamma = 2^{-10}$
ETR	- $n_{feature} = 1, 2, \ldots, 19, 20$ - $n_{tree} = 1, 2, \ldots, 99, 100$	- $n_{feature} = 1$ - $n_{tree} = 84$
MLP	- Activation function: *relu, logistic, tanh* - Solver: *lbfgs, sgd, adam* - $n_{neuron} = 1, 2, \ldots, 199, 200$	- Activation function: *relu* - Solver: *lbfgs* - $n_{neuron} = 170$
AdaBoost	- $n_{boosting} = 10, 11, \ldots, 29, 30$ - Learning rate: $0.001, 0.002, \ldots, 0.0099$	- $n_{boosting} = 10$ - Learning rate: 0.0076
GBR	- $n_{estimator} = 1,\ 2, \ldots, 9, 10$ - Learning rate: $0.01, 0.02, \ldots, 1.09, 1.10$	- $n_{estimator} = 10$ - Learning rate: 0.16
HGBR	- $n_{iteration} = 1, 2, \ldots 29, 30$ - $n_{depth} = 1, 2, \ldots, 19, 20$	- $n_{iteration} = 5$ - $n_{depth} = 1$

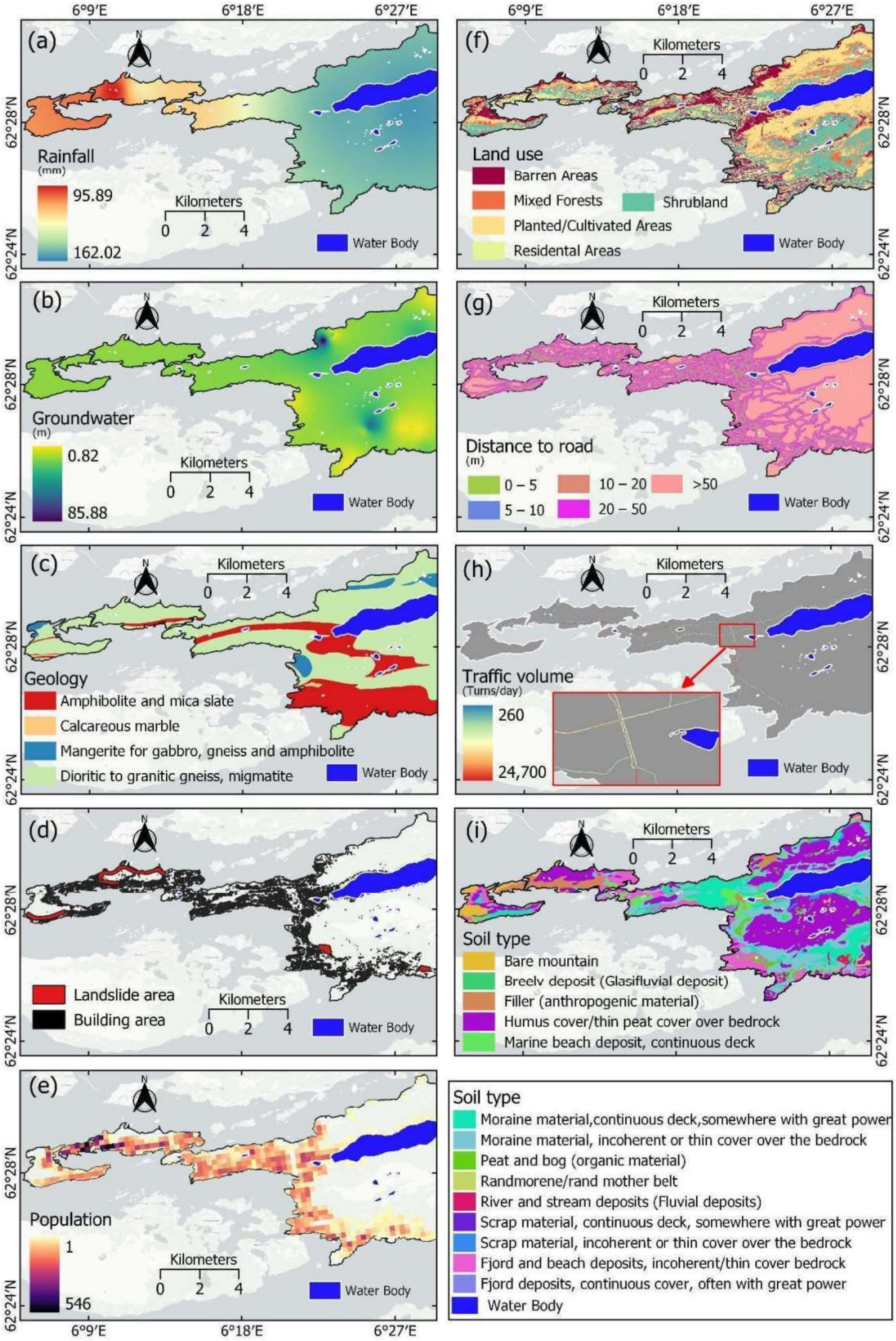

6°9′E
6°18′E
6°27′E
62°28′N
62°24′N
(a)
Rainfall
(mm)
95.89
162.02
Kilometers
0 2 4
Water Body
(b)
Groundwater
(m)
0.82
85.88
(c)
Geology
Amphibolite and mica slate
Calcareous marble
Mangerite for gabbro, gneiss and amphibolite
Dioritic to granitic gneiss, migmatite
(d)
Landslide area
Building area
(e)
Population
1
546
(f)
Land use
Barren Areas
Mixed Forests
Shrubland
Planted/Cultivated Areas
Residental Areas
(g)
Distance to road
(m)
0 – 5
5 – 10
10 – 20
20 – 50
>50
(h)
Traffic volume
(Turns/day)
260
24,700
(i)
Soil type
Bare mountain
Breelv deposit (Glasifluvial deposit)
Filler (anthropogenic material)
Humus cover/thin peat cover over bedrock
Marine beach deposit, continuous deck
Soil type
Moraine material,continuous deck,somewhere with great power
Moraine material, incoherent or thin cover over the bedrock
Peat and bog (organic material)
Randmorene/rand mother belt
River and stream deposits (Fluvial deposits)
Scrap material, continuous deck, somewhere with great power
Scrap material, incoherent or thin cover over the bedrock
Fjord and beach deposits, incoherent/thin cover bedrock
Fjord deposits, continuous cover, often with great power
Water Body

Figure A1. Maps of environmental factors: (**a**) Rainfall, (**b**) Groundwater, (**c**) Geology, (**d**) Landslide and building area, (**e**) Population, (**f**) Land use, (**g**) Distance to road, (**h**) Traffic volume, and (**i**) Soil type.

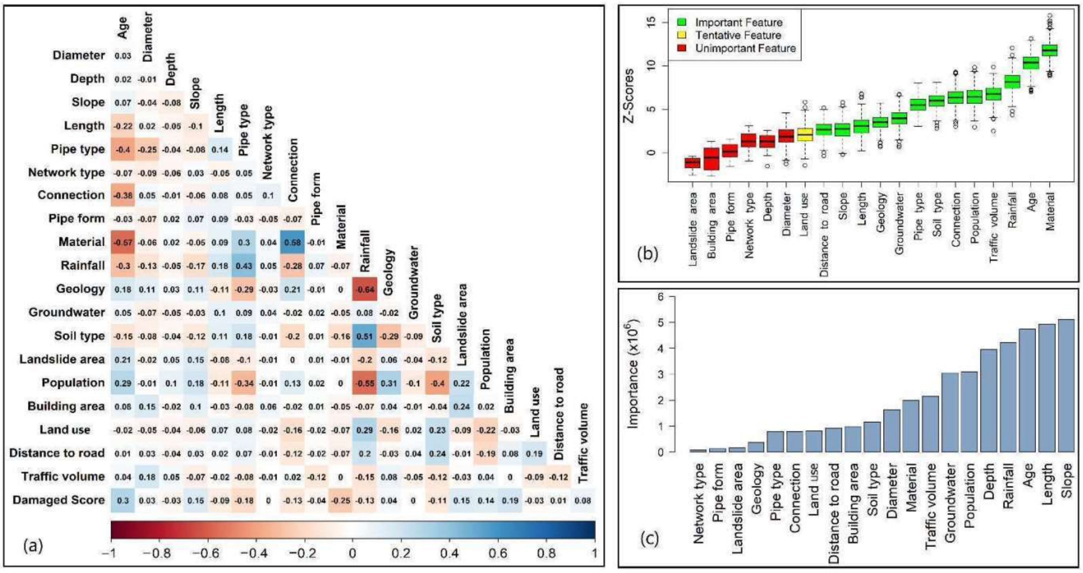


Figure A2. The features' importance: (**a**) Pearson's correlation, (**b**) Boruta, and (**c**) Random Forest.

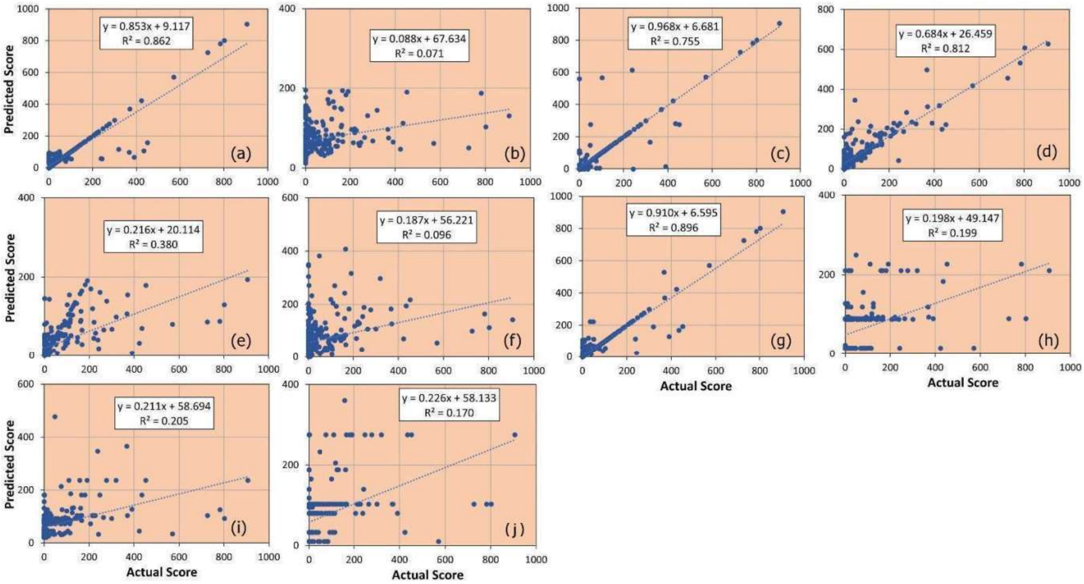


Figure A3. R^2 of the constructed machine learning models using the test dataset: (**a**) Gaussian Process Regression, (**b**) K-Nearest Neighbor, (**c**) Classification and Regression Trees, (**d**) Random Forest, (**e**) Support Vector Regression, (**f**) Multi-layer Perceptron Neural Network, (**g**) Extra Trees Regression, (**h**) AdaBoost, (**i**) Gradient Boosting, and (**j**) Histogram-Based Gradient Boosting.

References

1. Ana, E.V.; Bauwens, W. Modeling the structural deterioration of urban drainage pipes: The state-of-the-art in statistical methods. *Urban Water J.* **2010**, *7*, 47–59. [CrossRef]
2. Farkas, K.; Hillary, L.S.; Malham, S.K.; McDonald, J.E.; Jones, D.L. Wastewater and public health: The potential of wastewater surveillance for monitoring COVID-19. *Curr. Opin. Environ. Sci. Health* **2020**, *17*, 14–20. [CrossRef] [PubMed]
3. Sun, S.A.; Djordjević, S.; Khu, S.-T. A general framework for flood risk-based storm sewer network design. *Urban Water J.* **2011**, *8*, 13–27. [CrossRef]
4. Ana, E.; Bauwens, W.; Pessemier, M.; Thoeye, C.; Smolders, S.; Boonen, I.; De Gueldre, G. An investigation of the factors influencing sewer structural deterioration. *Urban Water J.* **2009**, *6*, 303–312. [CrossRef]
5. Anand, U.; Li, X.; Sunita, K.; Lokhandwala, S.; Gautam, P.; Suresh, S.; Sarma, H.; Vellingiri, B.; Dey, A.; Bontempi, E.; et al. SARS-CoV-2 and other pathogens in municipal wastewater, landfill leachate, and solid waste: A review about virus surveillance, infectivity, and inactivation. *Environ. Res.* **2022**, *203*, 111839. [CrossRef]
6. Yin, X.; Chen, Y.; Bouferguene, A.; Al-Hussein, M. Data-driven bi-level sewer pipe deterioration model: Design and analysis. *Autom. Constr.* **2020**, *116*, 103181. [CrossRef]
7. Beheshti, M.; Sægrov, S.; Ugarelli, R. Infiltration/inflow assessment and detection in urban sewer system. *Vannforeningen* **2015**, *1*, 24–34.
8. Susto, G.A.; Schirru, A.; Pampuri, S.; McLoone, S.; Beghi, A. Machine Learning for Predictive Maintenance: A Multiple Classifier Approach. *IEEE Trans. Ind. Inform.* **2015**, *11*, 812–820. [CrossRef]
9. Chughtai, F.; Zayed, T. Sewer pipeline operational condition prediction using multiple regression. In *Pipelines 2007: Advances and Experiences with Trenchless Pipeline Projects*; ASCE: Fairfax County, VA, USA, 2007; pp. 1–11.
10. Hawari, A.; Alkadour, F.; Elmasry, M.; Zayed, T. A state of the art review on condition assessment models developed for sewer pipelines. *Eng. Appl. Artif. Intell.* **2020**, *93*, 103721. [CrossRef]
11. Heydarzadeh, R.; Tabesh, M.; Scholz, M. Dissolved oxygen determination in sewers using flow hydraulic parameters as part of a physical-biological simulation model. *J. Hydroinforma.* **2021**, *24*, 1–15. [CrossRef]
12. Hadzilacos, T.; Kalles, D.; Preston, N.; Melbourne, P.; Camarinopoulos, L.; Eimermacher, M.; Kallidromitis, V.; Frondistou-Yannas, S.; Saegrov, S. UtilNets: A water mains rehabilitation decision-support system. *Comput. Environ. Urban Syst.* **2000**, *24*, 215–232. [CrossRef]
13. Tscheikner-Gratl, F.; Caradot, N.; Cherqui, F.; Leitão, J.P.; Ahmadi, M.; Langeveld, J.G.; Le Gat, Y.; Scholten, L.; Roghani, B.; Rodríguez, J.P.; et al. Sewer asset management—State of the art and research needs. *Urban Water J.* **2019**, *16*, 662–675. [CrossRef]
14. Fan, X.; Wang, X.; Zhang, X.; Asce Xiong Yu, P.E.F. Machine learning based water pipe failure prediction: The effects of engineering, geology, climate and socio-economic factors. *Reliab. Eng. Syst. Saf.* **2022**, *219*, 108185. [CrossRef]
15. Uddin, S.; Khan, A.; Hossain, M.E.; Moni, M.A. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med. Inform. Decis. Mak.* **2019**, *19*, 281. [CrossRef]
16. Hawari, A.; Firas, A.; Elmasry, M.; Zayed, T. Simulation-Based Condition Assessment Model for Sewer Pipelines. *J. Perform. Constr. Facil.* **2016**, *31*, 04016066. [CrossRef]
17. Chandrashekar, G.; Sahin, F. A survey on feature selection methods. *Comput. Electr. Eng.* **2014**, *40*, 16–28. [CrossRef]
18. Kuhn, M.; Johnson, K. An Introduction to Feature Selection. In *Applied Predictive Modeling*; Springer: New York, NY, USA, 2013; pp. 487–519. [CrossRef]
19. Caradot, N.; Riechel, M.; Rouault, P.; Caradot, A.; Lengemann, N.; Eckert, E.; Ringe, A.; Clemens, F.; Cherqui, F. The influence of condition assessment uncertainties on sewer deterioration modelling. *Struct. Infrastruct. Eng.* **2020**, *16*, 287–296. [CrossRef]
20. Bairaktaris, D.; Delis, V.; Emmanouilidis, C.; Frondistou-Yannas, S.; Gratsias, K.; Kallidromitis, V.; Rerras, N. Decision-Support System for the Rehabilitation of Deteriorating Sewers. *J. Perform. Constr. Facil.* **2007**, *21*, 240–248. [CrossRef]
21. Hansen, B.D.; Jensen, D.G.; Rasmussen, S.H.; Tamouk, J.; Uggerby, M.; Moeslund, T.B. General Sewer Deterioration Model Using Random Forest. In Proceedings of the 2019 IEEE Symposium Series on Computational Intelligence (SSCI), Xiamen, China, 6–9 December 2019; pp. 834–841.
22. Nusinovici, S.; Tham, Y.C.; Chak Yan, M.Y.; Wei Ting, D.S.; Li, J.; Sabanayagam, C.; Wong, T.Y.; Cheng, C.-Y. Logistic regression was as good as machine learning for predicting major chronic diseases. *J. Clin. Epidemiol.* **2020**, *122*, 56–69. [CrossRef]
23. Song, X.; Liu, X.; Liu, F.; Wang, C. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. *Int. J. Med. Inform.* **2021**, *151*, 104484. [CrossRef]
24. Population, C. Municipality in Møre og Romsdal (Norway). Available online: https://www.citypopulation.de/en/norway/admin/m%C3%B8re_og_romsdal/1507__%C3%A5lesund/ (accessed on 10 February 2022).
25. Climate, D. Ålesund Climate: Average Temperature, Weather by Month, Ålesund Water Temperature—Climate-Data.org. Available online: https://en.climate-data.org/europe/norway/m%C3%B8re-og-romsdal/alesund-9937/ (accessed on 20 April 2022).
26. Kvitsjøen, J.; Paus, K.; Bjerkholt, J.T.; Fergus, T.; Lindholm, O. Intensifying rehabilitation of combined sewer systems using trenchless technology in combination with low impact development and green infrastructure. *Water Sci. Technol.* **2021**, *83*, 2947–2962. [CrossRef] [PubMed]
27. Hanssen-Bauer, I.; Drange, H.; Førland, E.; Roald, L.; Børsheim, K.; Hisdal, H.; Lawrence, D.; Nesje, A.; Sandven, S.; Sorteberg, A. Climate in Norway 2100—A Knowledge Base for Climate Adaptation. In *Background information to NOU Climate Adaptation (In Norwegian: Klima i Norge 2100. Bakgrunnsmateriale til NOU Klimatilplassing)*; Norsk Klimasenter: Oslo, Norway, 2017.

28. Laakso, T.; Kokkonen, T.; Mellin, I.; Vahala, R. Sewer Condition Prediction and Analysis of Explanatory Factors. *Water* **2018**, *10*, 1239. [CrossRef]
29. Belief, E. GIS based spatial modeling to mapping and estimation relative risk of different diseases using inverse distance weighting (IDW) interpolation algorithm and evidential belief function (EBF) (Case study: Minor Part of Kirkuk City, Iraq). *Int. J. Eng. Technol.* **2018**, *7*, 185–191.
30. Sánchez-Espinosa, A.; Schröder, C. Land use and land cover mapping in wetlands one step closer to the ground: Sentinel-2 versus landsat 8. *J. Environ. Manag.* **2019**, *247*, 484–498. [CrossRef]
31. Schober, P.; Boer, C.; Schwarte, L.A. Correlation coefficients: Appropriate use and interpretation. *Anesth. Analg.* **2018**, *126*, 1763–1768. [CrossRef]
32. Adler, J.; Parmryd, I. Quantifying colocalization by correlation: The Pearson correlation coefficient is superior to the Mander's overlap coefficient. *Cytom. Part A* **2010**, *77A*, 733–742. [CrossRef]
33. Masrur Ahmed, A.A.; Deo, R.C.; Feng, Q.; Ghahramani, A.; Raj, N.; Yin, Z.; Yang, L. Deep learning hybrid model with Boruta-Random forest optimiser algorithm for streamflow forecasting with climate mode indices, rainfall, and periodicity. *J. Hydrol.* **2021**, *599*, 126350. [CrossRef]
34. Nanda, M.A.; Seminar, K.B.; Maddu, A.; Nandika, D. Identifying relevant features of termite signals applied in termite detection system. *Ecol. Inform.* **2021**, *64*, 101391. [CrossRef]
35. Liu, H.; Zhou, M.; Liu, Q. An embedded feature selection method for imbalanced data classification. *IEEE/CAA J. Autom. Sin.* **2019**, *6*, 703–715. [CrossRef]
36. Dewi, C.; Chen, R.-C. Random forest and support vector machine on features selection for regression analysis. *Int. J. Innov. Comput. Inf. Control* **2019**, *15*, 2027–2037.
37. Gibson, N.P.; Aigrain, S.; Roberts, S.; Evans, T.M.; Osborne, M.; Pont, F. A Gaussian process framework for modelling instrumental systematics: Application to transmission spectroscopy. *Mon. Not. R. Astron. Soc.* **2012**, *419*, 2683–2694. [CrossRef]
38. Meng, L.; Zhang, J. Process Design of Laser Powder Bed Fusion of Stainless Steel Using a Gaussian Process-Based Machine Learning Model. *JOM* **2020**, *72*, 420–428. [CrossRef]
39. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
40. Pulido, E.S.; Arboleda, C.V.; Rodríguez Sánchez, J.P. Study of the spatiotemporal correlation between sediment-related blockage events in the sewer system in Bogotá (Colombia). *Water Sci. Technol.* **2019**, *79*, 1727–1738. [CrossRef]
41. Zhang, J.; Li, B.; Fan, X.; Wang, Y.; Chen, F. Sewer Corrosion Prediction for Sewer Network Sustainability. In *Humanity Driven AI: Productivity, Well-being, Sustainability and Partnership*; Chen, F., Zhou, J., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 181–194. [CrossRef]
42. Lall, U.; Sharma, A. A Nearest Neighbor Bootstrap For Resampling Hydrologic Time Series. *Water Resour. Res.* **1996**, *32*, 679–693. [CrossRef]
43. Yao, Z.; Ruzzo, W.L. A Regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data. *BMC Bioinform.* **2006**, *7*, S11. [CrossRef]
44. Kohli, S.; Godwin, G.T.; Urolagin, S. *Sales Prediction Using Linear and KNN Regression*; Springer Nature Singapore Pte Ltd.: Singapore, 2020; pp. 321–329.
45. Syachrani, S.; Jeong, H.S.D.; Chung, C.S. S. Decision Tree–Based Deterioration Model for Buried Wastewater Pipelines. *J. Perform. Constr. Facil.* **2013**, *27*, 633–645. [CrossRef]
46. Larose, D.T.; Larose, C.D. *Discovering Knowledge in Data: An Introduction to Data Mining*; John Wiley & Sons: New York, NY, USA, 2014; Volume 4.
47. Kumar, S.S.; Shaikh, T. Empirical Evaluation of the Performance of Feature Selection Approaches on Random Forest. In Proceedings of the 2017 International Conference on Computer and Applications (ICCA), Doha, Qatar, 6–7 September 2017; pp. 227–231.
48. Li, Y.; Zou, C.; Berecibar, M.; Nanini-Maury, E.; Chan, J.C.W.; van den Bossche, P.; Van Mierlo, J.; Omar, N. Random forest regression for online capacity estimation of lithium-ion batteries. *Appl. Energy* **2018**, *232*, 197–210. [CrossRef]
49. Trafalis, T.B.; Ince, H. Support vector machine for regression and applications to financial forecasting. In Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium, Como, Italy, 27 July 2000; Volume 346, pp. 348–353.
50. Wauters, M.; Vanhoucke, M. Support Vector Machine Regression for project control forecasting. *Autom. Constr.* **2014**, *47*, 92–106. [CrossRef]
51. Smola, A.J.; Schölkopf, B. A tutorial on support vector regression. *Stat. Comput.* **2004**, *14*, 199–222. [CrossRef]
52. Orhan, U.; Hekim, M.; Ozer, M. EEG signals classification using the K-means clustering and a multilayer perceptron neural network model. *Expert Syst. Appl.* **2011**, *38*, 13475–13481. [CrossRef]
53. Geurts, P.; Ernst, D.; Wehenkel, L. Extremely randomized trees. *Mach. Learn.* **2006**, *63*, 3–42. [CrossRef]
54. Hong, H.; Liu, J.; Bui, D.T.; Pradhan, B.; Acharya, T.D.; Pham, B.T.; Zhu, A.X.; Chen, W.; Ahmad, B.B. Landslide susceptibility mapping using J48 Decision Tree with AdaBoost, Bagging and Rotation Forest ensembles in the Guangchang area (China). *CATENA* **2018**, *163*, 399–413. [CrossRef]
55. Kégl, B. The return of AdaBoost. MH: Multi-class Hamming trees. *arXiv* **2013**, arXiv:1312.6086.

56. Bentéjac, C.; Csörgő, A.; Martínez-Muñoz, G. A comparative analysis of gradient boosting algorithms. *Artif. Intell. Rev.* **2021**, *54*, 1937–1967. [CrossRef]
57. Friedman, J.H. Greedy Function Approximation: A Gradient Boosting Machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
58. Ayyadevara, V.K. Gradient Boosting Machine. In *Pro Machine Learning Algorithms: A Hands-On Approach to Implementing Algorithms in Python and R*; Apress: Berkeley, CA, USA, 2018; pp. 117–134. [CrossRef]
59. Aljamaan, H.; Alazba, A. Software defect prediction using tree-based ensembles. In Proceedings of the 16th ACM international conference on predictive models and data analytics in software engineering, Virtual, 8–9 November 2020; pp. 1–10.
60. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. Lightgbm: A highly efficient gradient boosting decision tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 9.
61. Kabir, G.; Balek, N.B.C.; Tesfamariam, S. Sewer Structural Condition Prediction Integrating Bayesian Model Averaging with Logistic Regression. *J. Perform. Constr. Facil.* **2018**, *32*, 04018019. [CrossRef]
62. Vazquezl, M.Y.L.; Peñafiel, L.A.B.; Muñoz, S.X.S.; Martinez, M.A.Q. *A Framework for Selecting Machine Learning Models Using TOPSIS*; Springer Nature Switzerland AG: Cham, Switzerland, 2020; pp. 119–126.
63. Behzadian, M.; Khanmohammadi Otaghsara, S.; Yazdani, M.; Ignatius, J. A state-of the-art survey of TOPSIS applications. *Expert Syst. Appl.* **2012**, *39*, 13051–13069. [CrossRef]
64. Chakraborty, S. TOPSIS and Modified TOPSIS: A comparative analysis. *Decis. Anal. J.* **2022**, *2*, 100021. [CrossRef]
65. Ihaka, R.; Gentleman, R. R: A Language for Data Analysis and Graphics. *J. Comput. Graph. Stat.* **1996**, *5*, 299–314. [CrossRef]
66. Taylor, R. Interpretation of the correlation coefficient: A basic review. *J. Diagn. Med. Sonogr.* **1990**, *6*, 35–39. [CrossRef]
67. Bui, K.-T.T.; Torres, J.F.; Gutiérrez-Avilés, D.; Nhu, V.-H.; Bui, D.T.; Martínez-Álvarez, F. Deformation forecasting of a hydropower dam by hybridizing a long short-term memory deep learning network with the coronavirus optimization algorithm. *Comput.—Aided Civ. Infrastruct. Eng.* **2022**, *37*, 1368–1386. [CrossRef]
68. Bellon-Maurel, V.; Fernandez-Ahumada, E.; Palagos, B.; Roger, J.-M.; McBratney, A. Critical review of chemometric indicators commonly used for assessing the quality of the prediction of soil attributes by NIR spectroscopy. *TrAC Trends Anal. Chem.* **2010**, *29*, 1073–1081. [CrossRef]
69. Krawczyk, B. Learning from imbalanced data: Open challenges and future directions. *Prog. Artif. Intell.* **2016**, *5*, 221–232. [CrossRef]

Paper II

Lam Van Nguyen, Dieu Tien Bui, and Seidu Razak (2022). Comparison of Machine Learning Techniques for Condition Assessment of Sewer Network. *IEEE Access,* 10, 124238-124258. https://doi.org/10.1109/ACCESS.2022.3222823.

IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received 10 October 2022, accepted 9 November 2022, date of publication 17 November 2022,
date of current version 30 November 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3222823

APPLIED RESEARCH

Comparison of Machine Learning Techniques for Condition Assessment of Sewer Network

LAM VAN NGUYEN[1,2], DIEU TIEN BUI[3], AND RAZAK SEIDU[1]

[1]Smart Water and Environmental Engineering Group, Department of Ocean Operations and Civil Engineering, Faculty of Engineering, Norwegian University of Science and Technology, 6025 Ålesund, Norway
[2]Department of Geodesy, Faculty of Geomatics and Land Administration, Hanoi University of Mining and Geology, Hanoi 100000, Vietnam
[3]GIS Group, Department of Business and IT, University of South-Eastern Norway, 3800 Bø i Telemark, Norway

Corresponding author: Lam Van Nguyen (lam.v.nguyen@ntnu.no)

This work was supported by the Ålesund Municipality and the Norwegian University of Science and Technology (NTNU), Norway, through the Smart Water Project under Grant 90392200.

ABSTRACT Assessment of sewer condition is one of the critical steps in asset management and support investment decisions; therefore, condition assessment models with high accuracy are important that can help utility managers and other authorities correctly assess the current condition of the sewage network and effectively initiate maintenance and rehabilitation strategies. The main objective of this research is to assess the potential application of machine learning (ML) algorithms for predicting the condition of sewer pipes with a case study in Ålesund city, Norway. Nine physical factors (i.e., age, diameter, depth, slope, length, pipe type, material, pipe form, and connection type) and ten environmental factors (i.e., rainfall, geology, landslide area, building area, population, land cover, groundwater, traffic volume, distance to road, and soil type) were used to assess the sewer conditions employing seventeen ML models. After processing the sewer inspections, 1159 of 1449 individual pipelines were used to train the sewer condition model. The performance of ML models was validated using the 290 remaining inspected sewer pipes. The area under the Receiver Operating Characteristic (AUC-ROC) curve and accuracy (ACC) showed that the Random Forest (AUC-ROC = 77.6% and ACC = 78.3%) is a sensitive model for predicting the condition of sewer pipes in the study area. Based on the Random Forest model, maps of predicted conditions of sewers were generated that may be useful for utilities and water managers to establish future sewer system maintenance strategies.

INDEX TERMS Geographic information system, machine learning, predictive maintenance, sewer network, sewer condition assessment.

I. INTRODUCTION

The collection, transport, treatment, and discharge of stormwater, and wastewater are the main roles of sewage networks in urban areas [1]. Stormwater and wastewater collection systems play a critical role in minimizing the negative effects of floods during heavy rainfall events and protecting the environment from contamination [2]. However, due to intrinsic and extrinsic factors, sewer networks are subjected to deterioration, breakages, and collapse during their lifespan with dire consequences for infrastructure, the environment, and public health [3].

Studies have shown an increasing rate of breakages and collapse of sewer networks as a corollary of limited rehabilitation investments, climate change, and rapid urbanization [2]. By 2040, investment needs for water infrastructure in Norway have been previously estimated at €28 billion [4], and this amount was recently raised to €32 billion [5]. Consequently, maintenance strategies are being implemented to increase the life cycle of the sewer network and reduce the expenditure on replacement and rehabilitation [6]. To achieve this, condition models for the prediction of sewer conditions can be valuable tools to support the maintenance, rehabilitation and to investment decision strategies for the sewer network [7].

For condition modeling of sewer pipes, a determination of contributing factors of sewer conditions is important.

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Liu.

According to a review, factors influencing sewer conditions can be categorized into: *physical* (e.g., age, diameter, and length), *environmental* (e.g., rainfall, soil type, and groundwater), and *operational* factors (e.g., sediment level, flow rate, and infiltration) [8]. Besides, the roles of the aforementioned factors on the model prediction reliability are not equal; therefore, assessing the significance of the contributing factors is a pivotal task that may enhance the predictive ability of the models [9].

Geographic Information System (GIS) can play an important role in data management, modeling, and visualization in condition assessment. GIS can be applied to store, manage, calculate, visualize, and analyze spatial and non-spatial information and data on the contributing factors of sewer conditions in different layers [10]. Based on the GIS database, prediction models of the future condition of sewer pipes can be autonomously constructed and updated.

Condition models can be classified into *physical*, *statistical*, and *machine learning* [11], [12], [13], [14]. In the physical models, parameters related to conditions of the sewer pipes (e.g., material, diameter, type of effluent, etc.,) are employed to fit mathematical equations to the sewer's status [15]. Therefore, these models are effective for sewer network analysis during the construction period and initial operational phases. However, the scarcity of data needed for the simulation of deterioration mechanisms is one of the limitations of these models [13].

Sewer deterioration is a complex process affected by many factors, therefore statistical models are likely to have more advantages compared to the physical models in terms of calculating speed and straightforward function form when the monitoring time-series data is long enough [16]. Successful applications of statistical models in sewer deterioration assessment have been reported in the literature [17], [18], [19], [20]. These models are based on some assumptions that need to be satisfied to achieve highly accurate performance [13]. However, these assumptions are generally difficult to achieve with the deterioration process. For example, the distance between consecutive conditions is constant or the sewer status in each condition should be a normal distribution [8]. According to Zamanian, et al. [21], sewer deterioration is a non-linear process, and it is difficult for statistical models to predict this process with high accuracy.

Machine learning (ML) models can capture the linear and non-linear relationships between input factors and sewer condition state, even if these relationships are unclear or when data is incomplete [22]. Additionally, these models can effectively deal with different types of inputs and outputs, including numeric, nominal, or categorical [23]; therefore, they are applied in various studies involving sewer condition prediction [14], [24]. However, the accuracy of deterioration prediction models using ML algorithms needs to be improved by increasing the number of input factors and inspections or using an adequately distributed dataset [6].

Although different ML algorithms are used to model the condition of sewer pipes, their accuracies are dissimilar due to different characteristics in the study area, data quality, variation, and randomness between studies and used algorithms [25]. As a result, no ML model is the best for modeling sewer conditions in all areas. Besides, a comprehensive comparison between different types of ML models for modeling the condition of sewer pipelines is still missing. This study is an attempt to partially fill this gap in the literature by exploring and verifying the potential application of ML algorithms for sewer condition assessment. The significance of input factors was briefly analyzed to provide helpful information for water engineers/managers in prioritizing significant factors of the sewer condition in maintenance strategies in Ålesund city, Norway.

II. THE STUDY AREA AND GIS DATABASE

A. DESCRIPTION OF THE STUDY AREA

The sewer network of Ålesund, a coastal city located in the west of Norway, was used in this study. The city has an area of approximately 607.3 km^2 and lies between latitudes of 6°05'08'' N and 6°40'56'' N, and between longitudes of 62°25'07'' E and 62°30'37'' E (FIGURE 1).

Ålesund city is in an area heavily influenced by ocean currents with cold, rainy winters and cool summers. The variation in temperatures throughout the year is 13.6 °C with average temperatures of the coldest month (February) and the warmest month (August) being −0.6 °C and 13.0 °C, respectively [26]. The city is also located in a high rainfall density region with an average rainfall of 2100 mm per year. The average rainfall in the driest month (May) and the wettest month (December) are 104 mm and 230 mm, respectively. Along with the general trend of climate change, the weather in the city is affected by unavoidable fluctuations in temperature, precipitation, and extreme weather events that put pressure on the sewer network system [27], [28].

B. GENERAL DESCRIPTION OF DATA USED

Based on the literature [29] and data availability, a total of ten physical factors (i.e., age, diameter, depth, slope, length, pipe type, material, network type, pipe form, and connection type) and ten environmental factors (i.e., rainfall, geology, landslide area, building area, population, land cover, groundwater, traffic volume, distance to road, and soil type) were considered in this study. It should be stated that operational factors including, but not limited to, flow rate, blockages, infiltration, or inflows were not considered because of data unavailability at the time of this study.

1) PHYSICAL FACTORS

Physical factors of sewer pipe networks are elements that relate to the pipeline's physical characteristics and components. These factors are considered indispensable for pipe deterioration rate or remaining useful life [30]. Data on the physical characteristics of pipes are generally recorded,

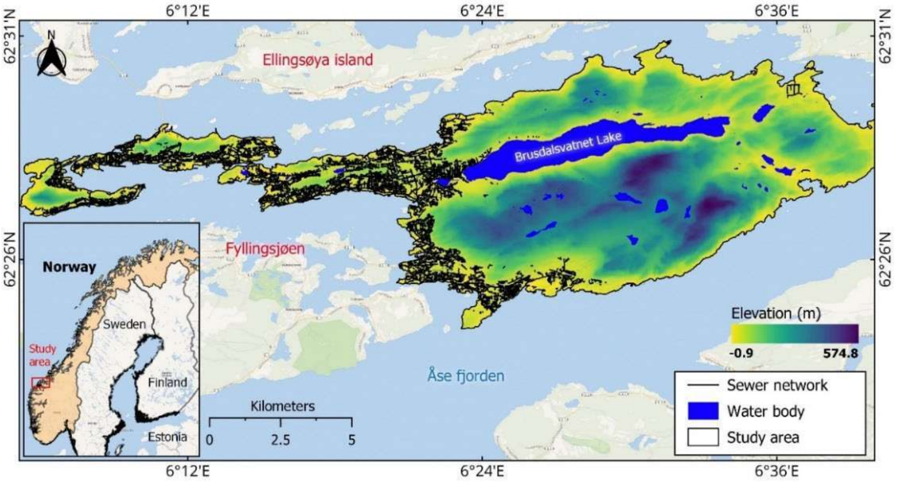


FIGURE 1. The sewer network in the study area.

updated, and managed by water utilities and/or relevant agencies. In this study, the physical factors were provided in the tabular datasets by the Department of Water and Sewage of Ålesund in 2021. Eight physical factors (i.e., age, diameter, length, pipe type, material, network type, pipe form, and connection type) are available in the dataset, the remaining factors (i.e., depth and slope) are obtained from Digital Elevation Model (DEM) with pixels of 5 m × 5 m using a GIS tool. An overview of the physical factors used in this study is shown in FIGURE 2. The number on top of the bar chart represents the number of sewer pipes corresponding to pipe type.

The age of sewer pipes was considered one of the most significant factors affecting the sewer condition process [8]. This factor immediately affects the pipe deterioration after the sewer is installed, and the aging speed is quicker during the operation [31]. In this study, the age of the sewer pipes was calculated as the difference between the installation year and the inspection year. The oldest sewer pipelines were installed in the 1900s and the newest pipes were replaced/set up in 2020 (FIGURE 2a).

The influence of material on the sewer condition process of sewer pipes is well established in previous studies [3], [6], [32]. For example, although concrete pipes are significantly resistant to abrasion, they are vulnerable to the corrosive action of hydrogen sulfide [33]. The materials of the sewer pipes are shown in FIGURE 2b.

Pipe diameters affect the deterioration process. For instance, larger pipes are less affected by deterioration compared to smaller ones [17]. Detailed information on sewer pipes according to their diameter is shown in FIGURE 2c. Pipes in shallow depths are more vulnerable than the deeper ones because of stresses from surface load, road traffic, illegal connection, tree root intrusion, or road maintenance/construction activities [34]. The depths of the sewer pipes were unavailable in the database and were therefore computed as the distance from the ground surface to the mid-point of the pipes (FIGURE 2d). The height of the ground surface was interpolated from the DEM.

The sewer pipe slope directly relates to water flow that mainly causes corrosion, sediment deposition, and clogging in the sewer pipes. For example, flat concrete pipes are more vulnerable due to hydrogen sulfide gas emissions because wastewater in these pipes cannot drain speedily [32]. The information on pipe slope was not available in the dataset and was computed as the difference between the inverted elevation of the start and end manholes using the GIS tool (FIGURE 2e). The start and end manholes were classified based on the flow direction of each pipe. When water flows from the start point to the endpoint, the slope value is positive when the start point is higher than the endpoint and vice versa.

The sewer network in the study area consists of three different types: wastewater, stormwater, and combined pipes. The effect of pipe type on pipe condition has been established in previous studies. For instance, combined sewers are more likely to be deteriorated than sanitary pipes due to high potential infiltration and exfiltration during rainfall events [35]. Pipe length was shown as one of the factors affecting sewer conditions where pipes with long lengths had a higher probability of failure than shorter ones, and the failures often occurred at the connection positions [36]. Therefore, connection type was considered a factor for constructing the sewer condition model in this study. Some types of sewer connections are shown in FIGURE 2f. The sewer network with different forms and types can deteriorate differently. For example, clay pipes with circular shapes were easily prone to fractures [32]. Therefore, different pipe forms (FIGURE 2g)

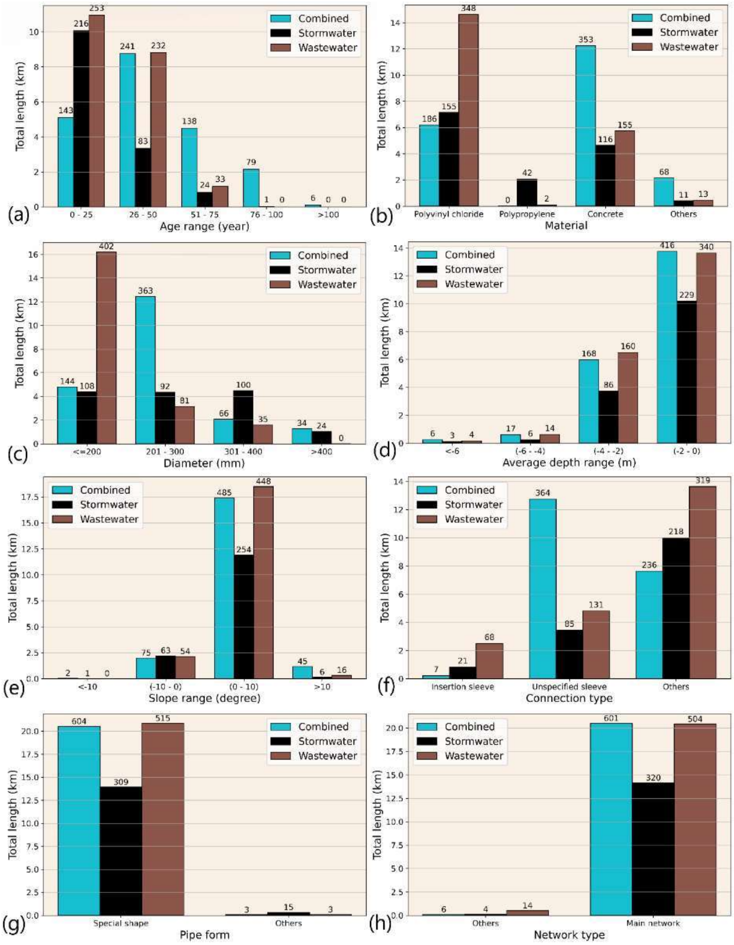


FIGURE 2. The physical characteristics of the sewage network.

and network types (FIGURE 2h) were considered in this study.

The sewer network in the study area contains about 31293 wastewater, stormwater, and combined pipelines with a total length of 703.0 km. After combining with the inspected data, a total of 1449 individual pipelines were used to model the sewer condition status. A summary of statistical indexes of physical variables is represented in TABLE 1.

TABLE 1. Summary of physical variables.

Physical variables	Type	Min	Max	Average	Std
Age (year)	Numeric	1.0	104.0	34.4	25.3
Diameter (mm)	Numeric	110.0	1000.0	248.4	98.6
Depth (m)	Numeric	-7.8	-0.1	-1.8	1.2
Length (m)	Numeric	1.0	177.5	38.6	21.3
Slope (°)	Numeric	-17.4	+34.6	+2.7	4.4
Pipe type	Categorical	-	-	-	-
Network type	Categorical	-	-	-	-
Pipe form	Categorical	-	-	-	-
Connection	Categorical	-	-	-	-
Material	Categorical	-	-	-	-

2) ENVIRONMENTAL FACTORS

Environmental factors used in this study are mainly extrinsic elements that relate to the relative geo-location of the sewers. The data have been collected from many sources with different formats and spatial resolutions (TABLE 2). It is worth noting that TABLE 2 shows the original spatial resolution of the environmental factors, and these data were processed to transfer them into the same coordinate system, format, and spatial resolution. In this study, post-processed data were re-sampled to the same spatial resolution (5 m × 5 m) and transformed into a grid spatial database before running ML models.

Rainfall results in rising groundwater which leads sewer pipes to deteriorate more quickly [35]. In this study, rainfall

TABLE 2. Summary of the environmental factors used in this analysis.

Data	Spatial resolution	GIS data type	Assess link
Rainfall	-	Point*	https://klimaservicesenter.no
Geology	1:50000	Polygon**	https://www.kartverket.no/en
Landslide area	1:5000	Polygon**	https://www.kartverket.no/en
Population	250 m × 250 m	GRID**	https://www.kartverket.no/en
Land cover	10 m × 10 m	GRID***	https://scihub.copernicus.eu
Building area	1:5000	Polygon**	https://www.kartverket.no/en
Groundwater	-	Point****	https://www.ngu.no
Traffic volume	5 m × 5 m	GRID*****	https://www.vegvesen.no/en
Distance to road	5 m × 5 m	GRID**	https://www.kartverket.no/en
Soil type	1:50000	Polygon**	https://www.kartverket.no/en

Data source: (*): the Norwegian Climate Service Center;(**): the Norwegian Mapping Authority; (***): Copernicus Open Access Hub; (****): the Norwegian Geological Survey; (*****): the Norwegian Public Roads Administration.

TABLE 3. The hydrological stations used for rainfall interpolation.

Weather station name	Latitude (°)	Longitude (°)	Avg. rainfall (mm)	Period
Brusdalsvann	62.4666	6.4626	157.0	01.1907 - 12.1972
Brusdalsvann II	62.4654	6.4013	152.1	01.1973 - 12.2014
Skodje	62.5000	6.7004	139.8	01.1961 - 12.1979
Ålesund	62.4753	6.1511	105.8	01.1895 - 12.1930
Ålesund II	62.4737	6.1729	95.5	01.1908 - 12.1954
Ålesund III	62.4754	6.2017	125.9	01.1955 - 12.2004
Ørskog	62.4775	6.8167	130.7	01.1896 - 12.2019
Hildre	62.6016	6.3186	125.5	01.1970 - 12.2018
Vigra	62.5613	6.1113	113.7	01.1959 - 12.2019

data were obtained from annual average rainfall over several years at nine weather stations near the study area. Detailed information about these stations is shown in TABLE 3. A rainfall map was generated using data at the aforementioned stations, and the Inverse Distance Weighting (IDW) method, which is the most common spatial interpolation method [37], was used to interpolate rainfall values in the study area (FIGURE 3a).

The geological characteristics around a sewer pipe can affect its condition processes. For instance, it has been shown that changes in geological structures affect infiltration and groundwater in coastal urban areas, resulting in sewer deterioration [38]. Additionally, hydraulic conductivity in different geological types can affect sewer deterioration differently [39]. FIGURE 3b shows the geological map used as input in sewer condition assessment.

Landslide has been implicated in sewer network because of failures caused by road subsidence [40]. Similarly, sewer pipes under building areas are more vulnerable to deterioration than those found in non-built areas [8]. In this study, landslide and building areas were considered as input factors for constructing sewer condition models (FIGURE 3c).

Population density is considered a critical factor for sewer deterioration. For example, a large population may lead to a huge volume of wastewater discharge into the wastewater collection network, resulting in the deterioration of the system [21]. In this study, a population map was prepared based on the statistical data received from the Norwegian Mapping Authority (FIGURE 3d). Land cover affects soil infiltration rate, evapotranspiration, or surface runoff and has been considered a variable in water quality change that has a strong correlation with the current condition of sewer pipes [41]. In this study, five classes of land cover, which were obtained from the Sentinel-2 images level 1C by using the object-based classification [42], were used (FIGURE 3e).

Groundwater is considered an essential factor that influences sewer pipes [38], because groundwater at or above sewer pipes leads to water infiltration into the pipe, facilitating the deterioration processes. In addition, the availability of groundwater around the sewer pipe can destabilize the soil around the sewers leading to failures or collapses. In this study, a groundwater map was prepared using the IDW method and 31 drills data around the study area (FIGURE 3f).

Road traffic has been shown to have an impact on the deterioration process of sewers. Studies have shown that the condition of sewers located under roads as well as those in close proximity to roads are significantly affected [36]. In this study, traffic volume was calculated from the statistical data provided by the Norwegian Public Roads Administration (FIGURE 3g). There is no universal guideline for selecting the distance to road in modeling the sewer deterioration process. For instance, while Ahmadi, et al. [43] only considered pipes located under roads, the ratio of pipe length along the road was counted in the study by Yin, et al. [6]. Remarkably, Laakso, et al. [9] emphasized the pipes close to the tree (about 5 m) had a higher deteriorated degree compared to further ones, and the pipes far from roads will suffer from less influence compared to near ones. By using a similar approach for road distance, we consider a 5m-range road distance for the first road class; therefore, larger distances can be accepted for classifying further pipes into different road classes. In this study, five ordinal road classes were used based on the road's buffers of 0-5 m, 5-10 m, 10-20 m, 20-50 m, and >50 m (FIGURE 3h).

Soil type is one of the significant factors in the deterioration models because it affects runoff generation and groundwater and the influence of soil on sewers with larger sizes or buried deeper is more significant than the others [44]. In this study, 14 soil classes were used to construct the sewer condition models (FIGURE 3i).

III. BACKGROUND TO MACHINE LEARNING ALGORITHMS USED

Many ML algorithms have been proposed and applied not only in the water sector but also in other fields [7], [8], [9], [23], [45]. In this study, the following classification-based ML methods were selected based on their popular applications in classification problems.

A. CLASSIFICATION AND REGRESSION TREE

Classification And Regression Tree (CART) was first proposed by Breiman, et al. [46] to solve regression and

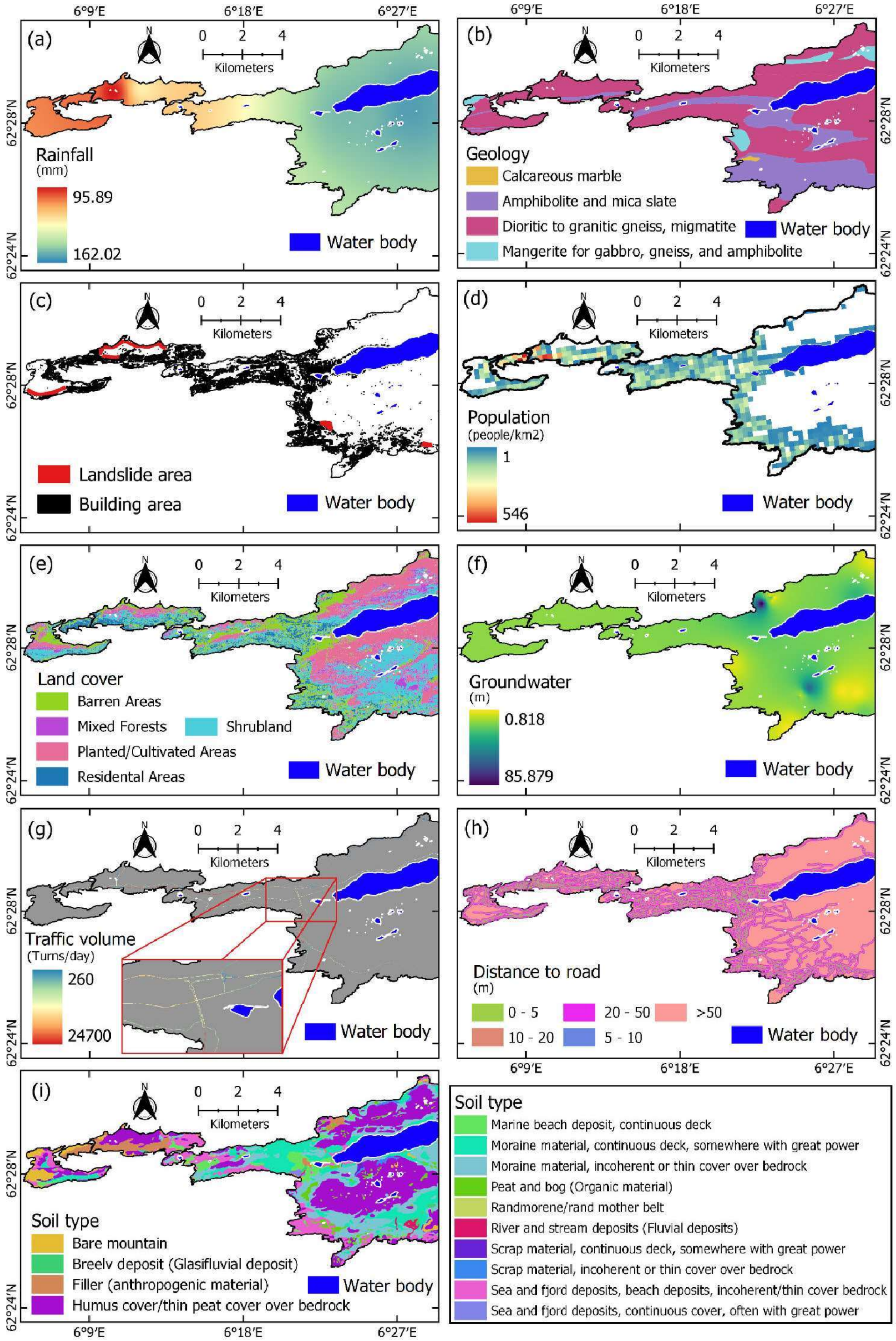

6°9′E
6°18′E
6°27′E
62°28′N
62°24′N
0 2 4
Kilometers
N
(a)
Rainfall
(mm)
95.89
162.02
Water body
(b)
Geology
Calcareous marble
Amphibolite and mica slate
Dioritic to granitic gneiss, migmatite
Mangerite for gabbro, gneiss, and amphibolite
Water body
(c)
Landslide area
Building area
Water body
(d)
Population
(people/km2)
1
546
Water body
(e)
Land cover
Barren Areas
Mixed Forests
Shrubland
Planted/Cultivated Areas
Residental Areas
Water body
(f)
Groundwater
(m)
0.818
85.879
Water body
(g)
Traffic volume
(Turns/day)
260
24700
Water body
(h)
Distance to road
(m)
0 - 5
20 - 50
>50
10 - 20
5 - 10
Water body
(i)
Soil type
Bare mountain
Breelv deposit (Glasifluvial deposit)
Filler (anthropogenic material)
Humus cover/thin peat cover over bedrock
Water body
Soil type
Marine beach deposit, continuous deck
Moraine material, continuous deck, somewhere with great power
Moraine material, incoherent or thin cover over bedrock
Peat and bog (Organic material)
Randmorene/rand mother belt
River and stream deposits (Fluvial deposits)
Scrap material, continuous deck, somewhere with great power
Scrap material, incoherent or thin cover over bedrock
Sea and fjord deposits, beach deposits, incoherent/thin cover bedrock
Sea and fjord deposits, continuous cover, often with great power

FIGURE 3. The environmental factors used in this study.

classification problems based on tree-based structures. In this method, the sewer dataset (also called the root node) was divided into binary values (good or bad condition) at each node using a series of recursive binary splits based on evaluating every possible predictor [47]. Finally, the predicted sewer's status was defined based on the most commonly occurring class of the node. The CART was selected in this study because this algorithm provided the largest information on the sewer status at each decision node using the input sewer factors [48]. This algorithm was used as a base classifier while constructing the ensemble techniques (e.g., AdaBoost or Gradient Tree Boosting).

B. RANDOM FOREST

Random Forest (RF) was developed by Breiman [49] to significantly improve classification accuracy by creating an ensemble of trees and letting them vote for the most popular class. In the RF model, the sewer input dataset was randomly split into classification trees, and the model was trained through bagging or bootstrap aggregating. The final sewer's condition status was obtained by aggregating the prediction from each tree. The RF model was applied for this study using the bootstrap technique to control the sub-sample size and get the average prediction from sub-decision trees to improve the predictive accuracy and control over-fitting.

C. ADABOOST CLASSIFICATION

AdaBoost method, which was introduced by Freund and Schapire [50], uses an adaptive re-sampling technique for controlling bias and variance to improve predictive performance. The AdaBoost randomly selects subsets from the sewer dataset; these subsets were assigned equal weights to implement a classifier for each iteration. The misclassified cases in the previous iteration will be reassigned with higher weights while the weights are kept for the correctly classified cases. A new normalized training subset is created, and a new iteration process continues. The iterative process is terminated if specific stopping criteria are satisfied, and the final sewer's status is the product of the weighted sum of all ensemble predictions.

D. GRADIENT TREE BOOSTING

Gradient boosting introduced by Friedman [51] sequentially fits a parameterized function (base learner) to pseudo residuals by least squares at each iteration using additive models. In Gradient Tree Boosting (GTB), a decision tree was used as a base learner. In the GTB model, a subset of the sewer dataset is randomly generated (without replacement) for each iteration. After that, this subsample is used in place for the full sample to fit the base classifier and update the model at the current iteration. The final sewer's condition status is obtained by minimizing the loss of function.

E. HISTOGRAM-BASED GRADIENT BOOSTING

Histogram-based Gradient Boosting (HGB) introduced by Guryanov [52] is a modification of the GTB and can increase the learning process and the model's prediction performance. This method divides the sewer training dataset into bins and constructs a histogram of feature values during the training phase. After every split decision tree, values of accumulated predictions of the sewer status are updated based on the deducted linear coefficient of split nodes. The iteration process is stopped when the stopping condition (e.g., the limit of tree depth or the number of leaves in the tree) is reached. Then, the sewer's condition status is defined using the best split points based on the feature histograms [53].

F. EXTREMELY RANDOMIZED TREES

Extremely Randomized Trees (ERT) is proposed by Geurts, et al. [54]; this algorithm splits nodes by making a small number of randomly chosen splits-points from the sewer dataset for each of the selected sewer condition status without re-sampling the dataset when building a tree. By using this approach, decision trees generated are entirely randomized whose structures are independent of the sewer's status. The sewer's status predicted by the single tree is aggregated to yield the final sewer's condition.

G. GAUSSIAN PROCESS

Gaussian Process (GP) model was introduced by Rasmussen [55] for classification and regression problems that generalize the Gaussian probability distribution. In the case of sewer condition status prediction, the sewer condition was transferred into $\{-1, +1\}$, a latent function f was used to predict the class membership probability for a new test pipe. The value of the function f was then mapped into the $[0, 1]$ interval using the probit function [56], where values of 0 and 1 denote the good and bad conditions of sewer pipes, respectively. Williams and Barber [57] introduced to use of Laplace's method for Gaussian approximation to the posterior over the latent function values. In this study, a Laplace method was applied to find a Gaussian approximation to the posterior because of its simplicity, scalability, and accuracy [58]. The predictive distribution of the sewer's status can be calculated by getting the weights of all possible predictions by their calculated posterior distribution [59]:

$$p\left(y^{*}=1|x^{*},\boldsymbol{X},\boldsymbol{y}\right)=\int_{f}^{*}\Phi\left(f^{*}\right)p\left(f^{*}|x^{*},\boldsymbol{X},\boldsymbol{y}\right)df^{*} \quad (1)$$

where $\boldsymbol{X}=[x_1,\ldots,x_n]^T$ and $\boldsymbol{y}=[y_1,\ldots,y_n]^T$ are vectors containing factors and sewer condition status, respectively; n is the number of sewer inspections; y^* and x^* are predicted sewer condition status and vector-containing factors of one sewer pipe, respectively; f^* and $\Phi(.)$ are variables corresponding to the test point x^* and the probit function, respectively.

H. GAUSSIAN NAIVE BAYES

The Gaussian Naive Bayes (GNB) classifies sewer status (good or bad condition) based on an assumption of having a Gaussian distribution on input factors using the Naive Bayes method [60]. The sewer status can be predicted using the Gaussian probability density function by substituting the parameters with the new input values [61]:

$$p\left(x_i|y\right) = \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{\left(x_i-\mu_y\right)^2}{2\sigma_y^2}} \tag{2}$$

where σ_y and μ_y are the variance and mean of the feature i^{th}, respectively; class y contains sewer condition status (good or bad condition).

I. BERNOULLI NAIVE BAYES

The Bernoulli Naive Bayes (BNB) classifies sewer status based on the Bayes theorem using sewer input data that are distributed according to multivariate Bernoulli distributions. The sewer status predicted by BNB is made based on the rule as follows [48]:

$$P\left(x_i|y\right) = P\left(i|y\right)x_i + \left(1 - P\left(i|y\right)\right)\left(1 - x_i\right) \tag{3}$$

where $P\left(x_i|y\right)$ is the likelihood of the features, x_i is the vector of the input factor the feature i^{th}, and y is sewer class (good or bad condition).

J. K-NEAREST NEIGHBORS

K-nearest neighbor (KNN) rule was first introduced by Cover and Hart [62] for classification problems. In the KNN model, the weight function is used to assess the degree of contribution of the nearer neighbors to the fit; the nearest neighbors are computed using search algorithms, the number of nearest neighbors is found using the grid-search method, and the distance metric is used to calculate the distance of one test observation from all the observations of the training dataset and find the nearest neighbors.

For sewer condition prediction, the distance from the sewer pipe x_i in the test dataset of each sample in the training dataset is computed. The top K points, which have the closest distance to x_i, are stored, and the status probability of sewer x_i is computed as follows [7]:

$$P\left(y = D, X = x_i\right) = \frac{1}{K}\sum\nolimits_{j\in A} I\left(y_j = D\right) \tag{4}$$

where $I\left(y_j = D\right)$ equals 1 if the instance y_j is in class D, otherwise, it equals 0, A is the dataset that contains K points, and D is the sewer status class (good or bad condition).

K. LOGISTIC REGRESSION

The logistic Regression (LR) model predicts the probability of the sewer condition status based on their relationship with input factors. The assumption of a linear relationship between factors and sewer condition status is unnecessary because this model uses the linear relationship between the logit of the input factors and the sewer status. The maximum likelihood method is generally used to estimate the intercept and coefficients based on the factors and sewer conditions. This method maximizes the probability of the sewer status given the fitted regression coefficients [63]. Although LR was designed for regression problems, this method was commonly used for classification problems (especially for binary classification) [64], [65].

L. RIDGE CLASSIFICATION

The Ridge regression method was introduced by Hoerl and Kennard [66] for solving the multicollinearity problem of covariates in samples. This method assumes that samples from each sewer condition class belong to a linear subspace, and a new test sample can be represented as a linear combination of class-specific training samples [67]. Ridge Classification (RC) algorithm is developed based on the Ridge regression, it converts the condition status of sewer pipes into $[-1, +1]$ and solves the problem as a regression task, minimizing the size of the coefficients by imposing a penalty, and the sewer condition class is assigned based on the highest value of the prediction result.

M. MULTI-LAYER PERCEPTRON NEURAL NETWORK

Multi-layer Perceptron Neural Network (MLP) is a fully connected class of feedforward Artificial Neural Networks (ANN). This network has three sequential layers: the input layer, the hidden layer, and the output layer. The number of neurons in the input layer equals the number of factors, two neurons in the output layer represent the expected sewer status (good or bad condition), and the number of hidden layers and hidden neurons is generally found by trial and error [68].

Before training the MLP model, each factor (i.e., physical and environmental factors) was assigned to each neuron and a bias unit was added to the input layer. Then, randomly generated weights were assigned for elements in the input layer, the weighted sums for neurons were calculated and the activation functions were used to transfer the results to the hidden layer. Similar processes were implemented in the hidden layer and the results were driven to the output layer. The error (the difference between the predicted sewer condition status and the measured condition) was calculated and minimized at the output layer. Finally, the derivation of the error function (loss function) with each weight in the network was determined and the model was updated. This was an iterative process over multiple epochs until the ideal weights were determined and the final sewer condition status was predicted based on these weights.

N. SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) was proposed by Cortes and Vapnik [69] to distinctly classify the data points using a hyperplane in N-dimensional space (N is the number of features). In the SVM model, the sewer pipe condition status is determined by maximizing the distance from the hyperplane to the data points of both good and bad conditions. The hyperplanes can be computed as follows [70]:

$$\begin{cases} y_i\left(\boldsymbol{w}.\emptyset^T\left(x_i\right)+\boldsymbol{b}\right) \geq 1-\varepsilon_i \\ \varepsilon_i \geq 0, i=1,2,\ldots,n \end{cases} \tag{5}$$

where n is the number of inspected pipes, $x_i \in \mathbb{R}^n$, $y_i \in \mathbb{R}$, x, and y are vectors that contain input factors and sewer's condition status respectively, $\boldsymbol{w}$ is the coefficient vector, $\boldsymbol{b}$ is and bias of the hyperplane in the feature space, $\emptyset$ is the non-linear mapping function, and ε_i are positive slack variables. The predicted condition status of the sewer pipe using the SVM is calculated as follows [71]:

$$\begin{cases} f\left(x\right)=sign\left(\sum_{i=1}^{n}\alpha_i y_i K\left(x_i,x\right)+b\right) \\ \sum_{i=1}^{n}\alpha_i y_i=0, 0 \leq \alpha_i \leq C, \forall i=1,2,\ldots,n \end{cases} \tag{6}$$

where auxiliary variables α_i are Lagrange multipliers, C is the regularization parameter, and $K\left(x, x_i\right)$is Kernel function.

IV. THE PROPOSED METHODOLOGY FOR SEWER CONDITION ASSESSMENT USING MACHINE LEARNING

The development of the sewer condition modeling for the study area involved the following interlinked steps: (1) Collecting physical characteristics and preprocessing auxiliary data to obtain environmental characteristics of the sewer pipes, (2) Dividing the inspected sewer pipes into training and validation datasets, (3) Eliminating redundant features, (4) Constructing conditional assessment models based on different ML algorithms, (5) Validating models' performance and accuracy, and (6) Preparation of condition maps of sewer pipelines. This procedure is shown in FIGURE 4.

A. GIS DATABASE

In this study, GIS was used to preprocess the environmental factors, which were highly related to spatial information. For instance, satellite images have been processed by GIS to create a land-use map using supervised classification in ArcGIS Pro software. Interpolated maps (i.e., rainfall, population, and groundwater) were computed using spatial analysis tools in GIS (e.g., raster calculator, interpolated function) to integrate environmental variables into sewer pipelines. Based on the information obtained from the GIS database, ML algorithms were applied for spatial modeling of the conditions of the sewer network.

Physical factors are normally recorded during the installation, operation, and maintenance of the sewer pipes. These data are managed by the local municipality or water agencies; therefore, these factors are easily assigned to each sewer pipe. In this study, the tabular data of physical variables were assigned for sewer pipe using GIS. However, environmental factors are collected from multi-source data (TABLE 2). Hence, environmental factors need to be aggregated for each sewer pipe.

From a spatial perspective, pipes in the sewer network are represented by "lines". However, a pipeline can cross many regions with different environmental characteristics (e.g., different land cover or soil type). Therefore, the location of the pipe geometry center is used to assign environmental factors. The data aggregation process is shown in FIGURE 4.

TABLE 4. The condition classes of pipe.

Damage class	Damage score	Sewer's condition	Aggregated class
Class 1	0 – 5	Very good status	Good condition
Class 2	6 – 10	Good status	
Class 3	11 – 20	Questionable status	
Class 4	21 – 50	Bad status	Bad condition
Class 5	>50	Very bad status	

In this study, the inspected grades were used as dependent variables for modeling the condition of the sewer pipes. The current conditions of the sewer pipes were assigned using damage scores obtained through the closed-circuit television (CCTV) method. Next, these damage scores were coded into damage classes representing the sewer conditions. According to Haugen and Viak [72], the conditions of sewer pipes in Norway are classified into five-grade scales based on their damage scores (TABLE 4).

There are different approaches for processing dependent variables to model the conditions of sewer pipes in the literature. For example, sewers in six-grade scales were aggregated into three grades [12]; in contrast, five grades of sewers were kept to develop models [73]. In this study, pipes in classes 1-2-3 were grouped into one class (good condition) and pipes in the remaining classes were aggregated into another class (bad condition) before building the condition models. Moreover, aggregating multi-output classes into smaller outputs will reduce the imbalance of the classification. The distribution of sewer classes according to age and material is shown in FIGURE 5, the data shows a slight imbalance in the dataset as a majority (approximately 62%) of inspected sewer pipes in Ålesund city are in good condition class.

After the GIS database was created, environmental factors were converted to raster format with a grid size of 5 m × 5 m in the WGS84-UTM32T (EPSG:32632). After that, the raster values were assigned for each pipe based on their geographical location. Categorical factors (i.e., pipe type, network type, pipe form, connection, geology, landslide area, land cover, building area, road class, and soil type) were coded by integer values. Furthermore, concrete, other, polypropylene, and polyvinyl chloride (PVC) pipes were coded by values 0, 1, 2, 3, and 4, respectively, for correlation analysis in this study.

B. PREPARATION OF TRAINING AND VALIDATION DATASETS

For the model development, a total of 1449 individual pipelines were used to train and validate the condition models of the sewage network. There is no universal guideline for

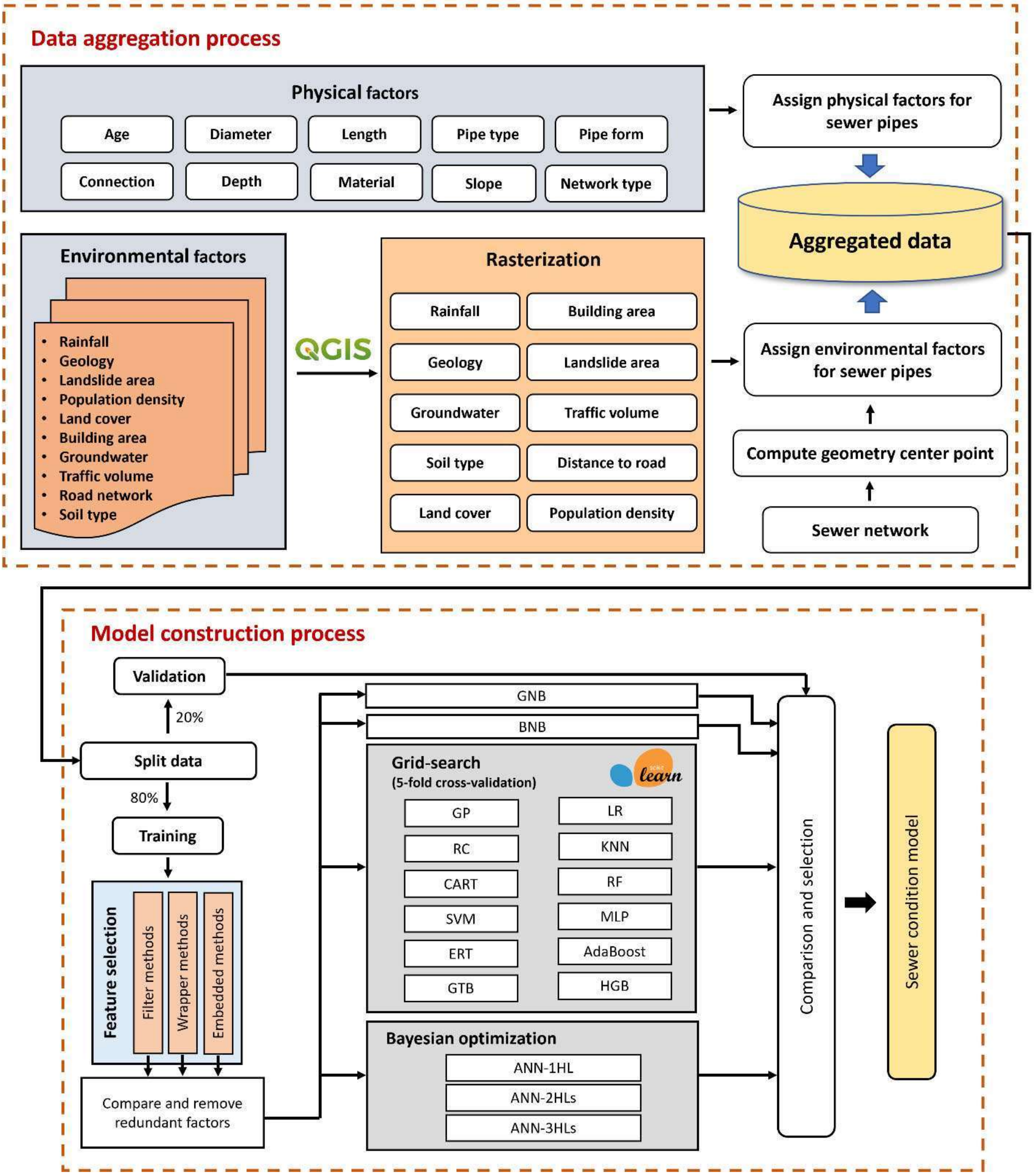


FIGURE 4. The framework for modeling the condition of a sewerage system.

choosing the ratio of training and validation datasets when modeling the condition of the sewage network. For example, a ratio of 80/20 was used for training/testing datasets [7]; in contrast, a ratio of 75/10/15 was used in another study to train, validate, and test the model [6]. In this study, this data was randomly divided with a ratio of 80% and 20% for training and testing datasets, respectively.

C. FEATURE SELECTION METHODS

In general, using redundant variables not only decreases the performance of ML models but also burdens computation. Feature selection techniques help to reduce dimensionality and clearly understand data [74]. Therefore, identifying and removing less significant factors before modeling is critical in preprocessing step [75].

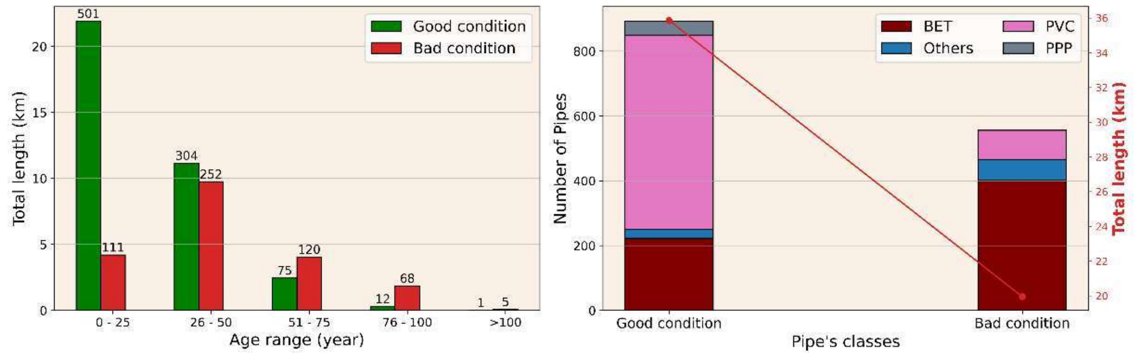


FIGURE 5. Different classes of sewer pipes in the study area.

The feature selection techniques used can generally be classified into three: filter, wrapper, and embedded methods. A brief description of the methods is as follows:

- *Filter methods* determine an optimal subset of variables mainly based on their statistical properties and relationship with the target variable. These methods do not remove the multicollinearity of features because they do not account for the interaction between variables [74]. Detailed information on the filter methods can be found in the study by Song, et al. [76].
- *Wrapper methods* select a subset of features by removing and adding the subsets accordingly based on the role of variables [77]. These methods often have higher performance than filter-based methods, these approaches however are more time-consuming [45]. Details on some wrapper methods can be found in the study by Nanda, et al. [78].
- *Embedded methods* apply the model-tuning process to perform feature selection [77]. These methods are a combination of the best qualities of filter and wrapper methods in which the variable selection process and classification have been implemented simultaneously using a learning algorithm [45]. Assessment of the importance of variables using embedded methods can be found in the study by Bhavan and Aggarwal [79].

In this study, six feature selection methods (two filter methods including Pearson's R (PR) and mutual information (MI), two wrapper methods including Boruta and Stepwise Feature Selection (SFS), and two embedded methods including Random Forest (RF) and Recursive Feature Elimination (RFE)) were used to assess the contribution of variables to ML models. The less important variables, which were defined by the majority of feature selection methods, were eliminated before constructing the ML models. The packages "Boruta", "stepAIC", "randomForest", "caret", and "kerlab" in the R Studio software were applied to implement the feature selection methods.

D. CONSTRUCTION OF SEWER CONDITION ASSESSMENT MODELS

The performance of the ML models highly depends on their hyperparameters. The typical hyperparameters of each ML model are shown in TABLE 5. The Scikit-learn and Keras libraries in Python were used to develop ML models in this study.

In this study, the MLP with a single hidden layer was investigated using the Scikit-learn library to predict the sewer status. The log-loss function was optimized in this model using stochastic gradient descent or Limited-memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) method because this method is especially appropriate for multi-variable optimization [80]. For comparison purposes, the multi-layer ANN architectures using the Keras library were also applied to predict sewer status. Hence, the ANN architectures with one, two, and three hidden layers were investigated. The Bayesian global optimization with Gaussian processes method was used for tuning some hyperparameters (e.g., the number of hidden layers, the number of neurons in the hidden layer, activation functions, and optimization functions) [81]. The ANN was built and trained using the Keras library in Python, the early-stopping technique was used to avoid over-fitting.

The grid-search method with 5-fold cross-validation was selected to find the optimal values of hyperparameters. The training dataset was randomly split into 5 equal-sized subsets and the cross-validation process was repeated 5 times for each of the five subsets to find the optimal solution. The optimum values of hyperparameters of each ML model were used to build the conditional assessment models.

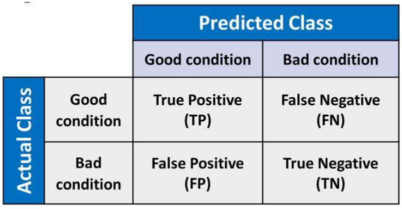
E. MODEL VALIDATION

In this study, the efficiency of the developed models was assessed using Geometric Mean (GM), Accuracy (ACC), F-Score, Matthew's correlation coefficient (MCC), the area under the Receiver Operating Characteristic curve (AUC-ROC), and the area under the Precision-Recall curve

TABLE 5. Summary of optimal parameters used in this study.

Model	Parameter and range	Optimal value	Avg. ACC (%)
GP	- Kernel function: *RBF, DP, MT, RQ, WK, ESSF*	- Kernel function: *MT* ($l_{GPC} = 1.0, \nu_{GPC} = 1.5$)	77.05
LR	- $C = 2^{-10}, 2^{-9}, \dots, 2^{9}, 2^{10}$ - Penalization: l_1, l_2 - Solver: *newton-cg, lbfgs, liblinear, sag, saga*	- $C = 2^{-5}$ - Penalization: l_2 - Solver: *newton-cg*	76.96
RC	- $\alpha = 0.1,0.2,\dots,0.9,1.0$	- $\alpha = 0.9$	77.05
GNB	-	-	73.69
BNB	-	-	74.55
KNN	- $K_{neighbor} = 1,2,\dots,99,100$ - Weight function: *uniform, distance* - Metric: *euclidean, manhattan, minkowski* - Search algorithm: *ball_tree, kd_tree, brute*	- $K_{neighbor} = 21$ - Weight function: *uniform* - Metric: *manhattan* - Search algorithm: *ball_tree*	78.69
DT	- Quality of a split: *gini, entropy* - $n_{feature} = 1,2,\dots,18,19$	- Quality of a split: *gini* - $n_{feature} = 17$	73.86
RF	- Quality of a split: *gini, entropy* - $n_{feature} = 1,2,\dots,18,19$ - $n_{tree} = 10,20,\dots,990,1000$	- Quality of a split: *entropy* - $n_{feature} = 19$ - $n_{tree} = 260$	78.08
SVM	- Kernel function: *linear, rbf, poly, sigmoid* - $C = 2^{-15}, 2^{-14}, \dots, 2^{4}, 2^{5}$ - $d = 1,2,\dots,9,10$ - $\gamma = 2^{-10}, 2^{-9}, \dots, 2^{2}, 2^{3}$	- Kernel function: *poly* - $C = 2^{3}$ - $d = 5$ - $\gamma = 2^{-3}$	77.74
MLP	- Activation function: *logistic, tanh, relu* - Solver: *lbfgs, sgd, adam* - $n_{neuron} = 1,2,\dots,199,200$	- Activation function: *tanh* - Solver: *adam* - $n_{neuron} = 99$	78.26
ERT	- Quality of a split: *gini, entropy* - $n_{feature} = 1,2,\dots,18,19$ - $n_{tree} = 10,20,\dots,990,1000$	- Quality of a split: *entropy* - $n_{feature} = 2$ - $n_{tree} = 660$	77.39
AdaBoost	- $n_{boosting} = 10,20,\dots,990,1000$ - $Learning\ rate = 0.1,0.2,\dots,0.9,1.0$	- $n_{boosting} = 40$ - $Learning\ rate = 0.5$	77.39
GTB	- $n_{estimator} = 10,20,\dots,990,1000$ - $Learning\ rate = 0.1,0.2,\dots,0.9,1.0$	- $n_{estimator} = 20$ - $Learning\ rate = 0.1$	78.52
HGB	- $n_{iteration} = 1,2,\dots,29,30$ - $Learning\ rate = 0.1,0.2,\dots,0.9,1.0$	- $n_{iteration} = 7$ - $Learning\ rate = 0.4$	78.34
ANN-1HL	- Activation function (hidden layer): *softmax, softplus, softsign, relu, tanh, sigmoid, hard_sigmoid, linear, selu, elu* - Optimizer: *SGD, RMSprop, Adagrad, Adadelta, Adam, Adamax, Nadam* - $n_{neuron} = 1,2,\dots,99,100$	- Activation function: *relu* - Optimizer: *RMSprop* - $n_{neuron} = 29$	80.24
ANN-2HLs		- Activation function: *softsign* - Optimizer: *Adagrad* - $n_{neuron}\ (1^{st}\ layer) = 73$ - $n_{neuron}\ (2^{nd}\ layer) = 95$	79.03
ANN-3HLs		- Activation function: *selu* - Optimizer: *Adagrad* - $n_{neuron}\ (1^{st}\ layer) = 82$ - $n_{neuron}\ (2^{nd}\ layer) = 21$ - $n_{neuron}\ (3^{rd}\ layer) = 2$	79.81

Abbreviations: l_{GP}: length-scale parameter of GP; ν_{RC}: smoothness function of the MT kernel; l_1: Lasso penalized regularization; l_2: Ridge penalized regularization; newton-cg: Newton Conjugate Gradient; lbfgs: Limited-memory Broyden–Fletcher–Goldfarb–Shanno, liblinear: Library for Large Linear Classification, sag: Stochastic Average Gradient (SAG), saga: variant of SAG; uniform: uniform weight function; distance: inverse distance weight function; euclidean: standard Euclidean metric; manhattan: manhattan metric; minkowski: minkowski metric; ball_tree: ball tree wrapped search algorithm; kd_tree: k-dimensional tree search algorithm; brute: brute-force search algorithm; gini: Gini impurity index; entropy: Entropy index; poly: polynomial kernel function; sigmoid: sigmoid kernel function; logistic: logistic sigmoid function; tanh: hyperbolic tan function; relu: rectified linear unit function; sgd: stochastic gradient descent; adam: Adaptive Movement Estimation; softplus: smooth approximation version of adam; hard_sigmoid: piece-wise linear approximation of the sigmoid function; Nadam: Adam with Nesterov momentum.



		Predicted Class	
		Good condition	Bad condition
Actual Class	Good condition	True Positive (TP)	False Negative (FN)
	Bad condition	False Positive (FP)	True Negative (TN)

FIGURE 6. Confusion matrix for binary classification.

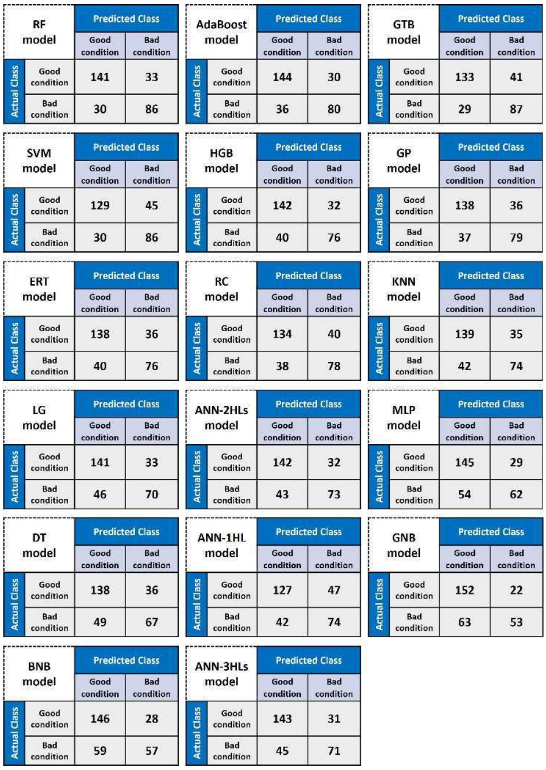
(AUC-PRC). These are expressed as follows (7)–(10), shown at the bottom of the page, whereas ACC and AUC-ROC are the most popular criteria for assessing the classification performance of ML algorithms, GM, F-Score, MCC, and AUC-PRC are sensitive to imbalanced datasets [82].

Other values including True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) are obtained from the confusion matrix for binary classification (FIGURE 6). The values of the confusion matrix (on the validation dataset) for the binary classification of each ML model are presented in FIGURE 7.

Because multiple assessment criteria were used, the TOPSIS (Technique for Order of Preference by Similarity to Ideal Solution) method for multiple-criteria decision-making was applied to rank the predictive performance of ML algorithms. This method proposed by Yoon and Hwang [83] is a multi-criteria decision analysis method that is based on the concept that the chosen alternative should have the shortest geometric distance from the positive ideal solution and the longest geometric distance from the negative ideal solution. This method was widely used to compare the performance of multiple ML algorithms using multiple criteria [84], [85]. In this study, the R package ''topsis'' introduced by Ihaka and Gentleman [86] was used to implement the TOPSIS method.

F. GENERATION OF SEWER CONDITION MAP

In general, the input factors affecting the sewer condition status are dynamic elements (e.g., rainfall or population). However, some other factors can be assumed to be unchanged over time. For example, a collapsed/damaged concrete pipe can be replaced by a newly similar concrete pipe, and similar things can happen with other factors such as diameter, pipe type, or network pipe. Therefore, in this study, we assume that there is only a fluctuation in rainfall, groundwater, and population density during the operational period of the sewer network while predicting future sewer condition status, the other factors are assumed as unchanged. The reason for choosing rainfall, groundwater, and population density as dynamic elements is that these factors are sensitive over time.



		Predicted Class	
Model	Actual Class	Good condition	Bad condition
RF model	Good condition	141	33
	Bad condition	30	86
AdaBoost model	Good condition	144	30
	Bad condition	36	80
GTB model	Good condition	133	41
	Bad condition	29	87
SVM model	Good condition	129	45
	Bad condition	30	86
HGB model	Good condition	142	32
	Bad condition	40	76
GP model	Good condition	138	36
	Bad condition	37	79
ERT model	Good condition	138	36
	Bad condition	40	76
RC model	Good condition	134	40
	Bad condition	38	78
KNN model	Good condition	139	35
	Bad condition	42	74
LG model	Good condition	141	33
	Bad condition	46	70
ANN-2HLs model	Good condition	142	32
	Bad condition	43	73
MLP model	Good condition	145	29
	Bad condition	54	62
DT model	Good condition	138	36
	Bad condition	49	67
ANN-1HL model	Good condition	127	47
	Bad condition	42	74
GNB model	Good condition	152	22
	Bad condition	63	53
BNB model	Good condition	146	28
	Bad condition	59	57
ANN-3HLs model	Good condition	143	31
	Bad condition	45	71

FIGURE 7. Confusion matrix of each machine learning model.

In this study, we assumed that the change in rainfall mainly causes the change in groundwater. Hence, the groundwater at the time t (GWL_t) was calculated using the interpolation method:

$$GWL_t = GWL_0 + \left(Rainfall_t - Rainfall_0\right) \tag{11}$$

where GWL_0 and $Rainfall_0$ are the groundwater and rainfall at the time t_0, and $Rainfall_t$ is rainfall at time

$$GM = \sqrt{\frac{TP}{TP+FN} \times \frac{TN}{TN+FP}} \tag{7}$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \tag{8}$$

$$F-Score = \frac{2TP}{2TP+FP+FN} \tag{9}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \tag{10}$$

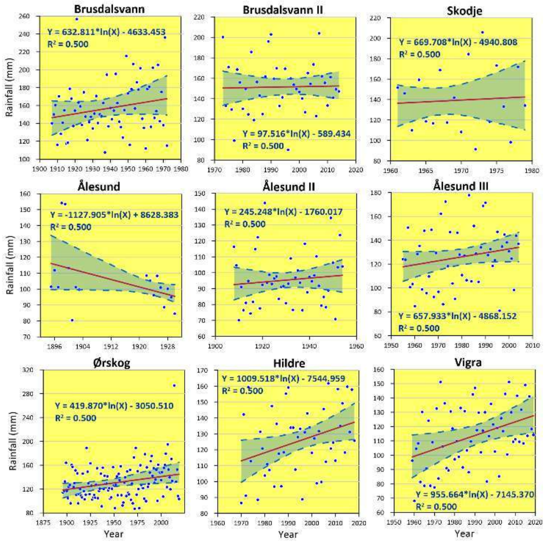


FIGURE 8. Fitting functions for rainfall interpolation.

$t(t = 2022, 2042, 2072)$. The value of each pixel of the groundwater map at time t was calculated in the following steps:

- *Step 1:* Determining the residual value of each pixel of the rainfall map between time t_0 and t: $Rainfall_t - Rainfall_0$. This step was implemented using the raster calculation function in GIS software.
- *Step 2:* Computing the value for each pixel using formula (11) to create the groundwater map at time t.

According to Worldometer [87], the annual population change (APC) in Ålesund city is 0.62% (2020-2021). To calculate the population in the years 2022, 2042, and 2072, we assumed that the population change is directly proportional to the APC. The value of pixel i in population density maps in the years 2022, 2042, and 2072 were calculated as follows:

$$P_t^i = P_{2018}^i \left[1 + 0.62\% \times (t - 2018)\right] \tag{12}$$

where P_{2018}^i and P_t^i are the population density in the year 2018 and at the time $t(t = 2022, 2042, 2072)$, respectively.

Rainfall at the weather stations in the years 2022, 2042, and 2072 was interpolated from historical measurements at the corresponding stations. The logarithm function was chosen to fit with the values. The coefficients of fitting functions and interpolated rainfall maps were shown in FIGURE 8. Then, maps of rainfall in the years 2022, 2042, and 2072 are created from interpolated rainfall values using the IDW method. Based on the two above steps, maps of interpolated groundwater and population in the years 2022, 2042, and 2072 are represented in FIGURE 9.

Predictive conditions of the sewer pipes in the future can be visualized on maps to provide a general overview of the status of the sewer system in the study area. These sewer condition maps can partly support utilities and water managers in determining the vulnerable regions affecting the condition of sewer pipes. QGIS, which is open-source software for GIS analysis, was used for data analysis and visualization.

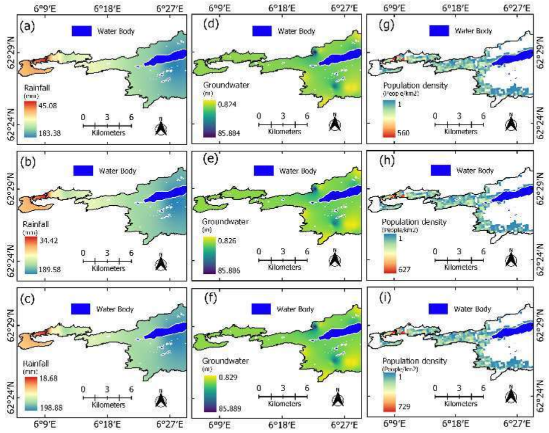


FIGURE 9. Maps of interpolated rainfall, groundwater, and population in the years 2022, 2042, and 2072.

V. RESULTS

A. ROLE OF FACTORS

As discussed in the above sections, different feature selection methods will produce different results for scoring the importance of each factor in developing the models. Hence, various algorithms of feature selection methods were investigated in this study, and the final decision to eliminate important factors was made based on the output results. Results of feature analysis based on the filter, wrapper, and embedded feature selection methods are shown in FIGURE 10, FIGURE 11, and FIGURE 12, respectively.

In the filter methods, the correlation analysis shows that material and age highly correlate with the sewer condition status (FIGURE 10a). More specifically, the material of sewer pipes has the highest correlation with their status (PR $= -0.54$), followed by the sewer's age (PR $= 0.46$), connection type (PR $= -0.35$), and pipe type (PR $= -0.33$).

It is evident that when the sewer's age increases, its condition deteriorates. The negative correlation of sewer material with the condition status indicates that concrete pipes are more durable than other pipes such as PVC pipes in the study area. The PR correlations of depth, diameter, network type, and distance to road are approximately equal to zero, indicating these factors have less influence on the condition of the pipes. The mutual information analysis shows sewer material and age to be the most significant factors in the sewer condition (FIGURE 10b). This analysis shows that pipe form and network type were insignificant factors.

The feature selection analysis using the wrapper methods is shown in FIGURE 11. The Boruta feature selection method reveals that the material of the sewer pipe is the most important factor for the condition assessment, followed by the age

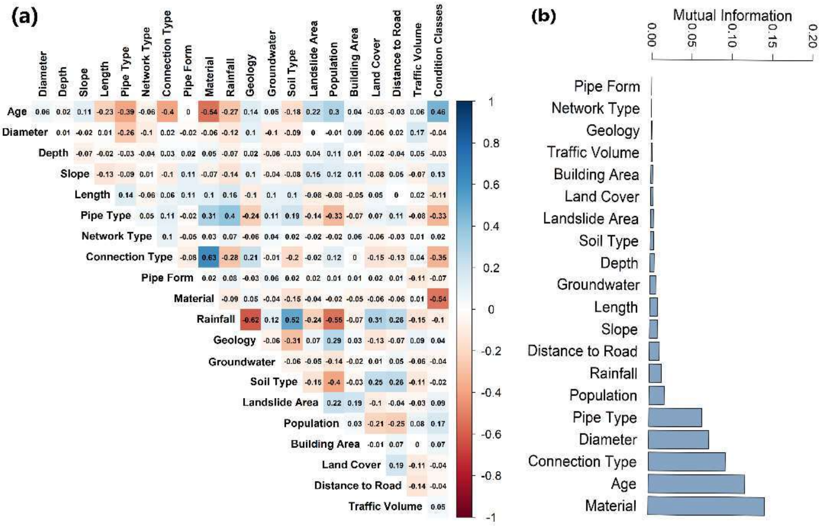


FIGURE 10. The filter feature selection methods: (a) PR correlation and (b) MI.

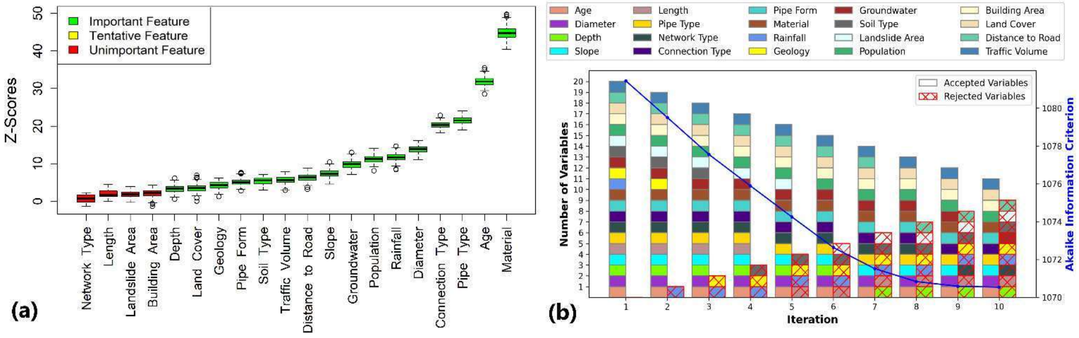


FIGURE 11. The wrapper feature selection methods: (a) Boruta and (b) SFS.

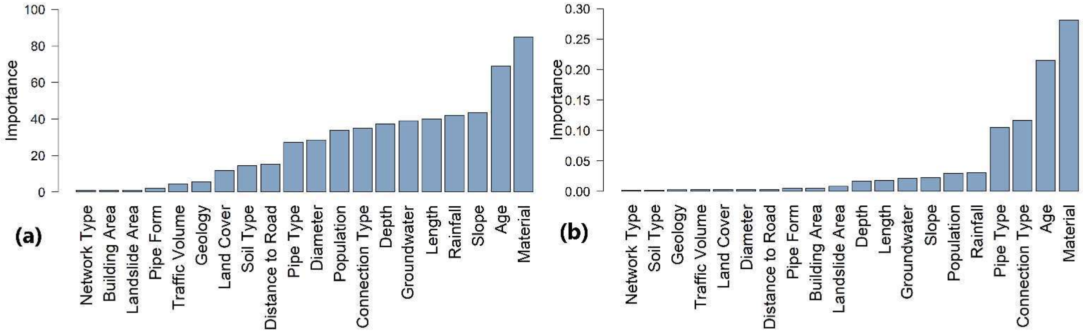


FIGURE 12. The embedded feature selection methods: (a) RF and (b) RFE.

of the sewer. Network type and length are assessed as the least important for this analysis (FIGURE 11a). The significance of factors for the condition assessment was assessed using the SFS method after ten iterations. FIGURE 11b shows that the

Akaike information criterion (AIC) value was achieved after ten iterations (about 1070.5). All factors were used to calculate AIC at the first iteration. The factors distance to road and geology were eliminated after the first and second iterations, respectively. Finally, material, age, diameter, slope, length, pipe type, connection type, pipe form, groundwater, building area, and traffic volume were accepted (FIGURE 11b).

FIGURE 12 shows the result of feature selection analysis using the embedded methods. For the RF method, sewer material was found to be the most significant factor, followed by the age of the sewer. Network type, building area, and landslide area were less significant (FIGURE 12a). Similar results were obtained by the REF method (FIGURE 12b).

Overall, all feature selection methods show that the material and age of sewer pipes are the most important factors. Based on the above results, network type was eliminated from the dataset before building the condition assessment models because the majority of feature selection methods assessed this variable as of less importance compared to others.

B. HYPERPARAMETERS OPTIMIZATION

Different ML models work with different parameters to generalize different data patterns. Hyperparameter tuning is used for optimal hyperparameters for the ideal model architecture. This study used the training dataset to select the best hyperparameters for each ML model using the grid-search method with a 5-fold cross-validation approach. The average accuracy was scored to define the best hyperparameters of each model. The tuned parameters, ranges of parameters, and their optimal values for each ML model are shown in TABLE 5. The accuracy of the ML models in TABLE 5 shows the average accuracy obtained from the grid-search method with a 5-fold cross-validation approach using the training dataset.

C. COMPARISON OF SEWER CONDITION MODELS

Performance prediction of ML models was generally assessed using the validation dataset based on the criteria and presented in TABLE 6.

It can be seen in TABLE 6 that the trees-based ML models (such as RF, AdaBoost, GTB, HGB, and ERT) have better performance than the others. In terms of the AI model, results show that the ANN architecture with 2 hidden layers (GM = 0.691, F-Score = 0.613, MCC = 0.398, AUC-ROC = 0.691, AUC-PRC = 0.707, and ACC = 71.72%) outperforms the single ANN architecture (GM = 0.684, F-Score = 0.624, MCC = 0.365, AUC-ROC = 0.684, AUC-PRC = 0.697, and ACC = 69.31%), and three-hidden layer ANN model produces the worst prediction (GM = 0.648, F-Score = 0. 562, MCC = 0.304, AUC-ROC = 0.648, AUC-PRC = 0.660, and ACC = 67.24%). The RF perform better in terms of all assessment criteria (GM = 0.776, F-Score = 0.732, MCC = 0.549, AUC-ROC = 0.776, AUC-PRC = 0.784, and ACC = 78.28%) indicating the most suitable condition assessment model.

TABLE 6. Prediction performance of used machine learning models in this analysis.

Model	Assessment criteria					
	GM	F-Score	MCC	AUC-ROC	AUC-PRC	ACC (%)
DT	0.685	0.612	0.379	0.685	0.699	70.69
RF	0.776	0.732	0.549	0.776	0.784	78.28
AdaBoost	0.758	0.708	0.522	0.759	0.771	77.24
GTB	0.757	0.713	0.507	0.757	0.765	75.86
HGB	0.735	0.679	0.478	0.736	0.748	75.17
ERT	0.734	0.678	0.472	0.734	0.746	74.83
GP	0.737	0.684	0.475	0.737	0.748	74.83
GNB	0.664	0.555	0.370	0.665	0.690	70.69
BNB	0.664	0.567	0.356	0.665	0.683	70.00
KNN	0.718	0.658	0.442	0.718	0.731	73.45
LG	0.707	0.639	0.424	0.707	0.721	72.76
RC	0.721	0.667	0.441	0.721	0.732	73.10
MLP	0.683	0.599	0.388	0.684	0.701	71.38
ANN-1HL	0.684	0.624	0.365	0.684	0.697	69.31
ANN-2HLs	0.691	0.613	0.398	0.691	0.707	71.72
ANN-3HLs	0.648	0.562	0.304	0.648	0.660	67.24
SVM	0.741	0.696	0.475	0.741	0.751	74.14

TABLE 7. The rank of machine learning models using the TOPSIS method.

Model	Score	Rank
RF	1.000	1
AdaBoost	0.883	2
GTB	0.838	3
SVM	0.713	4
HGB	0.706	5
GP	0.702	6
ERT	0.686	7
RC	0.571	8
KNN	0.564	9
LG	0.484	10
ANN-2HLs	0.370	11
MLP	0.323	12
DT	0.307	13
ANN-1HL	0.277	14
GNB	0.228	15
BNB	0.187	16
ANN-3HLs	0.015	17

The predictive performance of ML models is ranked using the TOPSIS method and presented in TABLE 7. The result shows that the RF is the best algorithm for modeling the sewer condition in the study area, followed by AdaBoost and GTB algorithms. Other algorithms are too simple (e.g., GNB or BNB) or too complex (e.g., ANN-3HLs) and they likely are not able to capture essential characteristics of the deterioration process in the sewer network in the study area.

D. SEWER CONDITION MAPS

The maps of the condition of sewer pipes in the year 2022, the next 20 years (2042), and the next 50 years (2072) in Ålesund city were created. In these maps, we assume there is no sewer pipes rehabilitation. For example, these pipes in bad condition in 2022 will maintain their conditions in 2042. FIGURE 13 shows the present status of the sewer network (in 2022) constructed using the RF model. The results show that the sewer pipes predicted in bad condition were largely

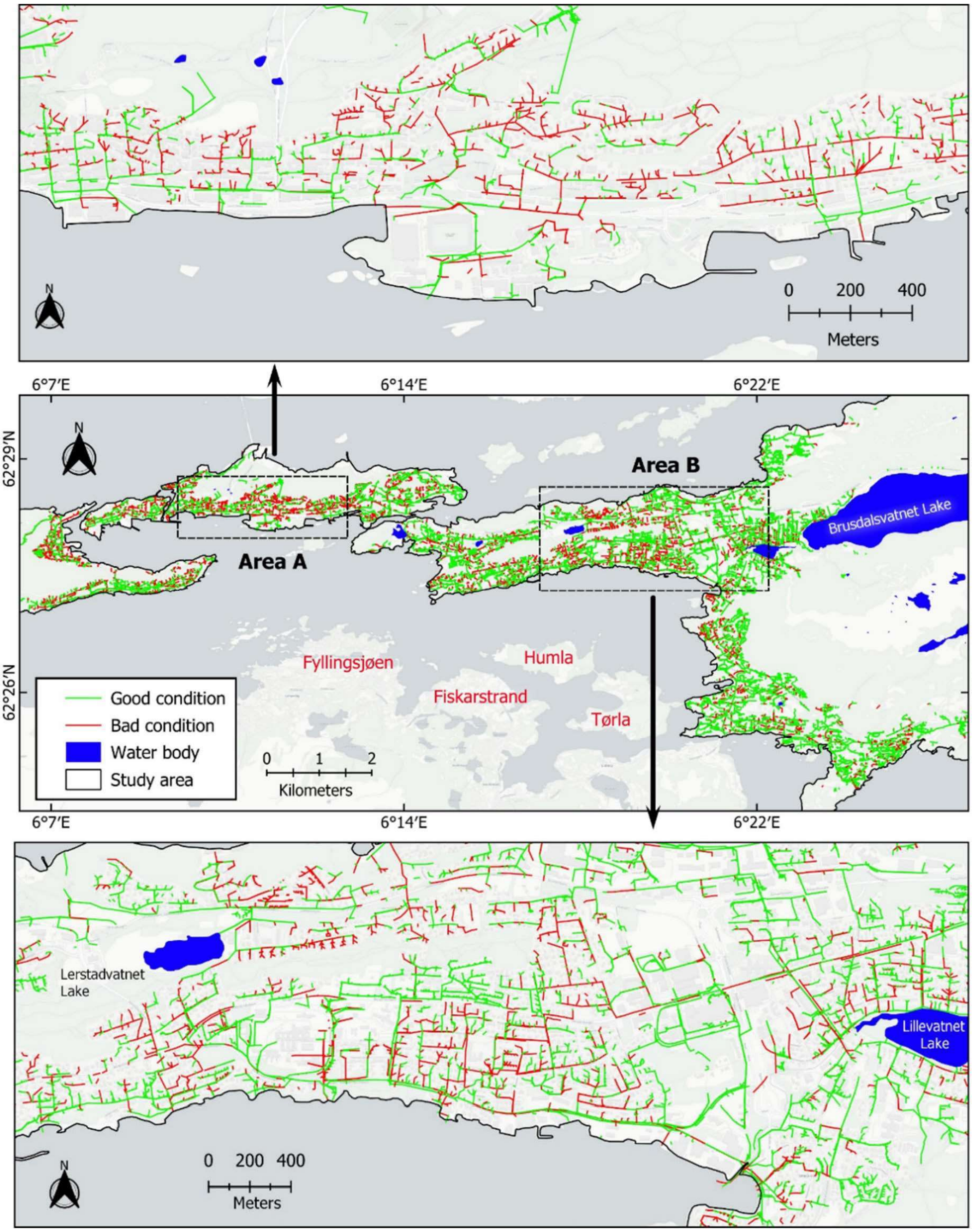


FIGURE 13. The sewer condition map in 2022 in the study area.

in the area marked A (the left-hand side rectangle), followed by the area marked B (the right-hand side rectangle). Due to confidential issues, maps of the condition status of the sewer network in the study area in the years 2042 and 2072 are not presented in this study. Interested readers can contact the authors to get detailed information.

FIGURE 14 shows the total length of sewers (in km) in each condition predicted in the years 2022, 2042, and 2072. The pie charts in the first row represent the number of predicted sewer pipes in each condition in the corresponding years. The results show that the number of sewers in bad condition after 50 years increased nearly two times,

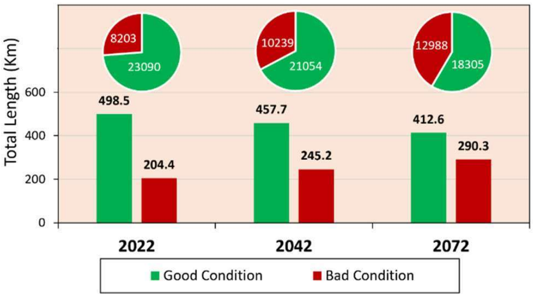


FIGURE 14. Summary of predicted classes in the study area.

from 8203 to 12988, corresponding to a total length of 204.4 km to 290.3 km. Moreover, it is worth noting that approximately half of the sewer pipes in the study will be in bad condition after 50 years. Maps of the sewer status received from the sewer condition model can support utilities and water managers in determining the spatial distribution of sewer pipe status in the city (FIGURE 13). Also, the length of pipes in each condition class shown in FIGURE 14 can help in investments in maintenance strategies.

VI. DISCUSSION

Sewer pipeline condition assessment is one of the critical steps in the water management process and a good condition model can support decision-making and maintenance strategies. In this research, the RF model was found to be more potent in predicting the sewer status in this study area.

Although all feature selection methods showed similarity in determining the two most important factors (material and age) and one less important factor (network type), however, the important degree of factors between methods is slightly different. This is likely due to the random feature of each method in splitting and combining subsets to optimize model performance [88].

Due to the unavailability of rainfall data in the study area, an interpolation map of rainfall was created from some weather stations near the study area. However, by using the IDW method, the accuracy of the interpolation rainfall map can be accepted for doing research on a large scale with annual time scales [89]. It is worth noticing that although the future rainfall map can be constructed from the climate projections [27], we used the linear method to interpolate annual average rainfall values at the weather stations and a rainfall map was created from these values. The main reason for doing this is that we want to apply the same approach for interpolating groundwater and population density maps.

One thing that should be paid attention to in this study is that some maps were established based on assumptions. For instance, the population density map was created based on the statistical data in 2018 and the future population density maps were created based on the annual average population change, or the assumption that changing groundwater only depends on the change of rainfall is only considered in this study. However, the change in these factors depends on different conditions and they should be considered in future studies. Another limitation of this study is that no operational factors are considered due to their unavailability at the time this study was undertaken. These factors can be accounted for in future studies to improve the model performance.

The final ML model has an accuracy of approximately 80%, indicating pretty good performance. This is because the deterioration of the sewer network is a complex process and depends on many different factors. Therefore, more pipe inspection data and factors should be considered to strengthen the predictive capability of the ML models.

In this study, pipe material and pipe age are the most important factors affecting the sewer pipe deterioration process. This conclusion is consistent with the result of Laakso, et al. [9] that found high-density polyethylene and reinforced concrete pipes were more durable than other materials. Age is a dynamic factor that immediately affects sewer deterioration as soon as the pipes are installed and it has been proved as the highest contributor to the deterioration process [90]. In contrast, network type (including wastewater, stormwater, and combined) is assessed as the lowest contribution to the model. This can be explained that most sewer pipes in the study are the main network type (FIGURE 2h).

VII. CONCLUSION

This study applied various ML algorithms for the assessment of sewer pipe conditions. A total of 1449 sewer pipelines derived from CCTV inspection were used to construct and verify the ML models. Six feature selection methods (i.e., filter, wrapper, and embedded methods) were applied to select the significant factors affecting sewer pipe conditions. The main conclusions from the study are:

- Sewer material is the most important factor affecting sewer pipe's condition status, followed by age. The sewer network type (stormwater, wastewater, and combined) was less important for the sewer condition in the study area.
- The RF model outperformed other ML models in modeling the sewer condition in the study area.
- Based on the RF model, maps of the condition of sewer pipes for the years 2022, 2042, and 2072 in Ålesund city were developed. These maps can be used as reference materials or documents in developing future maintenance strategies for the study area.
- The predictive performance of ML models can be improved by using more inspected sewer pipes as input for training ML models. Furthermore, other environmental and operational factors should be considered to improve the accuracy of the sewer condition model.

ACKNOWLEDGMENT

The authors would like to thank the Norwegian Mapping Authority, the Norwegian Climate Service Center, the Norwegian Water Resources and Energy Directorate, the Weather Atlas, the Mapping Authority, and the Norwegian Geological Survey for providing data for this research.

The data analysis and write-up thesis were operated as a part of the first author's Ph.D. studies at the Department of Ocean Operations and Civil Engineering, Faculty of Engineering, Norwegian University of Science and Technology, Norway.

REFERENCES

[1] K. Farkas, L. S. Hillary, S. K. Malham, J. E. McDonald, and D. L. Jones, "Wastewater and public health: The potential of wastewater surveillance for monitoring COVID-19," *Current Opinion Environ. Sci. Health*, vol. 17, pp. 14–20, Oct. 2020, doi: 10.1016/j.coesh.2020.06.001.

[2] J. Kvitsjøen, K. H. Paus, J. T. Bjerkholt, T. Fergus, and O. Lindholm, "Intensifying rehabilitation of combined sewer systems using trenchless technology in combination with low impact development and green infrastructure," *Water Sci. Technol.*, vol. 83, no. 12, pp. 2947–2962, Jun. 2021.

[3] E. Kuliczkowska, A. Kuliczkowski, and A. Parka, "Damages in vitrified clay sewers in service for 130–142 years," *Eng. Failure Anal.*, vol. 135, May 2022, Art. no. 106103, doi: 10.1016/j.engfailanal.2022.106103.

[4] M. Rostad and A. Kinei, "Finansieringsbehov i vannbransjen 2016–2040," Norsk Vann Rapport, Norsk Vann BA, Hamar, Norway, Tech. Rep. 223/2017, 2017, vol. 223. [Online]. Available: https://vannsenter.no/wp-content/uploads/2019/06/Finansieringsbehov-i-vannbransjen-2016-2040.Norsk-Vann.R223.pdf

[5] T. Breen. *Kronikk: Behov for Store Investeringer I Vann og Avløp (Chronicle: Need for Large Investments in Water and Wastewater)*. Norsk Vann BA. Accessed: Oct. 10, 2022. [Online]. Available: https://norskvann.no/behov-for-store-investeringer-i-vann-og-avlop/

[6] X. Yin, Y. Chen, A. Bouferguene, and M. Al-Hussein, "Data-driven bi-level sewer pipe deterioration model: Design and analysis," *Autom. Construct.*, vol. 116, Aug. 2020, Art. no. 103181, doi: 10.1016/j.autcon.2020.103181.

[7] X. Fan, X. Wang, X. Zhang, and P. E. F. A. X. B. Yu, "Machine learning based water pipe failure prediction: The effects of engineering, geology, climate and socio-economic factors," *Rel. Eng. Syst. Saf.*, vol. 219, Mar. 2022, Art. no. 108185, doi: 10.1016/j.ress.2021.108185.

[8] A. Hawari, F. Alkadour, M. Elmasry, and T. Zayed, "A state of the art review on condition assessment models developed for sewer pipelines," *Eng. Appl. Artif. Intell.*, vol. 93, Aug. 2020, Art. no. 103721, doi: 10.1016/j.engappai.2020.103721.

[9] T. Laakso, T. Kokkonen, I. Mellin, and R. Vahala, "Sewer condition prediction and analysis of explanatory factors," *Water*, vol. 10, no. 9, p. 1239, Sep. 2018, doi: 10.3390/W10091239.

[10] M. Wang, Y. Deng, J. Won, and J. C. P. Cheng, "An integrated underground utility management and decision support based on BIM and GIS," *Autom. Construct.*, vol. 107, Nov. 2019, Art. no. 102931, doi: 10.1016/j.autcon.2019.102931.

[11] C. Salihu, M. Hussein, S. R. Mohandes, and T. Zayed, "Towards a comprehensive review of the deterioration factors and modeling for sewer pipelines: A hybrid of bibliometric, scientometric, and meta-analysis approach," *J. Cleaner Prod.*, vol. 351, Jun. 2022, Art. no. 131460, doi: 10.1016/j.jclepro.2022.131460.

[12] N. Caradot, M. Riechel, P. Rouault, A. Caradot, N. Lengemann, E. Eckert, A. Ringe, F. Clemens, and F. Cherqui, "The influence of condition assessment uncertainties on sewer deterioration modelling," *Struct. Infrastruct. Eng.*, vol. 16, no. 2, pp. 287–296, Feb. 2020, doi: 10.1080/15732479.2019.1653938.

[13] F. Tscheikner-Gratl, N. Caradot, F. Cherqui, J. P. Leitão, M. Ahmadi, J. G. Langeveld, Y. Le Gat, L. Scholten, B. Roghani, J. P. Rodríguez, and M. Lepot, "Sewer asset management–state of the art and research needs," *Urban Water J.*, vol. 16, no. 9, pp. 662–675, 2019, doi: 10.1080/1573062X.2020.1713382.

[14] N. Caradot, M. Riechel, M. Fesneau, N. Hernandez, A. Torres, H. Sonnenberg, E. Eckert, N. Lengemann, J. Waschnewski, and P. Rouault, "Practical benchmarking of statistical and machine learning models for predicting the condition of sewer pipes in Berlin, Germany," *J. Hydroinformatics*, vol. 20, no. 5, pp. 1131–1147, Sep. 2018, doi: 10.2166/HYDRO.2018.217.

[15] R. Heydarzadeh, M. Tabesh, and M. Scholz, "Dissolved oxygen determination in sewers using flow hydraulic parameters as part of a physical-biological simulation model," *J. Hydroinformatics*, vol. 24, no. 1, pp. 1–15, Jan. 2022, doi: 10.2166/hydro.2021.051.

[16] B. Wei, L. Chen, H. Li, D. Yuan, and G. Wang, "Optimized prediction model for concrete dam displacement based on signal residual amendment," *Appl. Math. Model.*, vol. 78, pp. 20–36, Feb. 2020, doi: 10.1016/j.apm.2019.09.046.

[17] G. Vladeanu, J. Matthews, and M. Asce, "Wastewater pipe condition rating model using multicriteria decision analysis," *J. Water Resour. Planning Manage.*, vol. 145, no. 12, Dec. 2019, Art. no. 04019058, doi: 10.1061/(ASCE)WR.1943-5452.0001134.

[18] J. I. Sempewo and L. Kyokaali, "Comparative performance of regression and the Markov based approach in the prediction of the future condition of a water distribution pipe network amidst data scarce situations: A case study of Kampala water, Uganda," *Water Pract. Technol.*, vol. 14, no. 4, pp. 946–958, Dec. 2019, doi: 10.2166/WPT.2019.075.

[19] G. Kabir, N. B. C. Balek, S. Tesfamariam, and M. Asce, "Sewer structural condition prediction integrating Bayesian model averaging with logistic regression," *J. Perform. Constructed Facilities*, vol. 32, no. 3, Jun. 2018, Art. no. 04018019, doi: 10.1061/(ASCE)CF.1943-5509.0001162.

[20] A. Altarabsheh, M. Ventresca, and A. Kandil, "New approach for critical pipe prioritization in wastewater asset management planning," *J. Comput. Civil Eng.*, vol. 32, no. 5, Sep. 2018, Art. no. 04018044, doi: 10.1061/(ASCE)CP.1943-5487.0000784.

[21] S. Zamanian, J. Hur, and A. Shafieezadeh, "A high-fidelity computational investigation of buried concrete sewer pipes exposed to truckloads and corrosion deterioration," *Eng. Struct.*, vol. 221, Oct. 2020, Art. no. 111043, doi: 10.1016/j.engstruct.2020.111043.

[22] T. Ahmad, H. Chen, R. Huang, G. Yabin, J. Wang, J. Shair, H. M. A. Akram, S. A. H. Mohsan, and M. Kazim, "Supervised based machine learning models for short, medium and long-term energy prediction in distinct building environment," *Energy*, vol. 158, pp. 17–32, Sep. 2018, doi: 10.1016/j.energy.2018.05.169.

[23] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, p. 281, Dec. 2019, doi: 10.1186/s12911-019-1004-8.

[24] X. Li, F. Khademi, Y. Liu, M. Akbari, C. Wang, P. L. Bond, J. Keller, and G. Jiang, "Evaluation of data-driven models for predicting the service life of concrete sewer pipes subjected to corrosion," *J. Environ. Manage.*, vol. 234, pp. 431–439, Mar. 2019, doi: 10.1016/j.jenvman.2018.12.098.

[25] D. J. Kovacs, Z. Li, B. W. Baetz, Y. Hong, S. Donnaz, X. Zhao, P. Zhou, H. Ding, and Q. Dong, "Membrane fouling prediction and uncertainty analysis using machine learning: A wastewater treatment plant case study," *J. Membrane Sci.*, vol. 660, Oct. 2022, Art. no. 120817, doi: 10.1016/j.memsci.2022.120817.

[26] D. Climate. *Ålesund Climate: Average Temperature, Weather by Month, Ålesund Water Temperature—Climate-Data*. Accessed: Apr. 20, 2020. [Online]. Available: https://en.climate-data.org/europe/norway/m%C3%B8re-og-romsdal/alesund-9937/

[27] I. Hanssen-Bauer, H. Drange, E. Førland, L. Roald, K. Børsheim, H. Hisdal, D. Lawrence, A. Nesje, S. Sandven, and A. Sorteberg, "Climate in Norway 2100—A knowledge base for climate adaptation," in *Background Information to NOU Climate Adaptation*. Oslo, Norway: Norsk klimasenter, 2017.

[28] A. V. Dyrrdal and E. J. Førland. (2019). *Klimapåslag for Korttidsnedbør. Anbefalte Verdier for Norge (Climate Surcharge for Short-Term Precipitation. Recommended Values for Norway)*. Norsk klimaservicesenter, Norway. [Online]. Available: https://klimaservicesenter.no/

[29] B. Roghani, F. Cherqui, M. Ahmadi, P. Le Gauffre, and M. Tabesh, "Dealing with uncertainty in sewer condition assessment: Impact on inspection programs," *Autom. Construct.*, vol. 103, pp. 117–126, Jul. 2019, doi: 10.1016/j.autcon.2019.03.012.

[30] F. Shi. (2018). *Data-Driven Predictive Analytics for Water Infrastructure Condition Assessment and Management*. [Online]. Available: https://open.library.ubc.ca/collections/24/items/1.0372323

[31] T. Laakso, T. Kokkonen, I. Mellin, and R. Vahala, "Sewer life span prediction: Comparison of methods and assessment of the sample impact on the results," *Water*, vol. 11, no. 12, p. 2657, Dec. 2019. [Online]. Available: https://www.mdpi.com/2073-4441/11/12/2657

[32] E. Ana, W. Bauwens, M. Pessemier, C. Thoeye, S. Smolders, I. Boonen, and G. De Gueldre, "An investigation of the factors influencing sewer structural deterioration," *Urban Water J.*, vol. 6, no. 4, pp. 303–312, Oct. 2009, doi: 10.1080/15730620902810902.

[33] I. Bakry, H. Alzraiee, M. E. Masry, K. Kaddoura, and T. Zayed, "Condition prediction for cured-in-place pipe rehabilitation of sewer mains," *J. Perform. Constructed Facilities*, vol. 30, no. 5, Oct. 2016, Art. no. 04016016, doi: 10.1061/(ASCE)CF.1943-5509.0000866.

[34] M. M. Mohammadi, M. Najafi, S. Kermanshachi, V. Kaushal, and R. Serajiantehrani, "Factors influencing the condition of sewer pipes: State-of-the-art review," *J. Pipeline Syst. Eng. Pract.*, vol. 11, no. 4, p. 03120002, 2020, doi: 10.1061/(ASCE)PS.1949-1204.0000483.

[35] T.-Y. Kwak, S.-I. Woo, C.-K. Chung, and J. Kim, "Experimental assessment of the relationship between rainfall intensity and sinkholes caused by damaged sewer pipes," *Natural Hazards Earth Syst. Sci.*, vol. 20, no. 12, pp. 3343–3359, Dec. 2020, doi: 10.5194/nhess-20-3343-2020.

[36] B. Salman and O. Salem, "Modeling failure of wastewater collection lines using various section-level regression models," *J. Infrastruct. Syst.*, vol. 18, no. 2, pp. 146–154, 2012, doi: 10.1061/(ASCE)IS.1943-555X.0000075.

[37] E. Belief, "GIS based spatial modeling to mapping and estimation relative risk of different diseases using inverse distance weighting (IDW) interpolation algorithm and evidential belief function (EBF)(case study: Minor part of Kirkuk City, Iraq)," *Int. J. Eng. Technol.*, vol. 7, no. 4, pp. 91–185, 2018.

[38] X. Su, T. Liu, M. Beheshti, and V. Prigiobbe, "Relationship between infiltration, sewer rehabilitation, and groundwater flooding in coastal urban areas," *Environ. Sci. Pollut. Res.*, vol. 27, no. 13, pp. 14288–14298, May 2020, doi: 10.1007/s11356-019-06513-z.

[39] T. Liu, X. Su, and V. Prigiobbe, "Groundwater-sewer interaction in urban coastal areas," *Water*, vol. 10, no. 12, p. 1774, Dec. 2018, doi: 10.3390/w10121774.

[40] M. Y. Tebbouche, D. A. Benamar, H. M. Hassan, A. P. Singh, R. Bencharif, D. Machane, A. A. Meziani, and Z. Nemer, "Characterization of el kherba landslide triggered by the August 07, 2020, Mw = 4.9 Mila earthquake (Algeria) based on post-event field observations and ambient noise analysis," *Environ. Earth Sci.*, vol. 81, no. 2, p. 46, Jan. 2022, doi: 10.1007/s12665-022-10172-8.

[41] L. M. de Oliveira, P. Maillard, and E. J. de Andrade Pinto, "Application of a land cover pollution index to model non-point pollution sources in a Brazilian watershed," *CATENA*, vol. 150, pp. 124–132, Mar. 2017, doi: 10.1016/j.catena.2016.11.015.

[42] A. Sánchez-Espinosa and C. Schröder, "Land use and land cover mapping in wetlands one step closer to the ground: Sentinel-2 versus landsat 8," *J. Environ. Manage.*, vol. 247, pp. 484–498, Oct. 2019, doi: 10.1016/j.jenvman.2019.06.084.

[43] M. Ahmadi, F. Cherqui, J.-C. De Massiac, and P. Le Gauffre, "Influence of available data on sewer inspection program efficiency," *Urban Water J.*, vol. 11, no. 8, pp. 641–656, Nov. 2014, doi: 10.1080/1573062X.2013.831910.

[44] M. Beheshti, S. Sægrov, and R. Ugarelli, "Infiltration/inflow assessment and detection in urban sewer system," Norwegian Water Assoc. (Norsk vannforening), Oslo, Norway, Tech. Rep. 1, 2015. [Online]. Available: https://vannforeningen.no/wp-content/uploads/2015/01/Beheshti.pdf

[45] A. Nazir and R. A. Khan, "A novel combinatorial optimization based feature selection method for network intrusion detection," *Comput. Secur.*, vol. 102, Mar. 2021, Art. no. 102164, doi: 10.1016/j.cose.2020.102164.

[46] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Evanston, IL, USA: Routledge, 1984.

[47] H. Ebrahimy, B. Feizizadeh, S. Salmani, and H. Azadi, "A comparative study of land subsidence susceptibility mapping of Tasuj plane, Iran, using boosted regression tree, random forest and classification and regression tree methods," *Environ. Earth Sci.*, vol. 79, no. 10, p. 223, May 2020, doi: 10.1007/s12665-020-08953-0.

[48] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, and V. Dubourg, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Nov. 2011. [Online]. Available: https://scikit-learn.org/stable/

[49] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[50] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[51] J. H. Friedman, "Stochastic gradient boosting," *Comput. Statist. Data Anal.*, vol. 38, no. 4, pp. 367–378, Feb. 2002, doi: 10.1016/S0167-9473(01)00065-2.

[52] A. Guryanov, "Histogram-based algorithm for building gradient boosting ensembles of piecewise linear decision trees," in *Analysis of Images, Social Networks and Texts*, W. M. P. van der Aalst, V. Batagelj, D. I. Ignatov, M. Khachay, V. Kuskova, A. Kutuzov, S. O. Kuznetsov, I. A. Lomazova, N. Loukachevitch, A. Napoli, P. M. Pardalos, M. Pelillo, A. V. Savchenko, E. Tutubalina, Eds., Cham, Switzerland: Springer, 2019, pp. 39–50.

[53] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–9.

[54] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach. Learn.*, vol. 63, no. 1, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.

[55] C. E. Rasmussen, "Gaussian processes in machine learning," in *Summer School on Machine Learning*. Berlin, Germany: Springer, 2003, pp. 63–71, doi: 10.1007/978-3-540-28650-9_4.

[56] F. Rodrigues, F. Pereira, and B. Ribeiro, "Gaussian process classification and active learning with multiple annotators," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 433–441.

[57] C. K. I. Williams and D. Barber, "Bayesian classification with Gaussian processes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1342–1351, Dec. 1998.

[58] D. Zilber and M. Katzfuss, "Vecchia–Laplace approximations of generalized Gaussian processes for big non-Gaussian spatial data," *Comput. Statist. Data Anal.*, vol. 153, Jan. 2021, Art. no. 107081, doi: 10.1016/j.csda.2020.107081.

[59] M. Kuss, C. E. Rasmussen, and R. Herbrich, "Assessing approximate inference for binary Gaussian process classification," *J. Mach. Learn. Res.*, vol. 6, no. 10, pp. 1–26, 2005.

[60] A. H. Jahromi and M. Taheri, "A non-parametric mixture of Gaussian naive Bayes classifiers based on local independent features," in *Proc. Artif. Intell. Signal Process. Conf. (AISP)*, Oct. 2017, pp. 209–212, doi: 10.1109/AISP.2017.8324083.

[61] L. Cataldi, L. Tiberi, and G. Costa, "Estimation of MCS intensity for Italy from high quality accelerometric data, using GMICEs and Gaussian Naïve Bayes classifiers," *Bull. Earthq. Eng.*, vol. 19, no. 6, pp. 2325–2342, Apr. 2021, doi: 10.1007/s10518-021-01064-6.

[62] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. IT-13, no. 1, pp. 21–27, Jan. 1967.

[63] A. Y. C. Kuk and C.-H. Chen, "A mixture model combining logistic regression with proportional hazards regression," *Biometrika*, vol. 79, no. 3, pp. 531–541, 1992, doi: 10.1093/biomet/79.3.531.

[64] W. Książek, M. Gandor, and P. Pławiak, "Comparison of various approaches to combine logistic regression with genetic algorithms in survival prediction of hepatocellular carcinoma," *Comput. Biol. Med.*, vol. 134, Jul. 2021, Art. no. 104431, doi: 10.1016/j.compbiomed.2021.104431.

[65] T. Dokeroglu, A. Deniz, and H. E. Kiziloz, "A robust multiobjective Harris' hawks optimization algorithm for the binary classification problem," *Knowl.-Based Syst.*, vol. 227, Sep. 2021, Art. no. 107219, doi: 10.1016/j.knosys.2021.107219.

[66] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970, doi: 10.1080/00401706.1970.10488634.

[67] J. He, L. Ding, L. Jiang, and L. Ma, "Kernel ridge regression classification," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2014, pp. 2263–2267, doi: 10.1109/IJCNN.2014.6889396.

[68] U. Orhan, M. Hekim, and M. Ozer, "EEG signals classification using the K-means clustering and a multilayer perceptron neural network model," *Expert Syst Appl.*, vol. 38, no. 10, pp. 13475–13481, Sep. 2011, doi: 10.1016/j.eswa.2011.04.149.

[69] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Jul. 1995.

[70] A. Zendehboudi, M. A. Baseer, and R. Saidur, "Application of support vector machine models for forecasting solar and wind energy resources: A review," *J. Cleaner Prod.*, vol. 199, pp. 272–285, Oct. 2018, doi: 10.1016/j.jclepro.2018.07.164.

[71] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.

[72] H. J. Haugen and A. Viak. (2018). *Datafl yt—Klassifi Sering av Avløpsledninger*. Norwegian Water BA. [Online]. Available: https://docplayer.me/211256711-Norsk-vann-rapport-dataflyt-klassifisering-av-avlopsledninger.html

[73] J. Mashford, D. Marlow, D. Tran, and R. May, "Prediction of sewer condition grade using support vector machines," *J. Comput. Civil Eng.*, vol. 25, no. 4, pp. 283–290, 2011, doi: 10.1061/(ASCE)CP.1943-5487.0000089.

[74] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[75] Q. Al-Tashi, S. J. Abdulkadir, H. M. Rais, S. Mirjalili, and H. Alhussian, "Approaches to multi-objective feature selection: A systematic literature review," *IEEE Access*, vol. 8, pp. 125076–125096, 2020, doi: 10.1109/ACCESS.2020.3007291.

[76] X.-F. Song, Y. Zhang, D.-W. Gong, and X.-Y. Sun, "Feature selection using bare-bones particle swarm optimization with mutual information," *Pattern Recognit.*, vol. 112, Apr. 2021, Art. no. 107804, doi: 10.1016/j.patcog.2020.107804.

[77] J. Y.-L. Chan, S. M. H. Leow, K. T. Bea, W. K. Cheng, S. W. Phoong, Z.-W. Hong, J.-M. Lin, and Y.-L. Chen, "A correlation-embedded attention module to mitigate multicollinearity: An algorithmic trading application," *Mathematics*, vol. 10, no. 8, p. 1231, Apr. 2022, doi: 10.3390/math10081231.

[78] M. A. Nanda, K. B. Seminar, A. Maddu, and D. Nandika, "Identifying relevant features of termite signals applied in termite detection system," *Ecological Informat.*, vol. 64, Sep. 2021, Art. no. 101391, doi: 10.1016/j.ecoinf.2021.101391.

[79] A. Bhavan and S. Aggarwal, "Stacked generalization with wrapper-based feature selection for human activity recognition," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2018, pp. 1064–1068, doi: 10.1109/SSCI.2018.8628830.

[80] V. Q. Tran, V. Q. Dang, H. Q. Do, and L. S. Ho, "Investigation of ANN architecture for predicting residual strength of clay soil," *Neural Comput. Appl.*, vol. 34, no. 21, pp. 19253–19268, Nov. 2022, doi: 10.1007/s00521-022-07547-0.

[81] P.-I. Schneider, X. G. Santiago, C. Rockstuhl, and S. Burger, "Global optimization of complex optical structures using Bayesian optimization based on Gaussian processes," *Proc. SPIE*, vol. 10335, pp. 141–149, Jun. 2017.

[82] A. Tharwat, "Classification assessment methods," *Appl. Comput. Informat.*, vol. 17, pp. 168–192, Aug. 2021, doi: 10.1016/j.aci.2018.08.003.

[83] K. P. Yoon and C.-L. Hwang, *Multiple Attribute Decision Making: An Introduction*. Newbury Park, CA, USA: Sage, 1995.

[84] M. Y. L. Vazquezl, L. A. B. Peñafiel, S. X. S. Muñoz, and M. A. Q. Martinez, "A framework for selecting machine learning models using TOPSIS," in *Advances in Artificial Intelligence, Software and Systems Engineering* (Advances in Intelligent Systems and Computing). Cham, Switzerland: Springer, 2021, pp. 119–126.

[85] A. G. C. Pacheco and R. A. Krohling, "Ranking of classification algorithms in terms of mean–standard deviation using A-TOPSIS," *Ann. Data Sci.*, vol. 5, no. 1, pp. 93–110, Mar. 2018, doi: 10.1007/s40745-018-0136-5.

[86] R. Ihaka and R. Gentleman, "R: A language for data analysis and graphics," *J. Comput. Graph. Stat.*, vol. 5, pp. 299–314, Sep. 1996, doi: 10.1080/10618600.1996.10474713.

[87] Worldometer. *Norway Population (LIVE)*. Accessed: Mar. 8, 2022. [Online]. Available: https://www.worldometers.info/world-population/norway-population/

[88] M. Kuhn and K. Johnson, "An introduction to feature selection," in *Applied Predictive Modeling*. New York, NY, USA: Springer, 2013, pp. 487–519.

[89] X. Yang, X. Xie, D. L. Liu, F. Ji, and L. Wang, "Spatial interpolation of daily rainfall data for local climate impact assessment over greater Sydney region," *Adv. Meteorol.*, vol. 2015, Jul. 2015, Art. no. 563629, doi: 10.1155/2015/563629.

[90] Z. Khan, T. Zayed, and O. Moselhi, "Structural condition assessment of sewer pipelines," *J. Perform. Constructed Facilities*, vol. 24, no. 2, pp. 170–179, 2010, doi: 10.1061/(ASCE)CF.1943-5509.0000081.

LAM VAN NGUYEN received the B.Sc. and M.Sc. degrees in surveying and mapping engineering from the Hanoi University of Mining and Geology (HUMG), Vietnam, in 2011 and 2015, respectively. He is currently pursuing the Ph.D. degree with the Smart Water Laboratory, Norwegian University of Science and Technology (NTNU) in Ålesund Campus, Norway.

From 2001 to 2004, he was a Lecturer at the Department of Geodesy, Faculty of Geomatics and Land Administration, HUMG. His current project focuses on supporting the operational performance of sewers net by implementing machine learning algorithms for predictive maintenance. His research interests include geographic information system data processing, 3D visualization, and machine learning for wastewater/stormwater network maintenance.

DIEU TIEN BUI is currently a Full Professor at the GIS Group, Department of Business and IT, University of South-Eastern Norway (USN), Bø i Telemark, Norway. He has more than 200 publications, and out of them, more than 180 articles were published in science citation index (SCI/SCIE) indexed journals. His research interests include GIS and geospatial information science, remote sensing, and applied artificial intelligence and machine learning for natural hazards and environmental problems, such as landslide, flood, forest fire, ground biomass, and structural displacement.

RAZAK SEIDU received the Ph.D. degree in water and environmental engineering from the Norwegian University of Life Sciences (NMBU).

He is currently a Professor at the Department of Ocean Operations and Civil Engineering, Faculty of Engineering, Norwegian University of Science and Technology (NTNU), Ålesund Campus, Norway. He is the Leader of the Water and Environmental Engineering Research Group, NTNU. He has more than 15 years of experience in the water and sanitation sector. His research interests include smart water systems, water and wastewater treatment technologies, and stormwater modeling and management.

Paper III

Lam Van Nguyen and Seidu Razak (2023). Predicting sewer structural condition using hybrid machine learning algorithms. *Urban Water Journal*. https://doi.org/10.1080/1573062X.2023.2217430.

Urban Water Journal

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/nurw20

Predicting sewer structural condition using hybrid machine learning algorithms

L. V. Nguyen & S. Razak

To cite this article: L. V. Nguyen & S. Razak (2023): Predicting sewer structural condition using hybrid machine learning algorithms, Urban Water Journal, DOI: 10.1080/1573062X.2023.2217430

To link to this article: https://doi.org/10.1080/1573062X.2023.2217430

Published online: 30 May 2023.

URBAN WATER JOURNAL
https://doi.org/10.1080/1573062X.2023.2217430

Taylor & Francis
Taylor & Francis Group

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Predicting sewer structural condition using hybrid machine learning algorithms

L. V. Nguyen[a,b] and S. Razak[a]

[a]Department of Ocean Operations and Civil Engineering, Norwegian University of Science and Technology, Ålesund, Norway; [b]Department of Geodesy, Hanoi University of Mining and Geology, Hanoi, Vietnam

ABSTRACT

Predicting the structural condition of sewer pipes plays a vital role in the predictive maintenance of sewer pipes and renewal plans of many water utilities. This study explores the simultaneous utilization of physical and environmental features of sewer pipes in sewer structural condition prediction. Three (3) hybrid machine learning models which are the combination of Bagging (BG), Dagging (DG), and Rotation Forest (RotF) ensembles with a J48 Decision Tree (J48DT) based classifier were used to predict sewer pipe conditions in Ålesund city, Norway. The classification performance of the machine learning models was evaluated using the area under the receiver operating characteristic (AUC-ROC) and the area under the precision-recall (AUC-PRC) curves. The RotF-J48DT model had the highest (AUC-ROC = 0.857, AUC-PRC = 0.918) values, followed by the BG-J48DT, and the base classifier J48DT. The RotF-J48DT hybrid model should be considered when predicting the condition of sewer pipes in the study area.

ARTICLE HISTORY
Received 21 July 2022
Accepted 11 May 2023

KEYWORDS
sewer network; structural condition; machine learning; predictive maintenance

1. Introduction

A sewer system is an indispensable part of urban cities and plays a vital role in the collection and transport of wastewater and stormwater from residential and industrial areas to treatment plants. Sewer pipelines undergo different stresses during their lifespan, which contribute to their rapid deterioration with dire consequences for public health, property, and the environment (Sempewo and Kyokaali 2019). Wastewater usually contains chemical and microbial hazards that can escape through cracks in sewer pipelines and contaminate surrounding soil and groundwater (Farkas et al. 2020).

Failures in sewer systems have been reported in Europe. For example, the collapse of sewer pipes due to aging was reported in Paris and Bordeaux (Diab 2000). According to Kuliczkowska, Kuliczkowski, and Parka (2022), pipe failure is the main cause of street pavement collapse in residential areas. Venkatesh and Brattebø (2012) found the condition of wastewater pipelines in Oslo, Norway, to be poor due to deterioration, and were not able to perform their engineered functions effectively. Although there have been recent investments in sewer networks in Norway, some of these networks remain in poor condition with an annual renewal rate of 0.6% (RIF 2021). Many municipalities have therefore been looking for ways to intensify their renewal programs through the use of condition assessment methods, visual inspection techniques, and deterioration models (Fugledalen, Møller Rokstad, and Tscheikner-Gratl 2021).

The predictive ability of the structural condition and deterioration models significantly depends on the input variables. Based on the literature, the factors affecting the structural condition of sewer pipes are generally divided into physical, environmental, and operational factors (Mohammadi, Najafi, Vinayak, et al. 2019; Mohammadreza Malek Mohammadi et al. 2020). Age, diameter, material, depth, length, and slope are the most important physical factors (Hawari et al. 2020). The frequently used environmental factors are soil type, location, groundwater, and traffic volume (Laakso et al. 2018; Mohammadi, Najafi, Vinayak, et al. 2019; Mohammadreza Malek Mohammadi et al. 2020). Operational factors include preparation, cleaning, flow rate, infiltration and inflow, and pressure (Balekelayi and Tesfamariam 2019; Hawari et al. 2018). A recent review by Hawari et al. (2020) indicated that very few studies account for some environmental factors and operational factors as inputs of condition assessment models. Deterioration is a complex process and is a result of the interaction between several factors (Huu Dung Tran, 2007; Mohammadreza Malek Mohammadi et al. 2020). Therefore, considering more factors will provide more useful information in addressing sewer deterioration. One major challenge with this approach is the high redundancy and multicollinearity of features (input variables).

Reliable deterioration models enhance our understanding of the deterioration process and mechanism. This is critical for the evaluation of non-inspected pipe conditions and the forecast of their future state for rehabilitation strategies (Nicolas Caradot et al. 2017). This tool can help wastewater utilities to evaluate the non-inspected pipes' conditions and forecast the future state of the sewer pipes. Hawari et al. (2020) showed that exploring the relationship between the factors affecting the deterioration process is fundamental in building a good deterioration model for sewer condition prediction. Generally, deterioration models for sewer condition classification can be

CONTACT L. V. Nguyen lam.v.nguyen@ntnu.no

grouped into three (3) major categories namely physical, statistical, and machine learning methods (Mohammadi, Najafi, Vinayak, et al. 2019).

The physical models (or deterministic models) utilize only the physical properties and mechanics of sewer pipes to determine the extent of deterioration (Hawari et al. 2020; Tscheikner-Gratl et al. 2019). Several physical deterioration models have been reported in the literature. These include power function models (M. I. C. H. E. A. L. L. Doleac, Lackey, and Bratton 1980) and linear function models (Randall-Smith, Oliphant, and Russell 1992) to determine corrosion pit depth, UtilNets for reliability-based life prediction of buried grey cast iron in water mains (Hadzilacos et al. 2000), and ExtCorr for external corrosion estimation (Hawari et al. 2020). Hawari et al. (2020) argued that physical models are best suited for determining specific processes such as corrosion but are too simple to reflect a complex process such as deterioration. Additionally, suitable data types for physical modeling of deterioration are scarce and difficult to curate (E. V. Ana and Bauwens 2010). Some studies have proposed statistical models as an economic alternative to the physical models (Rajani and Kleiner 2001; Tscheikner-Gratl et al. 2019) to overcome some of the drawbacks.

Statistical models describe historical failure data's probabilistic nature as a random variable and estimate the best output (state) based on the condition of given data (Mohammadi, Najafi, Vinayak, et al. 2019). These statistical models include regression models (Bakry et al. 2016; Ngandu, Tesfamariam, and Asce 2019; Kabir et al. 2018; Mohammadrza Malek Mohammadi et al. 2019; Sempewo and Kyokaali 2019), Markov chains (Sempewo and Kyokaali 2019), cohort survival models (Nicolas Caradot et al. 2017), discriminant analysis (Vladeanu, Matthews, and Asce 2019; Alsaqqar, Hussein Khudair, and Karim Jbbar 2017), probabilistic models (Kleiner and Rajani 2001), and integrated methods (Kabir et al. 2018; Altarabsheh, Ventresca, and Kandil 2018; Hawari et al. 2016). Regression models are flexible and simple models for predicting the condition of sewer pipes that enhance interpretability vs explainability, however, the accuracy of these models can sometimes be low. Markov chains on the other hand create complex and chronological models with appreciable accuracy but determining the transitional probability matrix has always been a difficult challenge (Hawari et al. 2020). Additionally, the underlying assumption of normality is difficult to validate (Huu Dung Tran, 2007; Mohammadi, Najafi, Vinayak, et al. 2019). Machine learning (ML) models have been proposed as distribution-free alternatives to statistical models. These models include random forest (N. Caradot et al. 2018; Laakso et al. 2018; Vitorino et al. 2014), support vector machine (SVM) (Harvey and Arthur McBean 2014; H. D. Tran and Ng 2010), decision tree (Harvey and Arthur McBean 2014; Syachrani et al. 2013), or artificial neural network (El-Abbasy et al. 2014; H. D. Tran, Perera, and Ng 2009). These models explore the complex non-linear relationship between inputs and outputs (Hawari et al. 2020; Tsai, Miao-Ling, and Lin 2018).

Several previous studies have applied ML algorithms to study the sewer deterioration process. For example, Multinomial Logistic Regression and Artificial Neural Network models were developed to predict sanitary sewer pipes condition in a study by Atambo, Najafi, and Kaushal (2022), or Yin et al. (2020) used linear regression and a neural network to construct neighborhood-level and individual-level prediction models, respectively. However, a common point of these studies is that they only used single ML algorithms to construct sewer deterioration models, and hybrid ML models were not considered. Moreover, Tizmaghz, van Zyl, and Henning (2022) showed that each classification system exist weaknesses and no algorithm is perfect for all cases. Hence, finding a suitable sewer deterioration assessment model should be considered.

Many studies have shown that hybrid machine learning and metaheuristic algorithms are better than single ML methods because they enhance the capability of individual weak base algorithms to develop higher accuracy prediction models (Shirzadi et al. 2018, 2019). For instance, hybrid ensemble models outperformed a base classifier in the mapping of the groundwater potential zones or environmental hazards (Phong et al. 2021; Shahabi et al. 2020). The application of these kinds of hybrid ensemble ML models in the water field is still limited and seldom utilized, especially in predicting the structural condition of the sewage system.

The main objective of this study is to develop hybrid ensemble models for predicting the structural condition of sewer pipes using the physical and environmental factors affecting the deterioration process in Ålesund city, Norway. The hybrid ML model with higher performance (compared to the original ML model) can effectively support local water engineers, water managers, and relevant agencies in optimizing predictive maintenance strategies. Moreover, feature selection analysis in this study defines the most significant factors affecting sewer deterioration that provide useful information for local water agencies to prioritize their maintenance strategies. This study explores several hybrid ML models for predicting the condition of sewer pipelines. Specifically, the J48 Decision Tree (J48DT) algorithm is utilized as a base classifier and then combined with ensemble techniques namely Bagging (BG), Dagging (DG), and Rotation Forest (RotF) to develop hybrid ML models, namely BG-J48DT, DG-J48DT, and RotF-J48DT, for predicting the structural condition of sewer pipe in Ålesund city, Norway. Appropriate and suitable models for structural condition assessment will go a long way to help authorities and municipalities optimize maintenance strategies, reduce expenses, and strengthen the performance of the sewer network.

2. Materials and methods

2.1. Study area

The data used for this study were collected from the sewer network in Ålesund city, Norway. This city is located between longitudes 6°05′E and 6°42′E and latitudes 62°25′N and 62°32′E with an area of 633.6 km^2. The geographic location of the sewer network in Ålesund city is shown in Figure 1.

2.2. Data used

The sewer network consists of about 33,090 pipes with a total length of 760.4 km comprising concrete and polyvinyl chloride (PVC) as the main pipe materials. The condition of pipes was

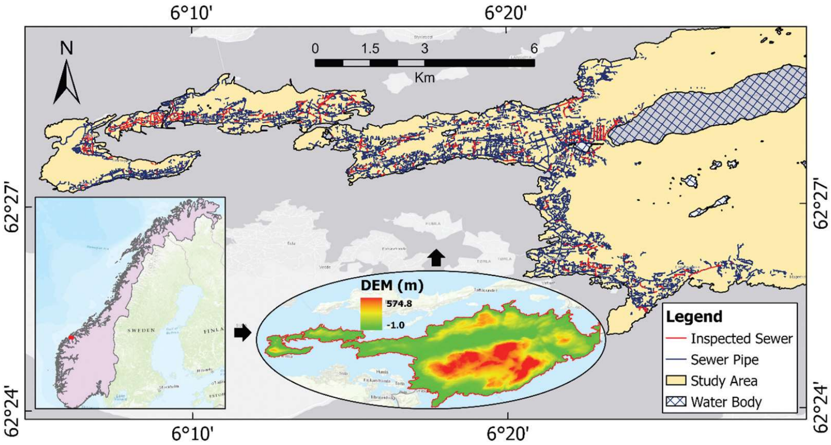


Figure 1. The sewer network in the study area.

Table 1. The condition classes of pipe.

Damage Class	Damage Score	State	Aggregated Class
Class 1	0–5	Very Good	Good Condition
Class 2	6–10	Good	
Class 3	11–20	Intermediate	Intermediate Condition
Class 4	21–50	Bad	Bad Condition
Class 5	>50	Very bad	

monitored by using the Closed-Circuit Television (CCTV) approach and damaged scores were assigned for sewer pipes to reflect their status (Nicolas Caradot et al. 2020; Bairaktaris et al. 2007). The status of sewer pipes was classified into five damaged classes based on the damage score obtained from the CCTV as shown in Table 1 (Haugen and Viak 2018). These damage classes were further grouped into three (3) aggregated classes namely good condition (class 1 or class 2), intermediate condition (class 3), and bad condition (class 4 or class 5). Previous studies have utilized these aggerated classes in sewer condition assessment since it facilitates comprehension of sewer condition states (Nicolas Caradot et al. 2020; Mohammadi, Najafi, Tabesh, et al.).

Sewer pipes without full properties in terms of physical or environmental values were eliminated from the database, this data was then compared to inspected data to select sewer pipes for training and validating models. As a result, a total of 1,335 inspected pipes were used in this study (Figure 2a). These data were curated via the CCTV method from 2012 to 2020. The physical attributes of the pipes of the selected pipes include age, diameter, depth, length, slope, pipe type, network type, pipe form, connection type, and material as shown in Table 2. The environmental factors considered for modeling the sewer pipe conditional process were calculated from auxiliary geospatial data. Detailed information is provided in Figure 3. These data were converted to pixels of 5 m × 5 m to prepare data for analyzing process. The map of rainfall is interpolated from monthly average precipitation from 9 weather stations within the Ålesund municipality using the inverse distance weighting method in ArcGIS Pro software.

2.3. Condition assessment

2.3.1. Boruta feature selection method

Condition assessment models of sewer pipes use multiple factors/features as independent variables. The current and future condition of any sewer pipe is a function of physical, environmental, and operational factors. Therefore, choosing significant factors before modeling is essential in reducing multicollinearity and redundancy amongst features. Many feature selection techniques have been proposed in the literature to assess the importance of independent variables. These feature selection methods include filter, wrapper, and embedded methods (Chandrashekar and Sahin 2014).

The Boruta algorithm, which is a wrapper method built around the Random Forest model, is a good candidate for dealing with both regression and classification problems (Kursa and Rudnicki 2010). Many studies used the Boruta for feature selection and showed that this algorithm is an effective method to reduce the dimensionality of the data set (Nanda et al. 2021; Bhavan and Aggarwal 2018). This algorithm distinguishes relevant variables or features into important, tentative, and unimportant categories based on a comparison of input variables' importance with output performance using a randomly permuted method. The main idea of the Boruta algorithm is to randomly create a copy of data, then classify the combination of copied versions with the original data. Then an iterative procedure is applied until every feature is classified as either important (accepted) or unimportant (rejected). The key steps for implementing the Boruta method are represented in Figure 4.

For a detailed description of the Boruta algorithm, readers are referred to (Kursa and Rudnicki 2010; Nanda et al. 2021). The Boruta method ranks the importance of each feature thereby eliminating unimportant factors and reducing multicollinearity before developing the sewer condition model.

2.3.2. Hybrid ensemble models

2.3.2.1. J48 decision tree classifier base model.

A decision tree is a classification model that comprises a root, decision

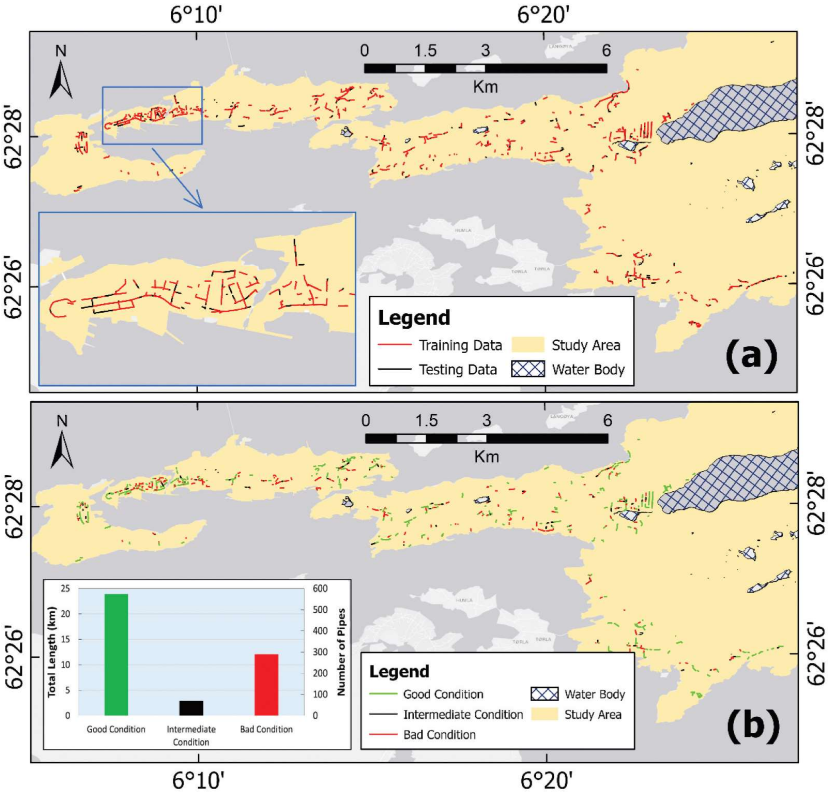


Figure 2. The maps of (a) Training and testing samples; (b) Condition class distribution.

Table 2. Input factors for building structural condition model for sewer pipe.

Physical factors	Type	Min	Max	Mean	Environmental factors	Spatial Resolution	GIS Data Type
Age	Numeric	1	104	32.31	Rainfall	-	Point
Diameter	Numeric	110	1000	251.07	Geology	1:50,000	Polygon
Depth	Numeric	−7.81	−0.01	−1.82	Landslide Area	1:5,000	Polygon
Length	Numeric	1.00	177.85	37.54	Population	250m × 250m	GRID
Slope	Numeric	−10.58	32.28	2.74	Land Cover	5m × 5m	GRID
Pipe Type	Categorical	-	-	-	Building Area	1:5,000	Polygon
Network Type	Categorical	-	-	-	Groundwater Level	-	Point
Pipe Form	Categorical	-	-	-	Traffic Volume	5m × 5m	GRID
Connection	Categorical	-	-	-	Distance to Road	5m × 5m	GRID
Material	Categorical	-	-	-	Soil Type	1:50,000	Polygon

nodes, leaf nodes, and branches (Bui et al. 2014). In a decision tree structure, one of the attributes represents a decision node and the class value is represented by a leaf node (Sahu and Mehtre 2015). The minimum number of instances per leaf and the confidence factor are two important user-defined parameters for building a decision tree classifier. A typical decision tree classifier is constructed in two steps: building and pruning. In the building step, the parameters influencing the classification accuracy of the decision tree are determined. In the pruning step, Laplace smoothing is used for probabilistic estimates of the leaves (Bui et al. 2014). Depending on the accuracy and efficiency desired, different algorithms can be used to generate decision trees. These dominant algorithms include Best First Tree (BFTree), Classification and Regression Trees (CART), Alternating Decision Tree (AD Tree), ID3, J48, and C4.5. A study by Lim, Loh, and Shih (2000) showed that the C4.5 family of algorithms represents the fastest algorithm for building decision trees with good accuracy. The J48 algorithm, which is slightly modified C4.5 in WEKA, is used in this study for building the decision tree base model. The steps for implementing the J48 Decision Tree (J48DT) are presented in Figure 5.

In Figure 5, $Entropy(Z)$ denotes the entropy of each attribute, and $Gain(Z, A)$ denotes the information gain of each split (Hilal et al. 2021); Z and A are represented the dataset and attributes, respectively; n and m are the number of partitions of A and the

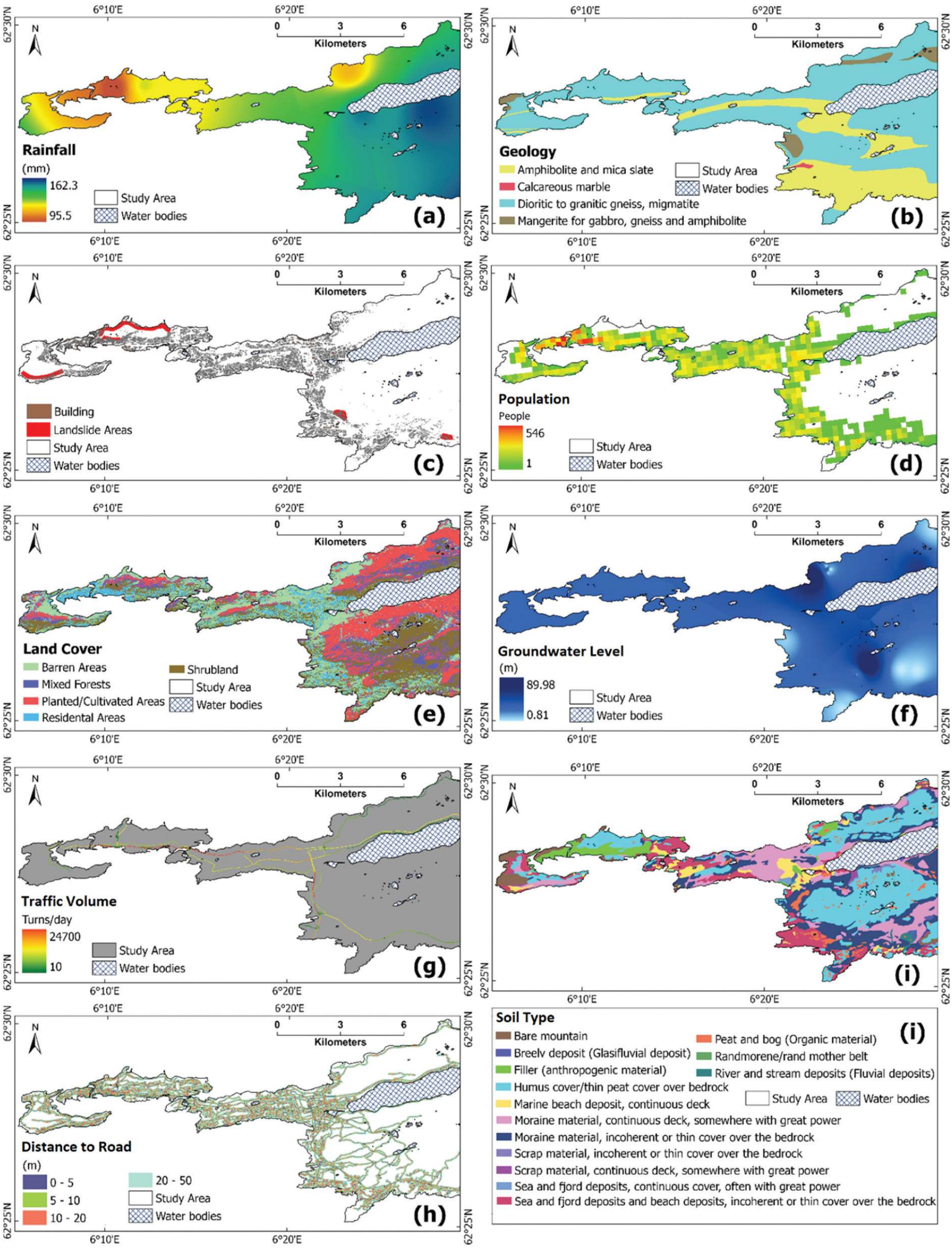

6°10'E
6°20'E
62°30'N
62°25'N
N
0 3 6
Kilometers
Rainfall
(mm)
162.3
95.5
Study Area
Water bodies
(a)
Geology
Amphibolite and mica slate
Calcareous marble
Dioritic to granitic gneiss, migmatite
Mangerite for gabbro, gneiss and amphibolite
Study Area
Water bodies
(b)
Building
Landslide Areas
Study Area
Water bodies
(c)
Population
People
546
1
Study Area
Water bodies
(d)
Land Cover
Barren Areas
Mixed Forests
Planted/Cultivated Areas
Residental Areas
Shrubland
Study Area
Water bodies
(e)
Groundwater Level
(m)
89.98
0.81
Study Area
Water bodies
(f)
Traffic Volume
Turns/day
24700
10
Study Area
Water bodies
(g)
Distance to Road
(m)
0 - 5
5 - 10
10 - 20
20 - 50
Study Area
Water bodies
(h)
(i)
Soil Type
Bare mountain
Breelv deposit (Glasifluvial deposit)
Filler (anthropogenic material)
Humus cover/thin peat cover over bedrock
Marine beach deposit, continuous deck
Moraine material, continuous deck, somewhere with great power
Moraine material, incoherent or thin cover over the bedrock
Scrap material, incoherent or thin cover over the bedrock
Scrap material, continuous deck, somewhere with great power
Sea and fjord deposits, continuous cover, often with great power
Sea and fjord deposits and beach deposits, incoherent or thin cover over the bedrock
Peat and bog (Organic material)
Randmorene/rand mother belt
River and stream deposits (Fluvial deposits)
Study Area
Water bodies
(i)

Figure 3. The maps of environment-related factors.

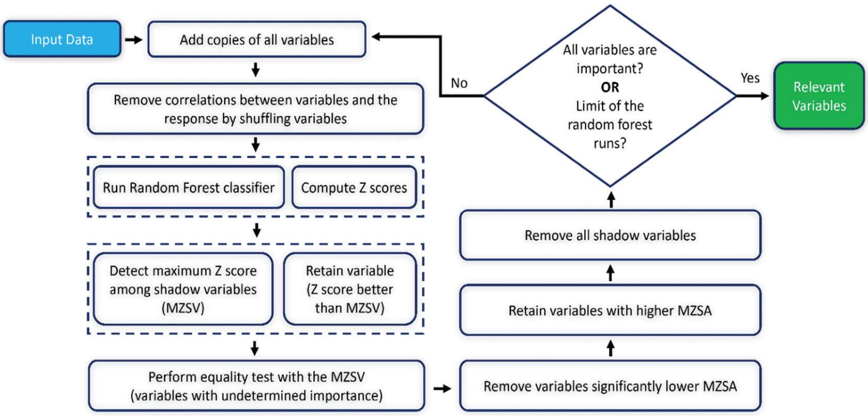

Input Data
Add copies of all variables
Remove correlations between variables and the response by shuffling variables
Run Random Forest classifier
Compute Z scores
Detect maximum Z score among shadow variables (MZSV)
Retain variable (Z score better than MZSV)
Perform equality test with the MZSV (variables with undetermined importance)
Remove variables significantly lower MZSA
Retain variables with higher MZSA
Remove all shadow variables
All variables are important? OR Limit of the random forest runs?
No
Yes
Relevant Variables

Figure 4. Framework for implementing the Boruta feature selection method.

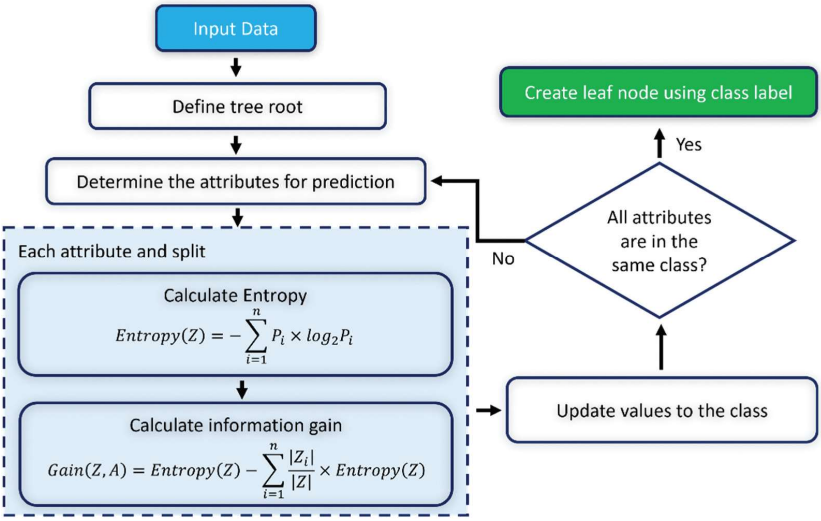

Input Data
Define tree root
Determine the attributes for prediction
Each attribute and split
Calculate Entropy
$Entropy(Z) = -\sum_{i=1}^{n} P_i \times log_2 P_i$
Calculate information gain
$Gain(Z,A) = Entropy(Z) - \sum_{i=1}^{n} \frac{|Z_i|}{|Z|} \times Entropy(Z)$
Update values to the class
All attributes are in the same class?
No
Yes
Create leaf node using class label

Figure 5. J48 Decision Tree overview.

number of classes respectively; Z_i and P_i represent the number of cases on partition *i* and the proportion of Z_i to *Z*, respectively.

2.3.2.2. Bagging ensemble. Bagging (Bootstrap Aggreg ating) was proposed by Breiman (1996) to raise the stability of models significantly classification problems by improving accuracy and reducing variance. There are three main steps implemented in this model:

- Creating multiple datasets: new sewer pipe points are created by randomly selecting samples with replacement (e.g. the individual sewer data points can be chosen more than one time) from the original training dataset.
- Building multiple J48DT classifiers: the J48DT algorithm is used to independently train using random subsets from the previous step. Each J48DT will predict sewer condition status from the subset.
- Combining classifier: the sewer condition status predictions of all the individual J48DT classifiers are combined to give a better classifier, usually with less variance compared to before. Finally, the final sewer condition status is defined using a plurality vote of those predictions from the J48DT models.
- The concept of the bagging ensemble method is shown in Figure 6.

2.3.2.3. Dagging ensemble. Ting and Witten (1997) proposed Dagging (Disjoint Aggregating) method to create random training subsets from the original training dataset using

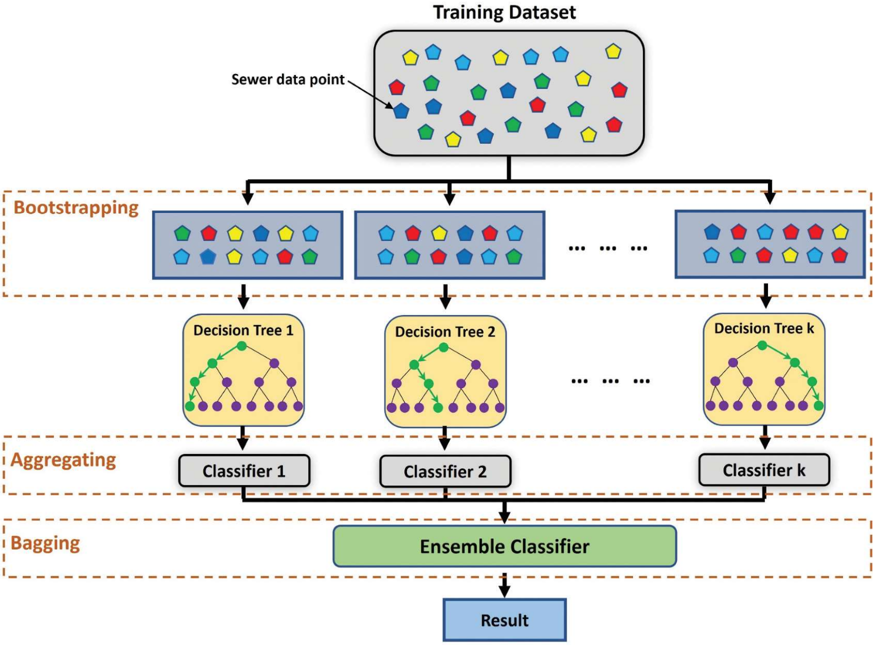


Figure 6. The framework of the Bagging ensemble method.

the disjoint sampling method (instead of the bootstrap sampling) without replacement.

The main steps for implementing the dagging ensemble are described as follows:

- New sewer pipe points are randomly created from the original training dataset without replacement (e.g. the individual sewer data points can be chosen only one time).
- Prediction of sewer status condition from each subset is obtained using the J48DT classifier.
- The plurality vote is used to aggregate results from the individual predictions obtained from each J48DT classifier, and the final sewer condition status is defined for each sewer data point.

2.3.2.4. *Rotation forest ensemble.*

Rotation Forest (RotF) was firstly introduced by Rodriguez, Kuncheva, and Alonso (2006) based on the idea of a random forest algorithm to improve the diversity and accuracy of the base classifier. In this method, the base classifiers (decision trees) are independently built and trained using the whole training dataset in a rotated feature space. Hyperplanes parallel to the feature axes are used to create classification regions while training decision trees and the final sewer condition status is computed based on the largest confidence for each status (Kuncheva and Rodríguez 2007).

Assume $x = [x_i, x_2, \ldots, x_m]^T$ is a vector that contains m factors of sewer pipes, $y = [y_1, y_2, \ldots, y_n]^T$represents sewer condition status vector where y_i is the sewer condition (i.e. good condition, intermediate condition, and bad condition) of the i^{th} sewer pipe and n is the number of inspected sewer pipes in the training dataset. Let K and F be the number of subsets and feature set of classifiers $D_1, D_2, \ldots, D_L$ in the ensemble (L is the number of classifiers in the ensemble), respectively, and X is the objects in the training dataset. Rodriguez, Kuncheva, and Alonso (2006) introduced steps for constructing the training dataset for classifier D_i as follows:

- For each classifier in the ensemble, randomly split the feature set F into K subsets: $F_{i,j}(i = 1, 2, \ldots, L, ; j = 1, 2, \ldots, K)$.
- For each subset $F_{i,j}$, let $X_{i,j}$ be the dataset X for the feature and eliminate from $X_{i,j}$ a random subset of classes, and randomly select a bootstrap sample from $X_{i,j}$ of size 75% of the data count. Run Principal Component Analysis (PCA) on the $M = m/K$ features and the selected subset of X. Store the coefficients of the principal components $a^1_{i,j}, a^2_{i,j}, \ldots, a^{M_j}_{i,j}$.
- Arrange the obtained vector with coefficients in a spare 'rotation' matrix R_i:

$$R_i = \begin{bmatrix} a^1_{i,1}, a^2_{i,1}, \ldots, a^{M_1}_{i,1} & [0] & \ldots & [0] \\ [0] & a^1_{i,2}, a^2_{i,2}, \ldots, a^{M_2}_{i,2} & \ldots & [0] \\ \ldots & \ldots & \ldots & \ldots \\ [0] & [0] & \ldots & a^1_{i,K}, a^2_{i,K}, \ldots, a^{M_K}_{i,K} \end{bmatrix} \quad (1)$$

- Rearrange the columns to match the order of features in F for constructing a rotation matrix R_i^a and build classifier D_i using the training set created by $\left(XR_i^a, y\right)$.
- Calculate the confidence of each class from the input x ($\mu_j(x)$) by the average combination method and assign x to the class with the largest confidence:

$$\mu_j(x) = \frac{1}{L}\sum\nolimits_{i=1}^{L} d_{i,j}\left(xR_i^a\right), j = 1, 2, \ldots, c \qquad (2)$$

The process for the RotF method is shown in Figure 7.

2.4. Model performance assessment

The performance of a classification model is usually evaluated using different metrics such as accuracy, sensitivity, specificity, F1-score, Matthew's Correlation Coefficient (MCC), Geometric Mean (GM), or graphical assessment methods e.g. receiver operating characteristics and/or precision-recall curves. The above-mentioned metrics are derived from a confusion matrix (Tharwat 2021). This matrix is the basic component for calculating model performance assessment metrics in binary classification problems and multi-class classification problems.

2.4.1. Confusion matrix

The confusion matrix is an $n \times n$ matrix whose elements C_{ij} correspond to the number of classes in grade i that are predicted to be in grade $(i, j \in \{1, \ldots, n\}), n$ is the number of classes. Table 3 shows the confusion matrix for a multi-class classification problem with 3 classes. Elements on the diagonal of the confusion matrix represent the number of samples that are correctly classified. Off-diagonal elements are the number of a sample that are incorrectly predicted.

Based on the confusion matrix, various model performance metrics including false negative (FN), false positive (FP), true positive (TP), and true negative (TN) can be computed as follows (Tharwat 2021):

$$FN_i = \sum\nolimits_{j=1}^{3} C_{ij}; (j \neq i) \qquad (3)$$

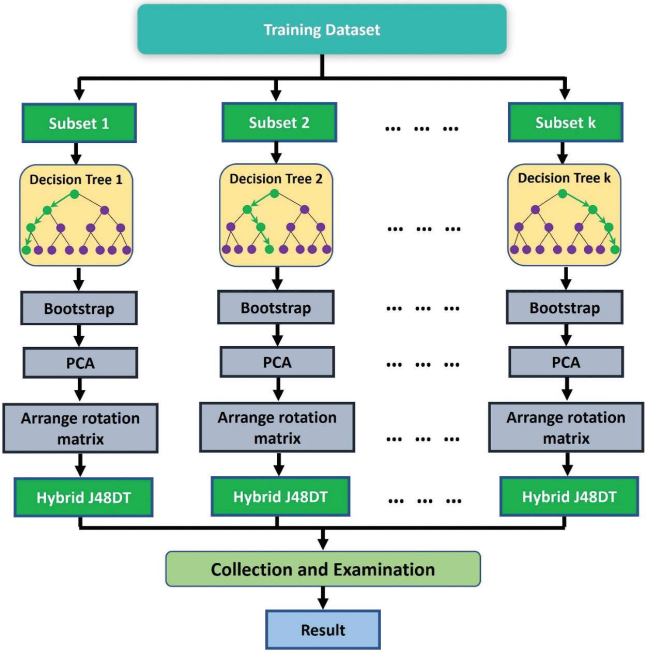


Figure 7. The framework of the rotation forest method.

Table 3. The confusion matrix for 3-class classification.

		Predicted Class		
		Good Condition	Intermediate Condition	Bad Condition
Actual Class	Good Condition	C_{11}	C_{12}	C_{13}
	Intermediate Condition	C_{21}	C_{22}	C_{23}
	Bad Condition	C_{31}	C_{32}	C_{33}

$$FP_i = \sum_{j=1}^{3} C_{ji};\ (j \neq i) \tag{4}$$

$$TP_i = C_{ii} \tag{5}$$

$$TN_i = \sum_{j=1}^{3} \sum_{k=1}^{3} C_{kj};\ (j, k \neq i) \tag{6}$$

- **Accuracy** (*Acc*): This metric is defined as a ratio between the number of correctly classified samples and the total number of samples. It is one of the most used measures for assessing classification performance. Based on the confusion matrix, the accuracy of the classification model is calculated as follows:

$$Acc = \frac{\sum_{i=1}^{3} C_{ii}}{\sum_{i=1}^{3} \sum_{j=1}^{3} C_{ij}} \tag{7}$$

- **F1-score**: This metric represents the harmonic mean of precision and recall, the value of 1 represents the highest classification performance and the value of 0 is the worst.

$$F1 - score_i = \frac{2TP_i}{2TP_i + FP_i + FN_i} \tag{8}$$

- **Geometric Mean (GM:)**This metric is the root of the product of class-wise sensitivity. For multi-class problems, this metric is a higher root of the product of the sensitivity of each class. It is most often used for evaluating the performance of classification with imbalanced data. The equation for calculating GM is described as follows:

$$GM_i = \sqrt{\frac{TP_i}{TP_i + FN_i} \times \frac{TN_i}{FP_i + TN_i}} \tag{9}$$

- **Matthew's Correlation Coefficient (MCC)**: This metric represents the correlation between the predicted and actual classifications. The coefficient of + 1 and −1 represents perfect and bad predictions, respectively. The value of zero represents a random prediction. The below equation describes the MCC formula for multiclassification:

$$MCC_i = \frac{TP_i \times TN_i - FP_i \times FN_i}{\sqrt{(TP_i + FP_i) \times (TP_i + FN_i) \times (TN_i + FP_i) \times (TN_i + FN_i)}} \tag{10}$$

Table 4. The model performance is based on the AROC values (Kritikos and Davies, 2015.).

AUC-ROC value	Model performance
<0.7	Poor
0.7–0.8	Satisfactory
0.8–0.9	Good
0.9–1.0	Excellent

2.4.2. Receiver operating characteristic

The Receiver Operating Characteristic (ROC) curve represents the relationship between sensitivity and specificity (Tharwat 2021). Each point in the ROC curve is generated by changing the threshold on the confidence score. The AUC-ROC is a threshold-independent metric that calculates the area under the ROC curve. The AUC-ROC score is in the range of zero to one and the ROC curve that has a larger AUC-ROC value will have better classification performance with the same class. Table 4 shows the success rate based on the AUC-ROC (Kritikos and Davies 2015).

2.4.3. Precision-recall curve

The Precision-Recall curve (PRC) represents the relationship between recall and precision. The PRC has been considered an alternative to the Receiver Operating Characteristics (ROC) curve for classification problems that have a large skew in the class distribution (Davis and Goadrich 2006). The AUC-PRC is a threshold-independent metric that calculates the area under the PRC curve.

2.5. Structural condition modeling framework

The data processing procedure for assessing sewer pipe status can be divided into several steps: 1) Collecting and pre-processing data, 2) Splitting data into training and testing data sets, 3) Building hybrid ML models, and 4) Validating and selecting structural conditions models. The flowchart for this procedure is shown in Figure 8.

To build and validate the structural condition models, the data was split into training, cross-validation, and testing datasets. The ratio for splitting training and testing datasets depends on the quantum of data available and the objectives of the study. In this study, we randomly split the dataset with a ratio of 70% for the training dataset and cross-validation (934 samples) and 30% for the testing dataset (401 samples) respectively. In the training and cross-validation dataset, the number of sewer pipes in good condition, intermediate condition, and

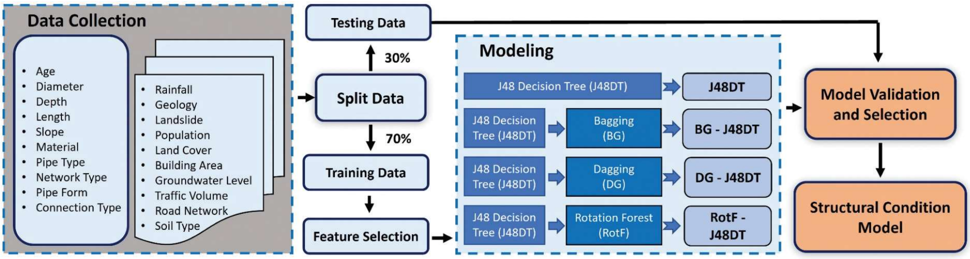


Figure 8. The framework for modeling the structural condition of sewer pipes.

bad condition are 574, 70, and 290 samples, respectively. The test dataset consists of 248 samples in good condition, 33 samples in intermediate condition, and 120 samples in bad condition.

A base classifier was created based on every sample dataset and several classifiers were obtained based on the training dataset. Finally, the outputs of individual classifiers are amalgamated via the voting process. The hybrid models use the J48 algorithm as the base classifier. In an attempt to avoid the subjective character of the hybrid models, the user-defined parameters (n_{inst}) and(n_{conf}) of the base classifier were not varied when applying ensembles. The optimal values for the user-defined parameters (n_{inst}) and(n_{conf}) were found using a grid search with 10-fold cross-validation. The best values for (n_{inst}) and (n_{conf}) are found as 2.0 and 0.25, respectively.

3. Results and discussions

3.1. Feature selection using the Boruta method

The importance of the features (variables) utilized in this study for sewer structural condition prediction is presented in Figure 9. The result showed that the material was adjudged as the most important factor affecting the structural deterioration process, followed by the age of the pipe. Six variables (landslide area, land cover, network type, building area, pipe form, and length) were unimportant for deterioration modeling and these factors were eliminated from the model. One tentative factor (geology) and the remaining factors were used to develop the structural condition models for the study area.

Yin et al. (2020) only used backward direction feature selection to eliminate insignificant factors and indicated that more advanced selection methods should be considered in the future to find significant factors. The significant values (p-value) from the logistic regression model were used to rank the important degree of factors in the study of Atambo, Najafi, and Kaushal (2022), this approach may be not an ideal solution in case an unbalanced dataset (Sanchez-Pinto et al. 2018). This study partly fills the above limitation by using the advanced wrapper method for defining the importance of input factors and eliminating insignificant factors before constructing ML models.

In this study, the Boruta feature selection method highlighted the material and age of the sewer pipelines as significant factors for modeling the structural condition of the pipes. This conclusion agrees with various studies in the literature (Mohammadi, Najafi, Tabesh, et al. ; Salman and Salem 2012; Baur and Herz 2002). For example, Mohammadi, Najafi, Tabesh, et al. () showed that pipe age was the most important factor, followed by the material and diameter of the pipe when building the condition prediction model of sanitary sewer pipe by applying the logistic regression model. The importance and absolute ranking of factors depend to some extent on the method utilized and local conditions. For example, Najafi and Kulandaivel (2005) concluded the diameter of the sewer was the most important based on an ANN model; whereas, pipe material was pointed as the most important factor based on the Back Propagation Neural Network and Probabilistic Neural Network models (Khan, Zayed, and Moselhi 2010). The age of sewer pipes was proved as the most important based on the binary logistic regression model in the study by Mohammadi, Najafi, Tabesh, et al. ().

In Ålesund city, sewer material significantly affects deterioration behaviors. Many sewer pipes in the study area are polyvinyl chloride (PVC) and concrete (BET), and these materials have a strong correlation with the conditioning process. For instance, PVC pipes are highly resistant to acidic and alkaline wastes and BET pipes work well with abrasion (Mohammadreza Malek Mohammadi et al. 2020). Additionally, installation and operation procedures ensure BET pipes are less affected by

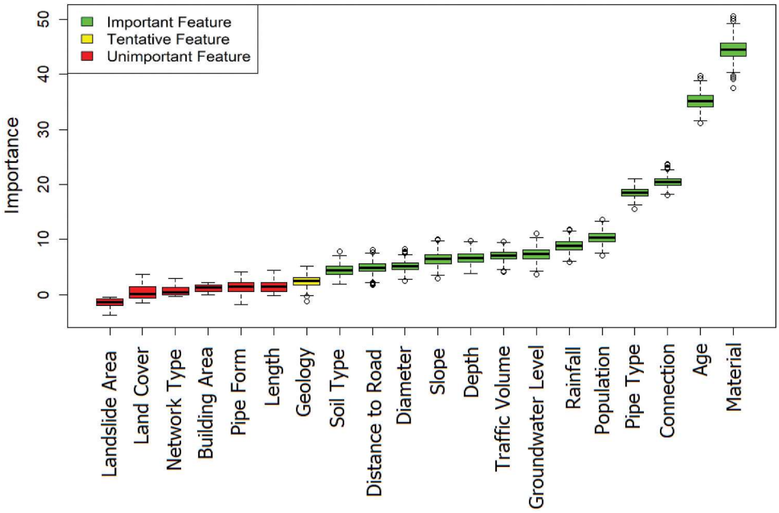


Figure 9. Feature selection for building conditional assessment model.

deterioration. More specifically, because of installation in controlled situations, BET pipes normally keep high quality and have good integrity. Moreover, reinforced steel in BET pipes makes them strong enough against structural deterioration and PVC pipes suffer excessively from traffic loads (Mohammadreza Malek Mohammadi et al. 2020), this mechanism is insignificant in this study as most pipes were not impacted by roads with high traffic volumes (Figure 3g).

The study shows the age of the pipe is an important factor in the deterioration process (Figure 9). This finding has been made in previous studies (Mohammadi, Najafi, Tabesh, et al. ; Khan, Zayed, and Moselhi 2010). The effect of aging on the condition of the sewer pipe begins immediately after the pipe is installed and it normally takes 44–65 years for pipes to change to poor condition (Laakso et al. 2018).

Although landslide directly affects the underground assets in general and the sewer network, it is not significant in this study. This can be explained by the landslide areas being small and the number of inspected sewer pipes in these areas not being significant (Figure 3c). In the same vein, the building area has been assessed to be a less important factor affecting the structural condition of sewer pipes in this study. This can be explained by the fact that most of the inspected pipes in the central area of Ålesund city were in good condition (Figure 2a) although found in locations of high building density (Figure 3c).

The length of the sewer pipes is considered an unimportant factor for this study area, this conclusion is in line with the result of Lubini and Fuamba (2011), in which the authors found the slope and length were not significant in their deterioration model.

3.2. Comparison of structural condition models

The predictive capability of the models for the structural condition of sewer pipes is assessed using the test dataset. The results of the statistical measures of the J48DT, BG-J48DT, DG-J48DT, and RotF-J48DT are represented in Table 5. Because there is a difference between the number of samples in the three output classes (class imbalance), accuracy may not be a reliable metric for assessing the overall classification performance (Haixiang et al. 2017). Therefore, some statistical measures such as GM, MCC, F1-score, AUC-ROC, and AUC-PRC have been used as alternative performance assessment metrics.

The results indicate that the statistical measures including GM, MCC, and F-measure have the highest values for good condition, which is a major class in the dataset, followed by the bad condition class. It can be seen that the MCC value of the DG-J48DT models is immeasurable when predicting samples in an intermediate condition indicating the bad performance for this class.

Figure 10 shows the AUC-ROC and AUC-PRC of four developed structural condition models in this study. Based on the definition of AUC values in Table 4, the developed ML models have good performance in predicting the structural condition of the sewer pipe in good and bad condition classes, but they have satisfactory performance in predicting samples in the intermediate condition class.

The results show that all three developed hybrid models improve classification performance compared to the base classifier (the J48DT model). More specifically, for the good condition class, the RotF-J48DT model have the highest classification performance (AUC-ROC = 0.857, AUC-PRC = 0.918), followed by BG-J48DT (AUC-ROC = 0.848, AUC-PRC = 0.907), DG-J48DT (AUC-ROC = 0.817, AUC-PRC = 0.880), and J48DT (AUC-ROC = 0.800, AUC-PRC = 0.765). Similarly, for the bad condition class prediction, the RotF-J48DT (AUC-ROC = 0.829, AUC-PRC = 0.635) is outperformed the BG-J48DT (AUC-ROC = 0.813, AUC-PRC = 0.574), DG-J48DT (AUC-ROC = 0.793, AUC-PRC = 0.577), and J48DT (AUC-ROC = 0.749, AUC-PRC = 0.522) models. All developed models have the lowest classification performance in predicting samples in the minor class (intermediate condition). However, the RotF-J48DT model (AUC-ROC = 0.673, AUC-PRC = 0.153) is better than the DG-J48DT (AUC-ROC = 0.637, AUC-PRC = 0.103), BG-J48DT (AUC-ROC = 0.586, AUC-PRC = 0.127), and J48DT (AUC-ROC = 0.522, AUC-PRC = 0.110) models. Additionally, in terms of weighted average values, the RotF-J48DT has higher values than other models indicating better classification performance. In conclusion, these ensemble models have better predictive power than the basic model. The RotF-J48DT ensemble model produces the best result for predicting the structural condition of sewer pipes in the study area.

Based on the area under the curve values, the hybrid models have higher performance compared to the base classifier. This conclusion is consistent with previous findings. For example, Miraki et al. (2019) showed that the novel classifier ensemble model, namely the Random Forest Classifier based on Random Subspace Ensemble, had

Table 5. Statistical measures of developed hybrid models in this analysis.

		Classes			
Structural Condition Models	Statistical Measures	Good Condition	Intermediate Condition	Bad Condition	ACC (%)
J48DT	GM	0.74	0.24	0.72	
	MCC	0.49	0.05	0.44	70.83
	F1-score	0.81	0.09	0.61	
BG-J48DT	GM	0.72	0.17	0.69	
	MCC	0.49	0.02	0.42	71.07
	F1-score	0.82	0.05	0.59	
DG-J48DT	GM	0.70	0.00	0.72	
	MCC	0.43	-	0.44	70.57
	F1-score	0.80	0.00	0.61	
RotF-J48DT	GM	0.70	0.17	0.70	
	MCC	0.44	0.02	0.42	70.07
	F1-score	0.80	0.05	0.60	

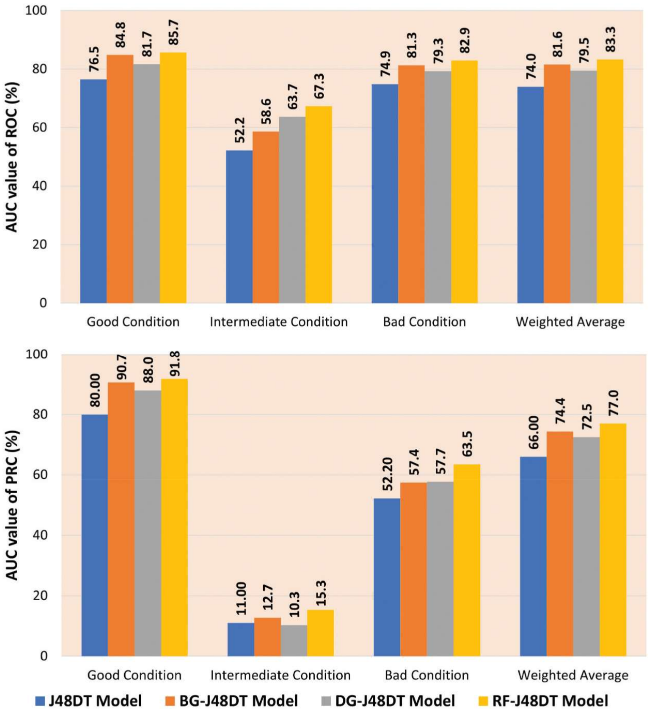


Figure 10. The AUC values of the developed models: (a) AUC-ROC, (b) AUC-PRC.

a higher predictive capability for groundwater potential mapping compared to other benchmark models. Additionally, the findings of Chen et al. (2019) indicated that two ensemble frameworks, namely Random subspace, and Bagging, produce a higher predictive performance than the base classifier, namely Reduced-error pruning trees. A similar conclusion was illustrated in the study by Phong et al. (2021) which shows the RealAdaBoost, Bagging, and Rotation Forest ensembles outperformed the functional tree base classifier.

Finally, the structural condition models developed in this study had a lower capability in predicting samples in the intermediate condition class (Figure 10). This can be explained that this is the minor class in the dataset (about 7.7% in the total of 957 training samples). Several studies have transformed the multi-classification problems into binary classification problems by clustering samples in classes 1–3 into one class (good condition) and the remaining samples into another class (bad condition) to improve classification performance (Mohammadi, Najafi, Tabesh, et al. ; E. Ana et al. 2009). In our study, we try to keep the basic characteristic of classes by converting five-grade scales into three-grade scales. This still allows water managers to correctly assess the importance of each class (pipes in class 1 and class 2 are good conditions, and pipes in class 4 and class 5 are bad conditions) and improves the classification performance of the structural condition models.

4. Conclusions

In this study, three ML hybrid models namely BG-J48DT, DG-J48DT, and RotF-J48DT based on the J48DT base classifier were investigated to predict the structural condition of sewer pipes in Ålesund city, Norway. The importance of input factors for modeling was assessed by applying the Boruta feature selection technique.

The results show that the material of sewer pipes is the most important factor affecting the structural condition of sewer pipes in the study area, followed by the age of the pipes. The landslide area, land cover, network type, building area, pipe form, and length of sewer pipe have the least influence on the structural condition in the study area.

Many model performance assessment measures including GM, MCC, F-Measure, AUC-ROC and AUC-PRC curves were used to evaluate the classification performance of the developed models. The three ensemble models have shown better prediction capability compared to the J48DT base classifier. The RotF-J48DT ensemble model is better at predicting all three condition classes comparing the remaining other ML models.

Although the ensemble models perform more effectively than the base classifier in predicting the structural condition of sewer pipes, the accuracy of these models is still limited (about 70%). Therefore, other ML models need to be investigated to improve classification performance and accuracy.

Acknowledgements

The authors would like to thank Ålesund Municipality for providing the data for this research. This research was funded by the Smart Water Project (No.: 90392200), financed by Ålesund Municipality and the Norwegian University of Science and Technology (NTNU).

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the Smart Water Project [90392200].

References

Alsaqqar, Awatif Soaded, Basim Hussein Khudair, and Rehab Karim Jbbar. 2017. "Rigid Trunk Sewer Deterioration Prediction Models Using Multiple Discriminant and Neural Network Models in Baghdad City, Iraq." *Journal of Engineering* 23 (8): 70–83. https://doi.org/10.31026/j.eng.2017.08.06.

Altarabsheh, Ahmad, Mario Ventresca, and Amr Kandil. 2018. "New Approach for Critical Pipe Prioritization in Wastewater Asset Management Planning." *Journal of Computing in Civil Engineering* 32 (5): 04018044–. https://doi.org/10.1061/(ASCE)CP.1943-5487.0000784.

Ana, E. V., and W. Bauwens. 2010. "Modeling the Structural Deterioration of Urban Drainage Pipes: The State-Of-The-Art in Statistical Methods." *Urban Water Journal* 7 (1): 47–59. https://doi.org/10.1080/15730620903447597.

Ana, E., W. Bauwens, M. Pessemier, C. Thoeye, S. Smolders, I. Boonen, and G. De Gueldre. 2009. "An Investigation of the Factors Influencing Sewer Structural Deterioration." *Urban Water Journal* 6 (4): 303–312. https://doi.org/10.1080/15730620902810902.

Atambo, Daniel Ogaro, Mohammad Najafi, and Vinayak Kaushal. 2022. "Development and Comparison of Prediction Models for Sanitary Sewer Pipes Condition Assessment Using Multinomial Logistic Regression and Artificial Neural Network." *Sustainability* 14 (9): 5549. https://doi.org/10.3390/su14095549.

Bairaktaris, D., V. Delis, C. Emmanouilidis, S. Frondistou-Yannas, K. Gratsias, V. Kallidromitis, and N. Rerras. 2007. "Decision-Support System for the Rehabilitation of Deteriorating Sewers." *Journal of Performance of Constructed Facilities* 21 (3): 240–248. https://doi.org/10.1061/(ASCE)0887-3828(2007)21:3(240).

Bakry, Ibrahim, Hani Alzraiee, Khalid Kaddoura, El Masry Mohamed, Tarek Zayed, and M. Asce. 2016. "Condition Prediction for Chemical Grouting Rehabilitation of Sewer Networks." *Journal of Performance of Constructed Facilities* 30 (6): 04016042–. https://doi.org/10.1061/(ASCE)CF.1943-5509.0000893.

Balekelayi, Ngandu, and Solomon Tesfamariam. 2019. "Statistical Inference of Sewer Pipe Deterioration Using Bayesian Geoadditive Regression Model." *Journal of Infrastructure Systems* 25 (3): 04019021. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000500.

Baur, R., and R. Herz. 2002. "Selective Inspection Planning with Ageing Forecast for Sewer Types." *Water Science & Technology* 46 (6–7): 389–396. https://doi.org/10.2166/wst.2002.0704 .

Bhavan, A., and S. Aggarwal. 2018. Stacked Generalization with Wrapper-Based Feature Selection for Human Activity Recognition. Paper presented at the 2018 IEEE Symposium Series on Computational Intelligence (SSCI). Bangalore, India: https://ieeexplore.ieee.org/document/8628830, 18-21 Nov. 2018.

Breiman, Leo. 1996. "Bagging Predictors." *Machine Learning* 24 (2): 123–140. https://doi.org/10.1007/BF00058655.

Bui, Dieu Tien, Biswajeet Pradhan, Inge Revhaug, and Chuyen Trung Tran. 2014. "A Comparative Assessment Between the Application of Fuzzy Unordered Rules Induction Algorithm and J48 Decision Tree Models in Spatial Prediction of Shallow Landslides at Lang Son City, Vietnam." In *Remote Sensing Applications in Environmental Research*, edited by Prashant K. Srivastava, Saumitra Mukherjee, Manika Gupta, and Tanvir Islam, pp. 87–111. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-05906-8_6.

Caradot, N., M. Riechel, M. Fesneau, N. Hernandez, A. Torres, H. Sonnenberg, E. Eckert, N. Lengemann, J. Waschnewski, and P. Rouault. 2018. "Practical Benchmarking of Statistical and Machine Learning Models for Predicting the Condition of Sewer Pipes in Berlin, Germany." *Journal of Hydroinformatics* 20 (5): 1131–1147. https://doi.org/10.2166/HYDRO.2018.217.

Caradot, Nicolas, Mathias Riechel, Pascale Rouault, Antoine Caradot, Nic Lengemann, Elke Eckert, Alexander Ringe, François Clemens, and Frédéric Cherqui. 2020. "The Influence of Condition Assessment Uncertainties on Sewer Deterioration Modelling." *Structure and Infrastructure Engineering* 16 (2): 287–296. https://doi.org/10.1080/15732479.2019.1653938.

Caradot, Nicolas, Hauke Sonnenberg, Ingo Kropp, Alexander Ringe, Stephane Denhez, Andreas Hartmann, and Pascale Rouault. 2017. "The Relevance of Sewer Deterioration Modelling to Support Asset Management Strategies." *Urban Water Journal* 14 (10): 1007–1015. https://doi.org/10.1080/1573062X.2017.1325497.

Chandrashekar, Girish, and Ferat Sahin. 2014. "A Survey on Feature Selection Methods." *Computers & Electrical Engineering* 40 (1): 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024.

Chen, Wei, Haoyuan Hong, Shaojun Li, Himan Shahabi, Yi Wang, Xiaojing Wang, and Baharin Bin Ahmad. 2019. "Flood Susceptibility Modelling Using Novel Hybrid Approach of Reduced-Error Pruning Trees with Bagging and Random Subspace Ensembles." *Journal of Hydrology* 575: 864–873. https://doi.org/10.1016/j.jhydrol.2019.05.089 .

Davis, Jesse, and Mark Goadrich. 2006. "The Relationship Between Precision-Recall and ROC Curves." In *Proceedings of the 23rd International Conference on Machine Learning*, 233–240. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery.

Diab, Y. G. 2000. "Maintenance of Urban Sewers in Europe: Diversity of Approaches, Tools and Strategies." National Conference on Environmental and Pipeline Engineering July 23-26, 2000. Kansas City, Missouri, United States. 33–39. https://doi.org/10.1061/40507(282)5.

Doleac, M. I. C. H. E. A. L. L., S. L. Lackey, and G. N. Bratton. 1980. Prediction of Time-To Failure for Buried Cast Iron Pipe. Paper presented at the Proceedings of American water works association annual conference. Denver, Colorado, United States. 21–28.

El-Abbasy, Mohammed S., Ahmed Senouci, Tarek Zayed, Farid Mirahadi, and Laya Parvizsedghy. 2014. "Artificial Neural Network Models for Predicting Condition of Offshore Oil and Gas Pipelines." *Automation in Construction* 45: 50–65. https://doi.org/10.1016/J.AUTCON.2014.05.003.

Farkas, Kata, Luke S. Hillary, Shelagh K. Malham, James E. McDonald, and David L. Jones. 2020. "Wastewater and Public Health: The Potential of Wastewater Surveillance for Monitoring COVID-19." *Current Opinion in Environmental Science & Health* 17: 14–20. https://doi.org/10.1016/j.coesh.2020.06.001.

Fugledalen, Thomas, Marius Møller Rokstad, and Franz Tscheikner-Gratl. 2021. "On the Influence of Input Data Uncertainty on Sewer Deterioration Models – a Case Study in Norway." *Structure and Infrastructure Engineering* 19 (8): 1064–1075. https://doi.org/10.1080/15732479.2021.1998142.

Hadzilacos, T., D. Kalles, N. Preston, P. Melbourne, L. Camarinopoulos, M. Eimermacher, V. Kallidromitis, S. Frondistou-Yannas, and S. Saegrov. 2000. "UtilNets: A Water Mains Rehabilitation Decision-Support System." *Computers, Environment and Urban Systems* 24 (3): 215–232. https://doi.org/10.1016/S0198-9715(99)00058-7.

Haixiang, Guo, Li Yijing, Jennifer Shang, Gu Mingyun, Huang Yuanyue, and Gong Bing. 2017. "Learning from Class-Imbalanced Data: Review of Methods and Applications." *Expert Systems with Applications* 73: 220–239. https://doi.org/10.1016/j.eswa.2016.12.035 .

Harvey, Robert Richard, and Edward Arthur McBean. 2014. "Comparing the Utility of Decision Trees and Support Vector Machines When Planning Inspections of Linear Sewer Infrastructure." *Journal of Hydroinformatics* 16 (6): 1265–1279. https://doi.org/10.2166/HYDRO.2014.007.

Haugen, Hans Jørgen, and Asplan Viak. 2018. "Dataflyt – Klassifisering av avløpsledninger." *Norwegian Water BA*. 40. https://docplayer.me/

211256711-Norsk-vann-rapport-dataflyt-klassifisering-av-avlopsledninger.html.

Hawari, Alaa, Firas Alkadour, Mohamed Elmasry, and Tarek Zayed. 2018. "Condition Assessment Model for Sewer Pipelines Using Fuzzy-Based Evidential Reasoning." *Australian Journal of Civil Engineering* 16 (1): 23–37. https://doi.org/10.1080/14488353.2018.1444333.

Hawari, Alaa, Firas Alkadour, Mohamed Elmasry, and Tarek Zayed. 2020. "A State of the Art Review on Condition Assessment Models Developed for Sewer Pipelines." *Engineering Applications of Artificial Intelligence* 93: 103721. https://doi.org/10.1016/j.engappai.2020.103721.

Hawari, Alaa, Alkadour Firas, Mohamed Elmasry, and Tarek Zayed. 2016. "Simulation-Based Condition Assessment Model for Sewer Pipelines." *Journal of Performance of Constructed Facilities* 31 (1): 04016066–. https://doi.org/10.1061/(ASCE)CF.1943-5509.0000914.

Hilal, Anwer Mustafa, Fahd N. Al-Wesabi, Masoud Alajmi, Majdy M. Eltahir, Mohammad Medani, Mesfer Al Duhayyim, Manar Ahmed Hamza, and Abu Sarwar Zamani. 2021. "Machine Learning-Based Decision Tree J48 with Grey Wolf Optimizer for Environmental Pollution Control." *Environmental Technology* 1–12. https://doi.org/10.1080/09593330.2021.2017491.

Kabir, Golam, Balekelay Celestin Balek Ngandu, Solomon Tesfamariam, and M. Asce. 2018. "Sewer Structural Condition Prediction Integrating Bayesian Model Averaging with Logistic Regression." *Journal of Performance of Constructed Facilities* 32 (3): 04018019–. https://doi.org/10.1061/(ASCE)CF.1943-5509.0001162.

Khan, Zafar, Tarek Zayed, and Osama Moselhi. 2010. "Structural Condition Assessment of Sewer Pipelines." *Journal of Performance of Constructed Facilities* 24 (2): 170–179. https://doi.org/10.1061/(ASCE)CF.1943-5509.0000081.

Kleiner, Yehuda, and Balvant Rajani. 2001. "Comprehensive Review of Structural Deterioration of Water Mains: Statistical Models." *Urban Water* 3 (3): 131–150. https://doi.org/10.1016/S1462-0758(01)00033-4.

Kritikos, Theodosios, and Tim Davies. 2015. "Assessment of Rainfall-Generated Shallow Landslide/debris-Flow Susceptibility and Runout Using a GIS-Based Approach: Application to Western Southern Alps of New Zealand." *Landslides* 12 (6): 1051–1075. https://doi.org/10.1007/s10346-014-0533-6.

Kuliczkowska, Emilia, Andrzej Kuliczkowski, and Anna Parka. 2022. "Damages in Vitrified Clay Sewers in Service for 130–142 Years." *Engineering Failure Analysis* 135: 106103. https://doi.org/10.1016/j.engfailanal.2022.106103.

Kuncheva, Ludmila I, and Juan J Rodríguez. 2007. An Experimental Study on Rotation Forest Ensembles. Paper presented at the International workshop on multiple classifier systems. Prague, Czech Republic.

Kursa, Miron B., and Witold R. Rudnicki. 2010. "Feature Selection with the Boruta Package." *Journal of Statistical Software* 36 (11): 1–13. https://doi.org/10.18637/jss.v036.i11.

Laakso, Tuija, Teemu Kokkonen, Ilkka Mellin, and Riku Vahala. 2018. "Sewer Condition Prediction and Analysis of Explanatory Factors." *Water 2018* 10 (9): 1239. https://doi.org/10.3390/W10091239.

Lim, Tjen-Sien, Wei-Yin Loh, and Yu-Shan Shih. 2000. "A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms." *Machine Learning* 40 (3): 203–228. https://doi.org/10.1023/A:1007608224229.

Lubini, Alain T, and Musandji. Fuamba. 2011. "Modeling of the Deterioration Timeline of Sewer Systems." *Canadian Journal of Civil Engineering* 38 (12): 1381–1390. https://cdnsciencepub.com/doi/pdf/10.1139/l11-103.

Miraki, Shaghayegh, Sasan Hedayati Zanganeh, Kamran Chapi, Vijay P. Singh, Ataollah Shirzadi, Himan Shahabi, and Binh Thai Pham. 2019. "Mapping Groundwater Potential Using a Novel Hybrid Intelligence Approach." *Water Resources Management* 33 (1): 281–302. https://doi.org/10.1007/s11269-018-2102-6.

Mohammadi, Mohammadreza Malek, Mohammad Najafi, Sharareh Kermanshachi, Vinayak Kaushal, and Ramtin Serajiantehrani. 2020. "Factors Influencing the Condition of Sewer Pipes: State-Of-The-Art Review." *Journal of Pipeline Systems Engineering and Practice* 11 (4): 03120002. https://doi.org/10.1061/(ASCE)PS.1949-1204.0000483.

Mohammadi, Mohammadrza Malek, Mohammad Najafi, Amir Tabesh, Jamie Riley, and Jessica Gruber. 2019. "Condition Prediction of Sanitary Sewer Pipes." *Pipelines 2019: Condition Assessment, Construction, and Rehabilitation - Proceedings of Sessions of the Pipelines 2019 Conference*. August 21-24, 2005. Houston, Texas, United States: 117–126. doi:10.1061/40800(180)61.

Mohammadi, Mohammadreza Malek, Mohammad Najafi, Kaushal Vinayak, Serajiantehrani Ramtin, Salehabadi Nazanin, and Ashoori Taha. 2019. "Sewer Pipes Condition Prediction Models: A State-Of-The-Art Review." *Infrastructures* 4 (4): 64. https://doi.org/10.3390/infrastructures4040064.

Najafi, Mohammad, and Guru Kulandaivel. 2005. "Pipeline Condition Prediction Using Neural Network Models." Pipelines 2005: Optimizing Design, Operations, and Maintenance. August 21-24, 2005. Houston, Texas, United States. 767–781. doi:10.1061/40800(180)61.

Nanda, Muhammad Achirul, Kudang Boro Seminar, Akhiruddin Maddu, and Dodi Nandika. 2021. "Identifying Relevant Features of Termite Signals Applied in Termite Detection System." *Ecological Informatics* 64: 101391. https://doi.org/10.1016/j.ecoinf.2021.101391.

Ngandu, Balekelayi, Solomon Tesfamariam, and M. Asce. 2019. "Statistical Inference of Sewer Pipe Deterioration Using Bayesian Geoadditive Regression Model." *Journal of Infrastructure Systems* 25 (3): 04019021–.

Phong, Tran Van, Binh Thai Pham, Phan Trong Trinh, Ly Hai-Bang, Vu Quoc Hung, Lanh Si Ho, Le Hiep Van, Lai Hop Phong, Mohammadtaghi Avand, and Indra Prakash. 2021. "Groundwater Potential Mapping Using GIS-Based Hybrid Artificial Intelligence Methods." *Groundwater* 59 (5): 745–760. https://doi.org/10.1111/gwat.13094.

Rajani, Balvant, and Yehuda Kleiner. 2001. "Comprehensive Review of Structural Deterioration of Water Mains: Physically Based Models." *Urban Water* 3 (3): 151–164. https://doi.org/10.1016/S1462-0758(01)00032-2.

Randall-Smith, M., R. Oliphant, and A. Russell. 1992. *Guidance Manual for the Structural Condition Assessment of Trunk Mains*. Swindon, UK: Water Research Centre: WRc.

RIF. 2021. Rådgivende Ingeniørers Forening. Oslo, Norway: Rådgivende Ingeniørers Forening (RIF). https://rif.no/wp-content/uploads/2019/08/Vann-Avl%C3%B8psanlegg.pdfAccessed June 15.

Rodriguez, Juan José, Ludmila I Kuncheva, and Carlos J Alonso. 2006. "Rotation Forest: A New Classifier Ensemble Method." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 28 (10): 1619–1630. https://doi.org/10.1109/TPAMI.2006.211.

Sahu, S., and B. M. Mehtre. 2015. Network Intrusion Detection System Using J48 Decision Tree. Paper presented at the 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI) Kochi, India, 10-13 Aug. 2015.

Salman, Baris, and Ossama Salem. 2012. "Modeling Failure of Wastewater Collection Lines Using Various Section-Level Regression Models." *Journal of Infrastructure Systems* 18 (2): 146–154. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000075.

Sanchez-Pinto, L. Nelson, Laura Ruth Venable, John Fahrenbach, and Matthew M. Churpek. 2018. "Comparison of Variable Selection Methods for Clinical Predictive Modeling." *International Journal of Medical Informatics* 116: 10–17. https://doi.org/10.1016/j.ijmedinf.2018.05.006.

Sempewo, Jotham Ivan, and Lydia Kyokaali. 2019. "Comparative Performance of Regression and the Markov Based Approach in the Prediction of the Future Condition of a Water Distribution Pipe Network Amidst Data Scarce Situations: A Case Study of Kampala Water, Uganda." *Water Practice & Technology* 14 (4): 946–958. https://doi.org/10.2166/WPT.2019.075.

Shahabi, Himan, Ataollah Shirzadi, Kayvan Ghaderi, Ebrahim Omidvar, Nadhir Al-Ansari, John J. Clague, Marten Geertsema, et al. 2020. "Flood Detection and Susceptibility Mapping Using Sentinel-1 Remote Sensing Data and a Machine Learning Approach: Hybrid Intelligence of Bagging Ensemble Based on K-Nearest Neighbor Classifier." *Remote Sensing* 12 (2): 266. https://doi.org/10.3390/rs12020266 .

Shirzadi, Ataollah, Karim Solaimani, Mahmood Habibnejad Roshan, Ataollah Kavian, Kamran Chapi, Himan Shahabi, Saskia Keesstra, Baharin Bin Ahmad, and Dieu Tien Bui. 2019. "Uncertainties of Prediction Accuracy in Shallow Landslide Modeling: Sample Size and Raster Resolution." *CATENA* 178: 172–188. https://doi.org/10.1016/j.catena.2019.03.017.

Shirzadi, Ataollah, Karim Soliamani, Mahmood Habibnejhad, Ataollah Kavian, Kamran Chapi, Himan Shahabi, Wei Chen, et al. 2018.

"Novel GIS Based Machine Learning Algorithms for Shallow Landslide Susceptibility Mapping." *Sensors* 18 (11): 3777. https://doi.org/10.3390/s18113777.

Syachrani, Syadaruddin, Hyung Seok, David Jeong, and Colin S. Chung. 2013. "Decision Tree–Based Deterioration Model for Buried Wastewater Pipelines." *Journal of Performance of Constructed Facilities* 27 (5): 633–645. https://doi.org/10.1061/(ASCE)CF.1943-5509.0000349.

Tharwat, Alaa. 2021. "Classification Assessment Methods." *Applied Computing & Informatics* 17 (1): 168–192. https://doi.org/10.1016/j.aci.2018.08.003.

Ting, Kai Ming, and Ian H Witten. 1997. "Stacking Bagged and Dagged Models." ICML '97: Proceedings of the Fourteenth International Conference on Machine Learning Pages. San Francisco, CA, United States: Morgan Kaufmann. 367–375. https://researchcommons.waikato.ac.nz/handle/10289/1072.

Tizmaghz, Z., J. E. van Zyl, and T. F. P. Henning. 2022. "Consistent Classification System for Sewer Pipe Deterioration and Asset Management." *Journal of Water Resources Planning and Management* 148 (5): 04022011. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001545.

Tran, Huu Dung. 2007. "Investigation of deterioration models for stormwater pipe systems." PhD thesis, Victoria University. https://vuir.vu.edu.au/1456/3/TRAN%20Huu%20Dung-thesis_nosignature.pdf.

Tran, H. D., and A. W. M. Ng. 2010. "Classifying Structural Condition of Deteriorating Stormwater Pipes Using Support Vector Machine." *Pipelines 2010: Climbing New Peaks to Infrastructure Reliability - Renew, Rehab, and Reinvest - Proc. of the Pipelines 2010 Conference*. Pipeline Division Specialty Conference 2010August 28 - September 1, 2010. Keystone, Colorado, United States. 386:857–866. https://doi.org/10.1061/41138(386)82.

Tran, H. D., B. J. C. Perera, and A. W. M. Ng. 2009. "Markov and Neural Network Models for Prediction of Structural Deterioration of Storm-Water Pipe Assets." *Journal of Infrastructure Systems* 16 (2): 167–171. https://doi.org/10.1061/(ASCE)IS.1943-555X.0000025.

Tsai, Chih-Fong, Li Miao-Ling, and Wei-Chao Lin. 2018. "A Class Center Based Approach for Missing Value Imputation." *Knowledge-Based Systems* 151: 124–135. https://doi.org/10.1016/j.knosys.2018.03.026.

Tscheikner-Gratl, Franz, Nicolas Caradot, Frédéric Cherqui, Joao P. Leitão, Mehdi Ahmadi, Jeroen G. Langeveld, Yves Le Gat, et al. 2019. "Sewer Asset Management – State of the Art and Research Needs." *Urban Water Journal* 16 (9): 662–675. https://doi.org/10.1080/1573062X.2020.1713382.

Venkatesh, G., and Helge Brattebø. 2012. "Assessment of Environmental Impacts of an Aging and Stagnating Water Supply Pipeline Network." *Journal of Industrial Ecology* 16 (5): 722–734. https://doi.org/10.1111/j.1530-9290.2011.00426.x.

Vitorino, D., S. T. Coelho, P. Santos, S. Sheets, B. Jurkovac, and C. Amado. 2014. "A Random Forest Algorithm Applied to Condition-Based Wastewater Deterioration Modeling and Forecasting." *Procedia Engineering* 89: 401–410. https://doi.org/10.1016/j.proeng.2014.11.205.

Vladeanu, G., John Matthews, and M. Asce. 2019. "Wastewater Pipe Condition Rating Model Using Multicriteria Decision Analysis." *Journal of Water Resources Planning and Management* 145 (12): 04019058–. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001134.

Yin, Xianfei, Yuan Chen, Ahmed Bouferguene, and Mohamed Al-Hussein. 2020. "Data-Driven Bi-Level Sewer Pipe Deterioration Model: Design and Analysis." *Automation in Construction* 116: 103181. https://doi.org/10.1016/j.autcon.2020.103181.

Paper IV

Lam Van Nguyen, Dieu Tien Bui, and Seidu Razak (2023). Utilization of Augmented Reality Technique for Sewer Condition Visualization. *Water*, *15*(24), 4232. https://doi.org/10.3390/w15244232.

water

Article

Utilization of Augmented Reality Technique for Sewer Condition Visualization

Lam Van Nguyen, Dieu Tien Bui and Razak Seidu

MDPI

https://doi.org/10.3390/w15244232

water

MDPI

Article

Utilization of Augmented Reality Technique for Sewer Condition Visualization

Lam Van Nguyen [1,*], Dieu Tien Bui [2] and Razak Seidu [1]

[1] Smart Water and Environmental Engineering Group, Department of Ocean Operations and Civil Engineering, Faculty of Engineering, Norwegian University of Science and Technology, 6025 Ålesund, Norway; rase@ntnu.no

[2] GIS Group, Department of Business and IT, University of South-Eastern Norway, 3800 Bø i Telemark, Norway; dieu.t.bui@usn.no

* Correspondence: lam.v.nguyen@ntnu.no

Abstract: Wastewater pipelines are largely buried underground, and techniques for assessing and visualizing their condition are critical for planning and rehabilitation. This paper introduces a framework for integrating Geographic Information System (GIS), 3D-creation platform, augmented reality (AR) techniques, and machine learning algorithms for the dynamic visualization of the condition of sewer networks. A sewer network in Ålesund City, Norway, was used as a case study, and the developed framework was implemented on an Android OS and Microsoft HoloLens. The results show the potential applications of the integrated framework of GIS, AR, and 3D models for sewer condition visualization. The positioning accuracy of the application for 2D objects is equivalent to that of well-designed GPS receivers (approximately 1–3 m), depending on the handheld device used. Loading and locating 3D objects will be limited by the performance of the devices used.

Keywords: augmented reality; 3D visualization; mobile application; sewer network system

check for updates

Citation: Nguyen, L.V.; Bui, D.T.; Seidu, R. Utilization of Augmented Reality Technique for Sewer Condition Visualization. *Water* **2023**, *15*, 4232. https://doi.org/10.3390/w15244232

Academic Editors: Carmen Teodosiu and Pankaj Kumar

Received: 1 November 2023
Revised: 29 November 2023
Accepted: 5 December 2023
Published: 8 December 2023

1. Introduction

The collection, transport, and treatment of wastewater play a significant role in protecting the environment and public health. Wastewater collection systems mainly comprise pipes of different materials that undergo degradation over time. In Norway, the condition of the sewer network in some municipalities is poor, with a low annual renewal rate of approximately 0.6% [1], and requiring significant investments [2]. To plan investments, deterioration models could be used to assess the current and future status of sewer pipes by accounting for intrinsic (e.g., size, age, or material) and extrinsic (e.g., rainfall, soil type, or population) factors. Models used for the assessment and prediction of sewer conditions include regression [3], classification [4], and hybrid models [5]. The results of these models have been presented in Geographic Information System (GIS) for visualization of the geolocation of pipes that need improvement. Although the visualization of pipe conditions in a GIS map provides a powerful interface, inaccuracies in geographical data may inhibit the usefulness of such maps for operation and maintenance tasks. Recent studies reveal that visualization through augmented reality is one of the most cost-effective and dependable approaches for the operation and maintenance of urban underground pipe networks [6,7].

Augmented reality (AR) is a technology that integrates digital/virtual information with the user's environment in real time. This technique enhances the visual perspectives of the physical real-world environment by using non-visual properties and capabilities of computing devices [8]. With the massive improvements in the computing power of computer software and hardware through Industry 4.0, AR is becoming one of the most promising technologies that in the future will support people in recognizing and experiencing real-world objects in a completely new way [9]. Moreover, AR technology has proven to be an effective supporting tool in product design, manufacturing, maintenance/inspection, and training activities [10]. However, applying AR technique to visualize objects in the water sector is still challenging and limited [11–14]. For example, Haynes, Hehl-Lange and Lange [14] indicated that the synchronization of 3D points displayed on device monitors and the corresponding points in the real world was one of the biggest challenges associated with the application of AR techniques. In addition, Mirauda, Erra, Agatiello and Cerverizzo [11] showed that handheld device orientation before use is essential, and the reliability of the AR application decreased over time because of drift, requiring that the device be reset during the experiment.

According to Hahmann and Burghardt [15], approximately 80% of all information available can be interpreted as spatial data or geodata. Therefore, profoundly digging into these data can provide valuable information for management and maintenance purposes. Data visualization can be easily implemented by using a variety of visualization packages: for example, the Microsoft Excel (https://www.microsoft.com/en-us/microsoft-365/excel) (accessed on 6 December 2023), the package Matplotlib (version: 3.8.2) in Python for 2D and 3D visualization [16], and the package ggplot2 (version: 3.4.4) in R [17]. One of the most significant disadvantages of the aforementioned packages is the ability to visualize data only on personal computers, which is not convenient for fieldwork. In terms of graphical visualization, spatiotemporal data can be generally visualized in two- or three-dimensional environments using three different techniques, including 2D maps, space-time cubes, or animations [18]. Based on the authors' knowledge, few mobile applications have implemented the AR technique for visualization in the water sector.

The visualization platform, or visual communication, supports human problem-solving and improves user decision-making performance by transforming non-visual objects into visual objects that are accessible to the human mind [19]. The users/managers cannot effectively control or monitor the system without deeply understanding the data. An effective visualization platform will support water engineers/managers in having a visual overview and correctly evaluating the system status to inform reasonable maintenance strategies [20]. The main aim of this work was to develop a mobile application integrated with AR technique to support sewer management through 2D/3D visualization of sewer conditions. By combining sewer conditions produced from predictive models and AR technology, water engineers/managers can quickly estimate a sewer's status in the field and reduce workloads compared to conventional methods such as digging or camera-based inspection. The study was conducted on sewer pipes in Ålesund city, Norway that were previously used for the development of deterioration models.

2. Theory and Methods

2.1. Literature Review

The application of AR in the management of underground utilities has received much attention in recent years. Huston, et al. [21] used building information modeling (BIM) and 3D Geographic Information System (GIS) for both design and spatial data during the planning, construction, and system lifecycle management phases of underground utilities. In that work, information technology and sensor signals were discussed to assess the state of urban underground infrastructure (e.g., water-related networks, natural gas, electric power) and provide reliable information for managers, planners, and users. A mask regional convolutional neural network and a 3D model were developed by Fang, et al. [22] to automatically detect, localize, and visualize sewer defects using a floating capsule robot.

A mobile interface connected to the internet using a client–server architecture was introduced by Schall, Zollmann and Reitmayr [13]. This AR system allows users to create, read, update, and delete data from a geospatial database. This mobile AR application improved workflows such as on-site planning, data capture and surveying, and on-site visualization. Similarly, AR systems allowing for editing and visualizing underground facilities' geographic and attribute data using Unity3D have been developed [23,24]. Pereira, et al. [25] introduced a combination of AR and a ground penetrating radar (GPR) technique for the mapping and assessment of underground infrastructure. This system reduces the limitations of current GPR systems that have degraded or unavailable Global Positioning System (GPS) signals in urban canyons and city tunnels, respectively.

An integration between state-of-the-art technologies, including mobility, Global Navigation Satellite System (GNSS), AR, and 3D GIS geo-database, was mentioned in the study by Jimenez, et al. [26] to guide utility field workers in visualizing buried infrastructure. The authors also indicated market analysis to identify appropriate business models based on the developed system. Although the work briefly showed the main aspects to be considered in future commercialization, it did not clearly explain the aforementioned integration.

An AR-based underground facility management system using Map API and JSON communication techniques was proposed by Kim [27] to provide, manage, and replace the location information from the GIS system for underground facilities. The primary limitations of this approach become apparent as the author delved into its insufficient scalability and the constraints posed by GIS data when integrated with the utilized LibGDX engine.

Li, Feng, Han and Liu [6] developed a mobile augmented reality-based framework to visualize static and dynamic data of a real-life urban gas pipe network. Tarek and Marzouk [28] developed a smartphone AR application to visualize infrastructure networks, thereby improving infrastructure operation and maintenance workflows. An application using a Microsoft HoloLens headset was proposed by Côté and Mercier [29] to visualize pipe maps on the road surface. Rahman, et al. [30] developed an integrated framework comprising machine learning (ML), sensor data, and AR techniques to manage a prawn farm. The results have proved that the AR technique (as well as its integration) is a promising tool to manage assets more efficiently.

The aforementioned studies indicate that augmented reality-based techniques are receiving more attention from researchers and have become one of the promising means of visualization in recent years. This study provides an augmented reality framework for the visualization of pipe conditions. The study introduces a workflow from data collection and processing to ML implementation to predict the condition of sewer pipes, along with a 3D model for sewer condition visualization. The study builds on earlier studies of the authors on sewer condition assessment and prediction using ML for the Ålesund Municipality in Norway [3–5]. The platform developed in this study will allow sewer infrastructure managers to visualize the condition of pipes on-site for planning and maintenance.

2.2. Data and 3D Model Preparation

Two main data types were used in the preparation phase of this application, including object-based data (3D models such as pipes and manholes) and their corresponding attribute-based data (such as material or status). The steps for data preparation are presented in Figure 1.

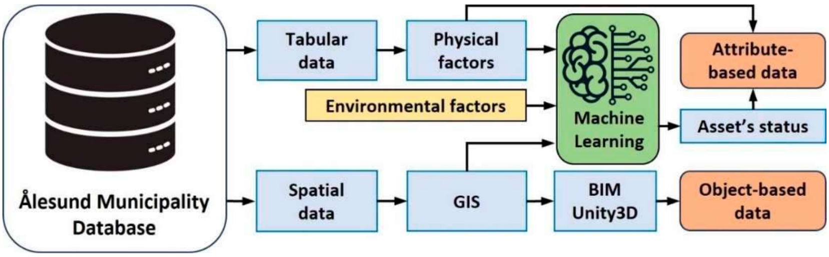


Figure 1. The workflow for data preparation.

Some initial attribute data (e.g., material or installation year) and geospatial data (such as network structure) were extracted from databases managed by the Ålesund Municipality. These data and an auxiliary environmental dataset were used to compute sewer conditions. Sewer conditions were divided into good, intermediate, and bad conditions based on the study of Haugen and Viak [31]. Specifically, for the Ålesund sewer network, 10 physical factors (i.e., age, diameter, depth, length, slope, material, pipe type, network type, pipe form, and connection type) and 10 environmental factors (i.e., rainfall, geology, landslide area, population density, land cover, building area, groundwater level, traffic volume, road network, and soil type) were aggregated and assigned for each sewer pipe. Then, some filter, wrapper, and embedded feature selection methods were applied to eliminate insignificant factors from the dataset. Finally, 10 regression-based ML models, 17 classification-based ML models, and 4 hybrid ML models were developed to estimate the conditions of sewer pipelines. The best output produced from the models was selected as the input for visualizing using the application developed in this study. For more information on the application of ML models for sewer condition assessment, readers are referred to [3–5].

Geospatial data on the sewers were imported into GIS software, including ArcGIS (version: 10.8.2) or QGIS (version: 3.10 LTR), and BIM software, including Trimble SketchUp (version: Desktop 2022.0) and Autodesk InfraWorks (version: 17.11.0), to create 3D models. SketchUp, owned by Trimble, Inc., is a suite of products for a broad range of drawing and design applications, including architectural, interior design, industrial and product design, landscape architecture, civil and mechanical engineering, theater, film, and video game development [32]. Trimble SketchUp is widely applied to many real-world problems, such as annual solar insolation potentials analysis [33], volcanic processes visualization [34], microclimatic mapping [35], or tree row simulation [36]. InfraWorks, developed by Autodesk, Inc. (San Francisco, CA, USA), is one of the most widely used software applications in BIM environments for the planning and designing of infrastructure projects such as sewer networks (Figure 2). In Autodesk InfraWorks, users can easily incorporate GIS data to collect large amounts of data that give more accurate information about the area they are modeling, whether a built-up or natural environment. InfraWorks has been used as an integrated application within BIM and GIS for the smart city concept [37], drainage system management [38], and infrastructure simulation [39].

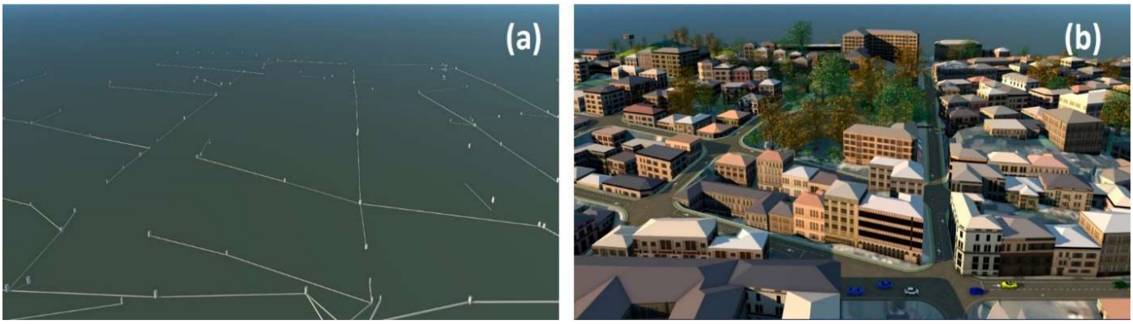
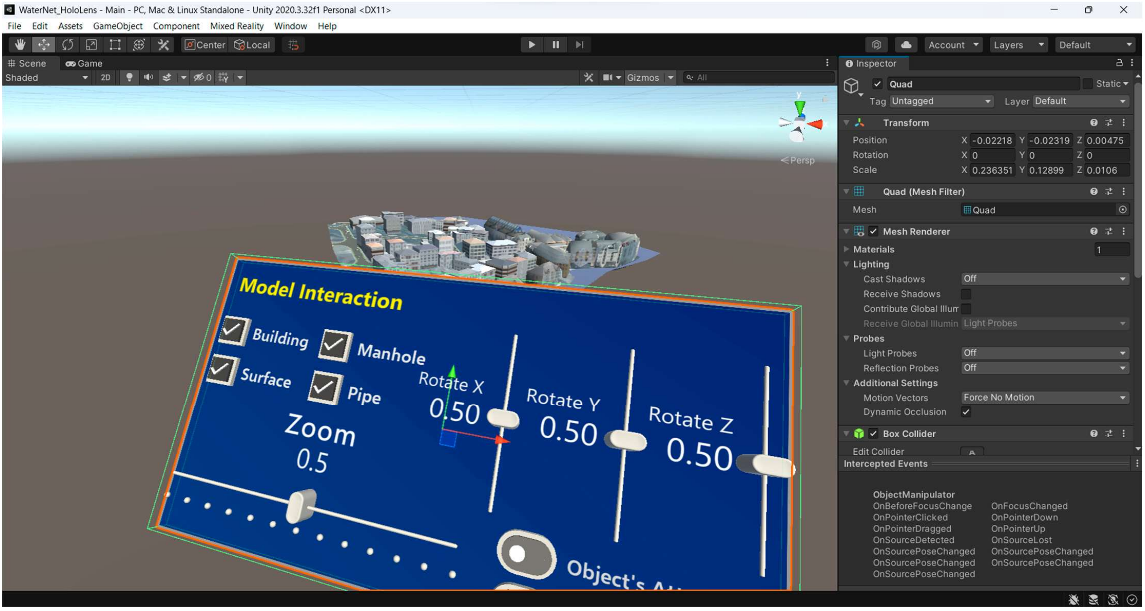


Figure 2. Illustration of the sewer network in the study area: (**a**) without surface; (**b**) with surface objects.

In this study, the 3D visualization platform was developed using Unity 3D, which is one of the most accessible and free tools for game developers [40]. Unity, which is developed by Unity Technologies, is a professional-quality game engine targeting a variety of platforms [41]. Unity 3D was employed to amalgamate 3D models acquired from Trimble SketchUp and Autodesk InfraWorks into the holographic environment. This integration included attributes extracted from the provided tabular dataset and predicted conditions from ML models. Unity 3D provides many comprehensive functions to implement AR applications [6]. Subsequently, the outcomes were visualized on both mobile devices and HoloLens. Additional data, such as road networks or building footprints, received from the Norwegian Mapping Authority (https://www.kartverket.no/en) (accessed on 20 March 2020) [4], were used to build the 3D model. These data were converted into shapefile (*.shp) format, which is an Environmental Systems Research Institute, Inc. (ESRI) (Redlands, CA, USA) vector data storage format for storing the location, shape, and attributes of geographic features.

An example of a 3D model of the study area in Unity is presented in Figure 3.



Inspector

☑ Quad ☐ Static

Tag Untagged Layer Default

Transform

Position	X	-0.02218	Y	-0.02319	Z	0.00475
Rotation	X	0	Y	0	Z	0
Scale	X	0.236351	Y	0.12899	Z	0.0106

Quad (Mesh Filter)

Mesh Quad

☑ Mesh Renderer

Materials 1

Lighting

Cast Shadows Off

Receive Shadows ☐

Contribute Global Illum ☐

Receive Global Illumin Light Probes

Probes

Light Probes Off

Reflection Probes Off

Additional Settings

Motion Vectors Force No Motion

Dynamic Occlusion ☑

☑ Box Collider

Edit Collider

Intercepted Events

ObjectManipulator

OnBeforeFocusChange	OnFocusChanged
OnPointerClicked	OnPointerDown
OnPointerDragged	OnPointerUp
OnSourceDetected	OnSourceLost
OnSourcePoseChanged	OnSourcePoseChanged
OnSourcePoseChanged	OnSourcePoseChanged
OnSourcePoseChanged	

Figure 3. An example of a 3D model in Unity.

2.3. Visualization Platform Development

The visualization platform developed in this study comprises three main segments: indoor, intermediate, and outdoor, as presented in Figure 4.

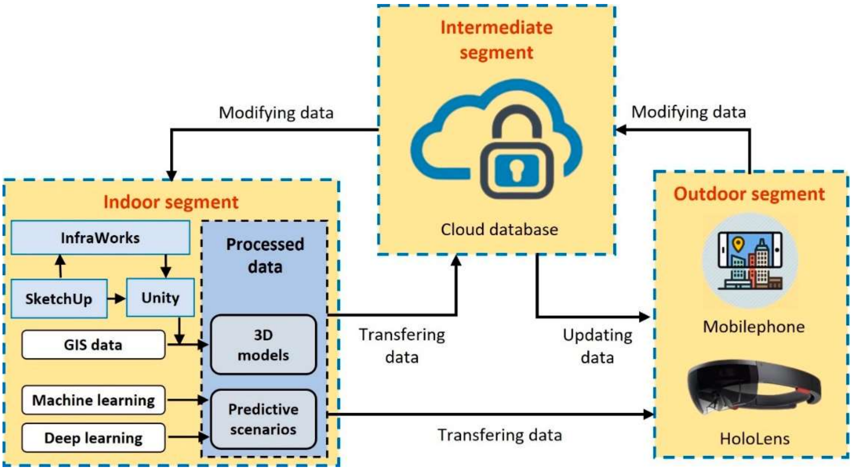


Figure 4. Overview of the integrated visualization platform.

The indoor segment involves 3D modeling and computation, focusing on the process implemented on personal and supercomputers. In this regard, 3D models of the sewer network (e.g., pipes, manholes, or pumps) were created and visualized using a high-performance computer system associated with the simulation programs [24,42]. Sewer conditions were predicted using ML and deep learning models based on input data, as described in Nguyen, Bui and Seidu [4]. The network components' output was coded based on their corresponding indexes for visualization purposes. While creating 3D objects, the names and indexes of original objects were maintained to merge with information obtained from the previous step. This information was reused if any new update was received from the intermediate segment. Finally, the models and auxiliary data were transformed into Unity 3D for simulation on smart devices such as smartphones and HoloLens.

The intermediate segment involves the processing of remote data. This segment employed cloud-based platforms to store data and have remote interactive connections with the computer or handheld devices. In this segment, the database created from the indoor segment was transferred to the host computer system on a cloud-based connection. Real-time data received from sensors in the sewers were updated and stored on the ***StaalCloud*** portal managed by Ålesund municipality. In this segment, the data will undergo initial processing to generate fundamental network details such as pipe index, velocity, and sewer condition status. After that, these generated attributes are assigned to corresponding objects based on their unique indexes, which are generated from the indoor segment. Following this process, the data can be transmitted to the outdoor segment for display on computer or handheld devices. Finally, the intermediate segment is configured to receive updated information from the outdoor segment, store it, and send it back to the indoor segment to update the system.

The outdoor segment involves visualizing and updating data. This segment contains devices that can visualize objects and their attributes in a hands-free manner. In this segment, the processed data obtained from the indoor segment and updated data received from the intermediate segment are used to enhance the user's experience via AR devices. The application directly reads information from computers or the cloud via a Wi-Fi network, matching them with corresponding objects and showing designated information. Any modification can be performed, and modified data in this segment can be transmitted back to the host computer in the indoor segment or the cloud database in the intermediate segment using an application programming interface (API) or equivalent protocols to update the database.

In our work, due to the limitations of accessing the cloud-based database (for data security reasons), we only tested two cases: (1) the connection between the indoor segment and the outdoor segment and (2) the ability to access, download, and visualize real-time data from the ***StaalCloud*** portal. The functions and performance for testing on the intermediate segment will need to be implemented in further investigations.

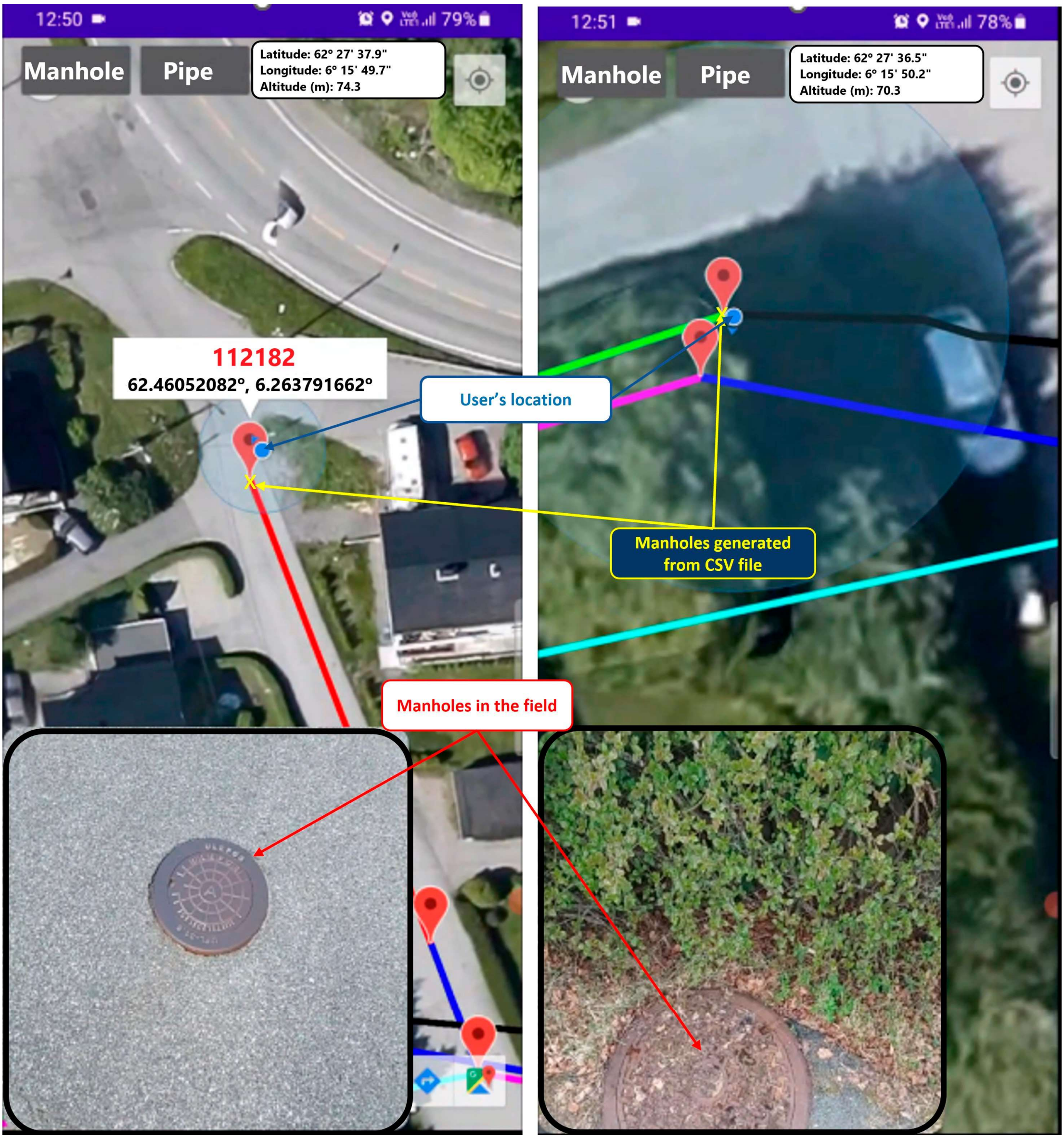
2.4. System Configuration

The ***StaalCloud*** portal is managed by the Ålesund municipality, and this is a platform to collect and store data from sensors. This portal was used to test the ability of the application to access, download, and visualize near real-time data. For the 3D visualization development platform, the Unity 3D game engine (version: 2021.3.15f1 Long Term Support), Android Studio (version: 2022.1.1), Java (version: 8), and C# (version: .NET Core 3.0, C# 8.0) programming languages were selected to develop and compile the application on mobile and Microsoft HoloLens devices.

In terms of hardware for running these applications, the Samsung Galaxy A42 5G (Android 10, 4 GB RAM, Qualcomm Snapdragon 690) [43] and Microsoft HoloLens (Windows 10, 64 GB Flash, 2 GB RAM) [44] were used. Except for the shape of objects (in 3D type), other attribute-related data were structured in the comma-separated values (CSV) format that is easily opened and modified by a text editor such as Notepad (version: 10240.0 or higher) or Microsoft Excel (https://www.microsoft.com/en-us/microsoft-365/excel) (accessed on 6 December 2023).

2.5. Accuracy Estimation

To assess the surveying accuracy of the application, experiments were performed at known reference points. Sewer networks are fundamentally composed of pipes and manholes [45]. While pipes are mainly invisible due to being covered by the ground surface, manholes can be easily recognized based on their cover on the ground. In this study, the accuracy of the mobile application was assessed based on the comparison of differences between the actual locations of manholes and their corresponding locations determined using the application (Figure 5). The accuracy of the HoloLens-developed application was not performed because of two reasons: (1) the first version of the HoloLens device used in this study does not support a GPS module inside, and (2) the development of the application on the HoloLens device aims to expand user interactions with holograms embedded into the real world.



12:50 79%

Latitude: 62° 27' 37.9"
Longitude: 6° 15' 49.7"
Altitude (m): 74.3

112182
62.46052082°, 6.263791662°

12:51 78%

Latitude: 62° 27' 36.5"
Longitude: 6° 15' 50.2"
Altitude (m): 70.3

Figure 5. Location of the manholes in the study area.

The surveyed locations of tested manholes were latitude and longitude values of corresponding points in the World Geodetic System 1984 (WGS-84). However, the actual manholes' locations in the database provided by Ålesund municipality were the horizontal coordinates of these points in the *EPSG:32632-WGS84/UTM Zone 32N* system; therefore, these geographical locations were transferred into this plane coordinate system. This transformation process was performed using ArcGIS Pro (version: 2.7.6) software developed by ESRI [46].

The locations of tested manholes with their names are shown in Figure 6; 20 different positions with different angles were implemented at each checked point to obtain measurements. The root mean square error (RMSE) was used to assess the accuracy of the developed application in locating manholes, as follows [6]:

$$RMSE = \sqrt{\frac{1}{20}\left(\sum_{i=1}^{20}\left(x_i^{obs} - x_i^{act}\right)^2 + \sum_{i=1}^{20}\left(y_i^{obs} - y_i^{act}\right)^2\right)} \quad (1)$$

where $\left(x_i^{act}, y_i^{act}\right)$ and $\left(x_i^{obs}, y_i^{obs}\right)$ are actual horizontal coordinates and observed horizontal coordinates at one tested manhole, respectively.

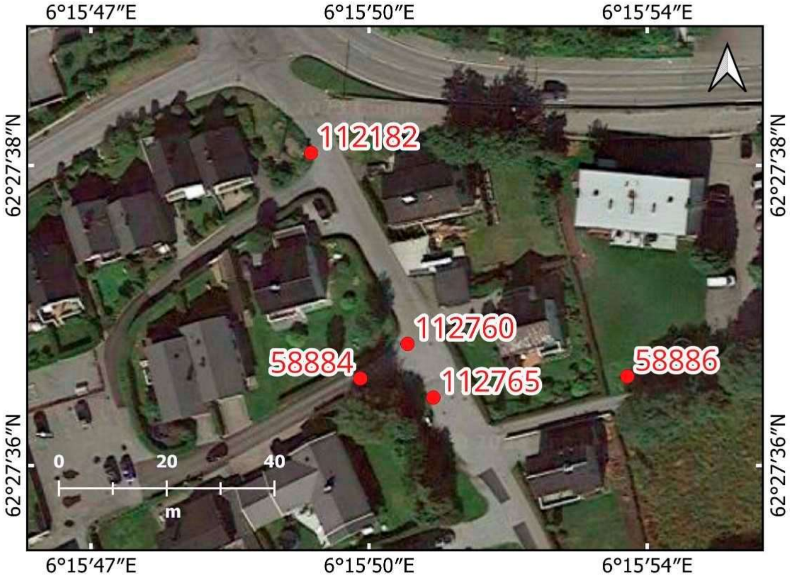


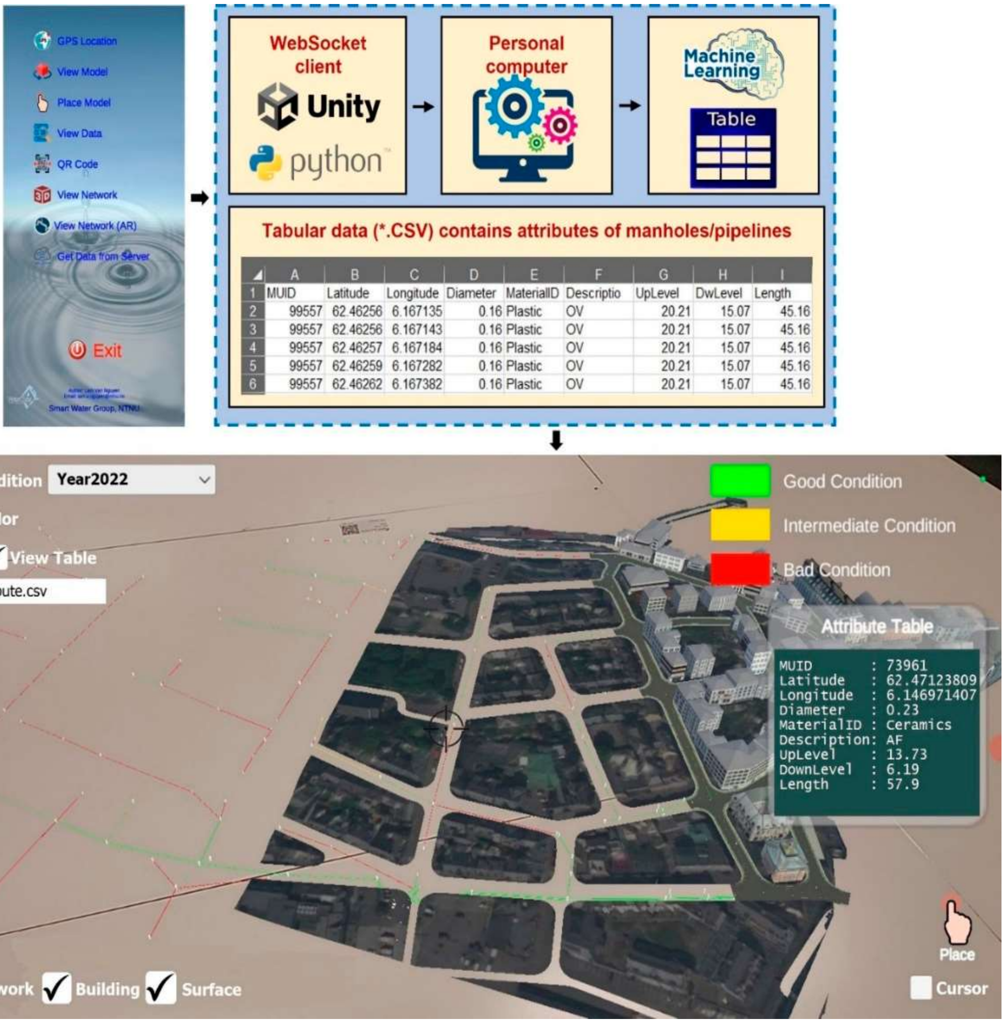
Figure 6. Locations of the tested manholes.

3. Results and Discussion

3.1. Visualization of Pipe Conditions with Mobile Application

The mobile application developed in this study allows the user to import attributes of manholes or pipes from a CSV file and visualize them in a real-time perspective. Moreover, the user can send commands from the mobile device to a personal computer (PC) via the WebSocket API protocol.

Based on a specific command received from the application on the mobile device, the PC will access a given tabular dataset or run pre-defined ML models to predict sewer condition status and produce the result. After that, these results are sent back to visualize on the mobile device. The process is illustrated in Figure 7.



	A	B	C	D	E	F	G	H	I
1	MUID	Latitude	Longitude	Diameter	MaterialID	Descriptio	UpLevel	DwLevel	Length
2	99557	62.46256	6.167135	0.16	Plastic	OV	20.21	15.07	45.16
3	99557	62.46256	6.167143	0.16	Plastic	OV	20.21	15.07	45.16
4	99557	62.46257	6.167184	0.16	Plastic	OV	20.21	15.07	45.16
5	99557	62.46259	6.167282	0.16	Plastic	OV	20.21	15.07	45.16
6	99557	62.46262	6.167382	0.16	Plastic	OV	20.21	15.07	45.16

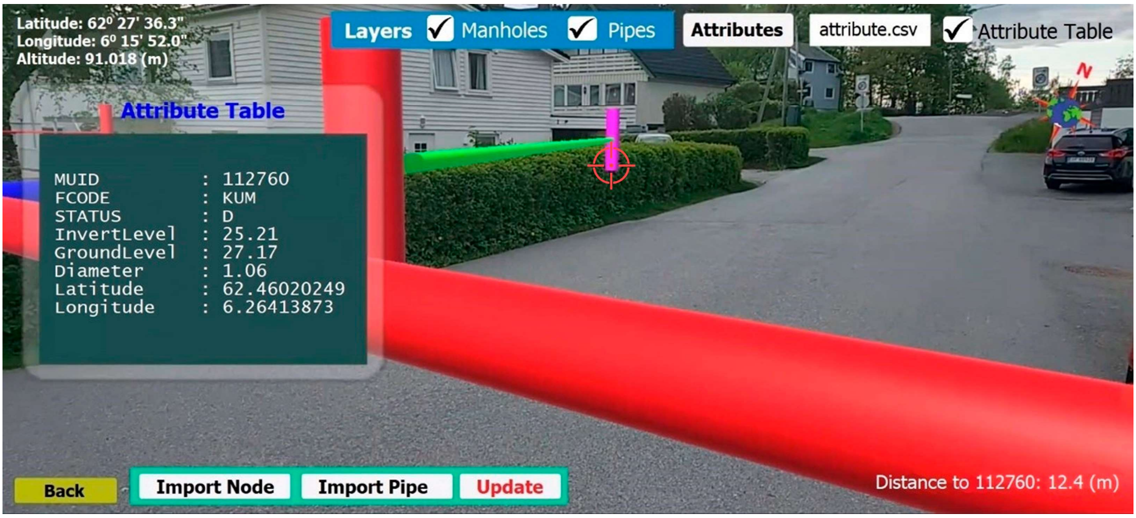
```
MUID       : 73961
Latitude   : 62.47123809
Longitude  : 6.146971407
Diameter   : 0.23
MaterialID : Ceramics
Description: AF
UpLevel    : 13.73
DownLevel  : 6.19
Length     : 57.9
```

Figure 7. Network visualization in the application.

An example of AR network visualization on a real scale is shown in Figure 8. After importing the given CSV files from the device's storage, the users can see attributes, including their condition, by moving the center point of the device screen onto objects of interest.

Figure 9 shows an example of real-time data access and visualization received from sensors through the ***StaalCloud*** portal. The values in this figure (i.e., water level and temperature) will automatically change based on the user's option.

The application also provides users with some extra functions, such as pinpointing locations of interest in the field using signals from satellites or visualizing given 3D objects. Specifically, the application allows the user to pinpoint the device's location in the WGS-84 coordinate reference system from the GPS signal. This function is useful to pinpoint problems in the field (e.g., crack locations or noticed points), and the output can be saved in the CSV format (Figure 10).



Latitude: 62º 27' 36.3"
Longitude: 6º 15' 52.0"
Altitude: 91.018 (m)

```
MUID          : 112760
FCODE         : KUM
STATUS        : D
InvertLevel   : 25.21
GroundLevel   : 27.17
Diameter      : 1.06
Latitude      : 62.46020249
Longitude     : 6.26413873
```

Distance to 112760: 12.4 (m)

Figure 8. AR network visualization in the application.

3.2. HoloLens Application

The inspiration for developing an application on HoloLens is to enhance the sense of authenticity of the user's experience by providing several types of natural interaction, such as gaze, gesture, voice, and spatial mapping [47]. In order to build an application that was compatible with the HoloLens device, several configurations were developed. For example, the Mixed Reality Toolkit (version: 2.8.3) and Microsoft Mixed Reality OpenXR Plugin API (version: 1.1.15) were installed [48]. An example of running this application on the HoloLens device is presented in Figure 11.

3.3. Accuracy Assessment

The differences in distances between actual positions and measured positions (in 20 iterations) at each manhole are presented in Figure 12 (circles in the figure represent outliers in the dataset).

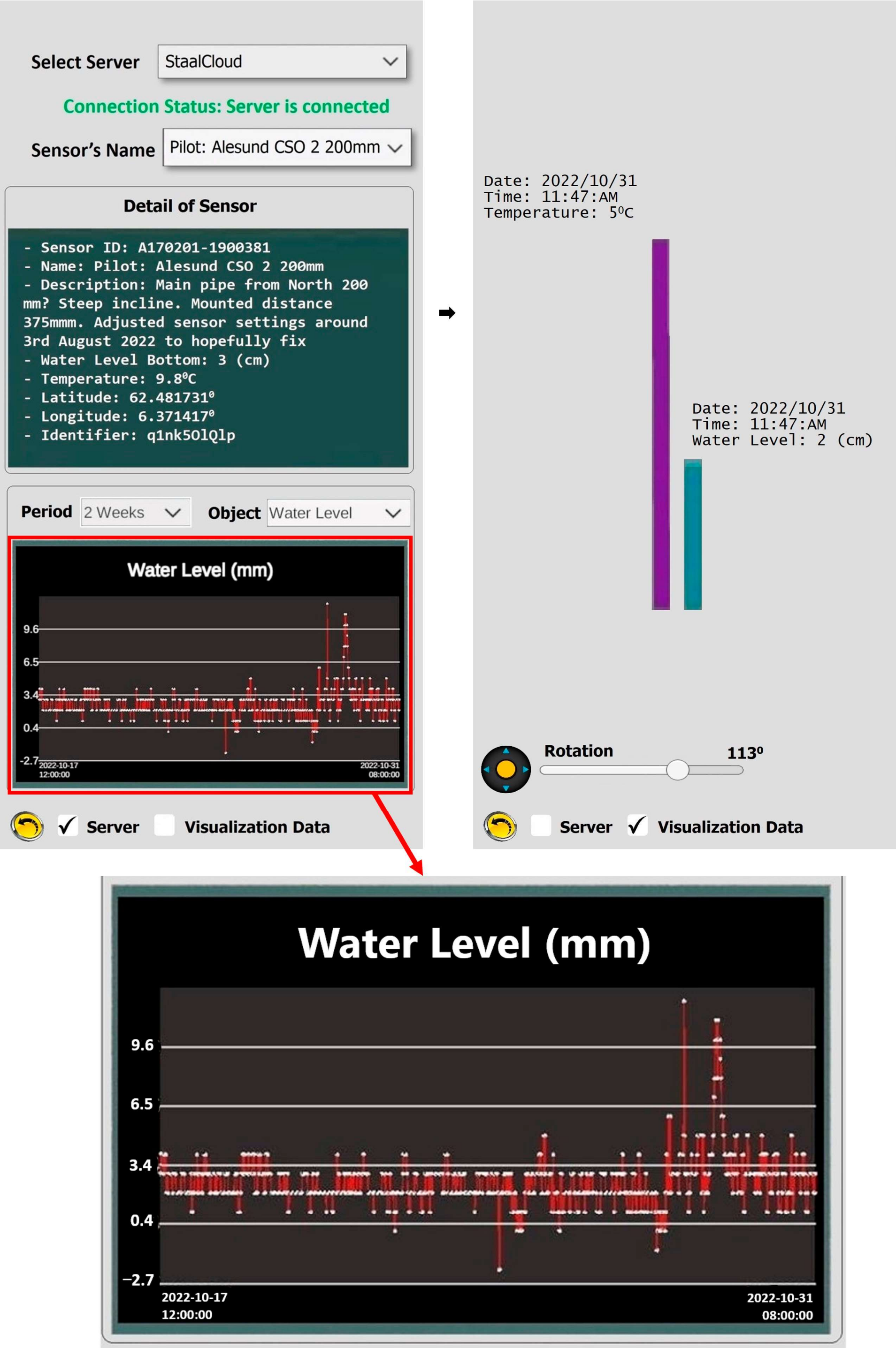
There were some outliers at manholes 58884, 112760, and 58886 (Figure 12). This may be explained by the existing high trees and electric wires that reduce signal strength from satellites to these manholes (Figure 13). For the manhole 112760, there was a high difference in the horizontal distances, although with no surrounding objects. This indicates that, in some cases, positioning accuracy with a smartphone can be unstable.

The average RMSE from the five tested manholes is shown in Table 1. The median RMSE from the tested manholes, with a 95% confidence interval, was 1.19–2.13 m for the application.

Table 1. Summary of RMSE at the tested manholes.

Manhole Name	112182	112760	58884	112765	58886
Mean RMSE (m)	1.15	2.19	1.80	1.59	1.56

As shown in Table 1, the positioning accuracy when using smartphones was significantly low, especially near trees or high buildings that blocked satellite signals [6]. Therefore, integrating handheld devices with external equipment such as orientation sensors or real-time kinematic (RTK) receivers should be considered to improve positional accuracy [13]. The application has only been developed for Android, and developments for iOS devices should be considered in future work.



Select Server: StaalCloud

Connection Status: Server is connected

Sensor's Name: Pilot: Alesund CSO 2 200mm

Detail of Sensor

```
- Sensor ID: A170201-1900381
- Name: Pilot: Alesund CSO 2 200mm
- Description: Main pipe from North 200
mm? Steep incline. Mounted distance
375mmm. Adjusted sensor settings around
3rd August 2022 to hopefully fix
- Water Level Bottom: 3 (cm)
- Temperature: 9.8ºC
- Latitude: 62.481731º
- Longitude: 6.371417º
- Identifier: q1nk5OlQlp
```

Period: 2 Weeks

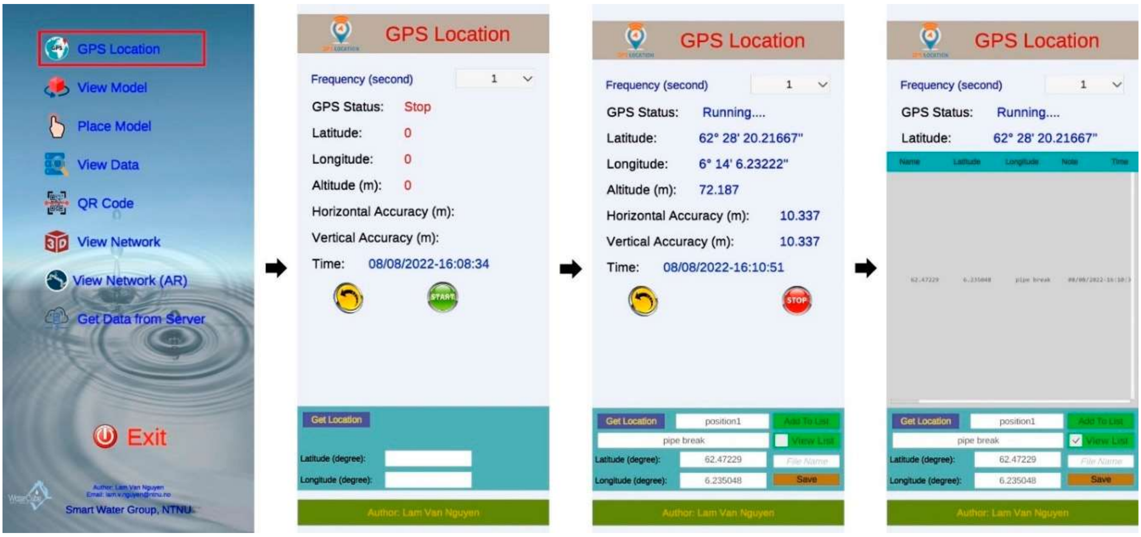
Object: Water Level

```
Date: 2022/10/31
Time: 11:47:AM
Temperature: 5ºC
```

```
Date: 2022/10/31
Time: 11:47:AM
Water Level: 2 (cm)
```

Rotation 113º

Figure 9. Real-time data access from the *StaalCloud* portal.



GPS Location

Frequency (second) 1

GPS Status: Stop

Latitude: 0

Longitude: 0

Altitude (m): 0

Horizontal Accuracy (m):

Vertical Accuracy (m):

Time: 08/08/2022-16:08:34

Latitude (degree):

Longitude (degree):

GPS Location

Frequency (second) 1

GPS Status: Running....

Latitude: 62° 28' 20.21667"

Longitude: 6° 14' 6.23222"

Altitude (m): 72.187

Horizontal Accuracy (m): 10.337

Vertical Accuracy (m): 10.337

Time: 08/08/2022-16:10:51

position1

pipe break

Latitude (degree): 62.47229

Longitude (degree): 6.235048

GPS Location

Frequency (second) 1

GPS Status: Running....

Latitude: 62° 28' 20.21667"

Name	Latitude	Longitude	Note	Time
	62.47229	6.235048	pipe break	08/08/2022-16:10:5

position1

pipe break

Latitude (degree): 62.47229

Longitude (degree): 6.235048

Figure 10. Pinpointing locations using GPS signals.

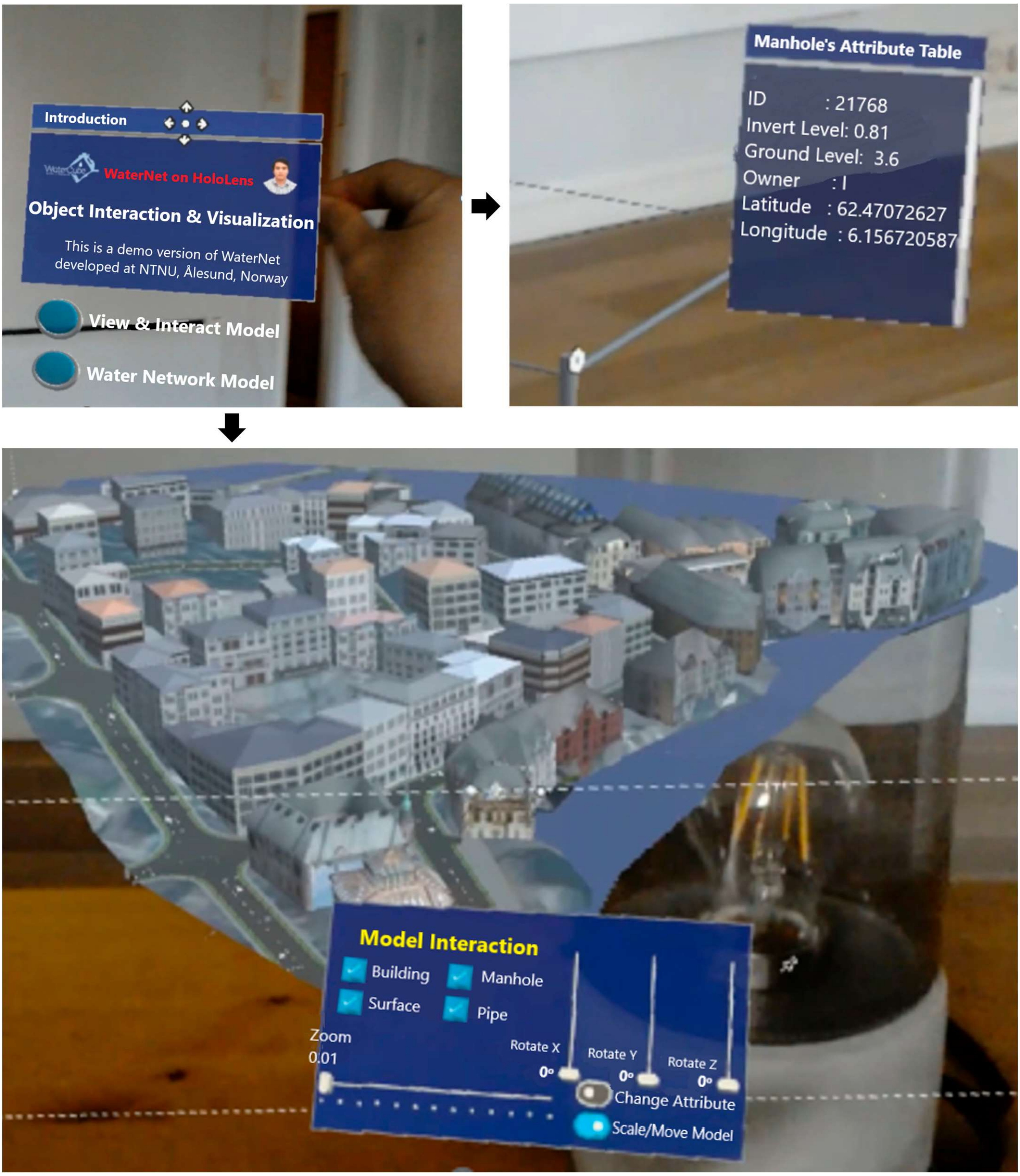
Due to the GPS error on smartphones and environmental objects [24], the positioning accuracy might not meet the desired expectations. Li, Feng, Han and Liu [6] highlighted that location-based AR applications using smartphones in previous studies achieved low accuracy, in a range of 3–10 m. The results in Table 1 show that the accuracy of the application developed in this study is equivalent to well-designed GPS receivers, which can achieve a positioning accuracy of 1–3 m [49].

Some approaches can be considered in future work to improve positioning accuracy for AR applications, such as using wireless technologies (image tracking, Bluetooth, radio frequency identification (RFID), or Wideband Code Division Multiple Access (WCDMA)/4G Long-term Evolution (LTE)) [50] or differential GPS (DGPS) techniques that can achieve centimeter-level accuracy [51].

In addition to the limitation of GPS positioning accuracy, the graphic processing capacity of mobile devices is also a drawback for the use of AR applications on a large scale. In this work, the mobile device sometimes did not work properly while loading the 3D model, possibly due to a heavy model load.

The main objective of this study was to demonstrate the promising application of smartphones and AR to visualize sewer networks and their properties using an integrated platform. In order to obtain a comprehensive assessment of accuracy using the smartphone for this purpose, many field measurements must be investigated. Moreover, differences in positioning accuracy on different mobile devices will be significant because of dissimilar hardware structures and positioning sensors [6,52].

Using this application in fieldwork requires that users pay attention to some things during the operation. For example, the function "*GPS Location*" for pinpointing geographical locations requires that mobile devices continuously receive a GPS signal, which consumes a lot of power and quickly drains their batteries [11]. Given the long duration of field investigations, users will need to be equipped with more batteries or external power sources.



WaterNet on HoloLens

Object Interaction & Visualization

This is a demo version of WaterNet developed at NTNU, Ålesund, Norway

Manhole's Attribute Table

ID : 21768

Invert Level: 0.81

Ground Level: 3.6

Owner : I

Latitude : 62.47072627

Longitude : 6.156720587

Model Interaction

Zoom 0.01

Rotate X 0°

Rotate Y 0°

Rotate Z 0°

Figure 11. Network visualization on HoloLens.

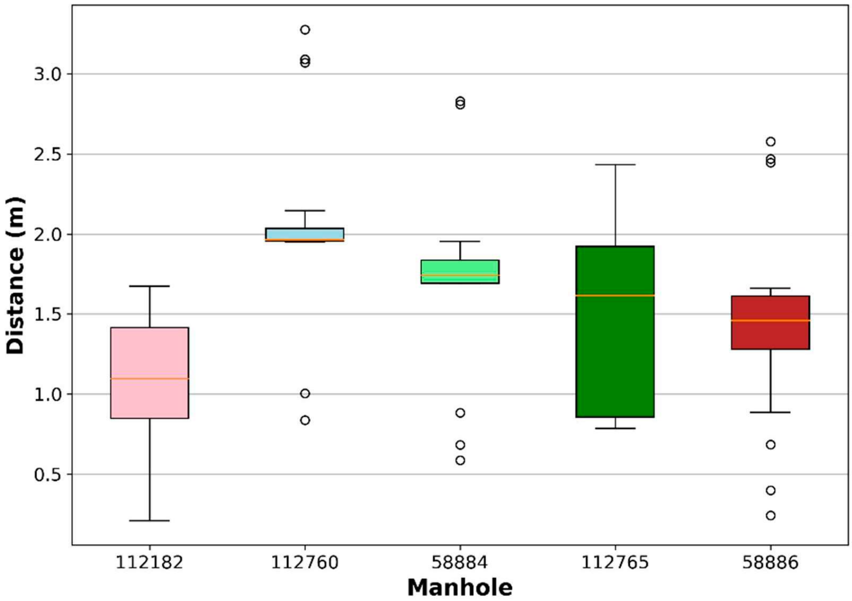


Figure 12. The differences in the horizontal distances for each tested manhole.

Minimizing the delay in field data acquisition is one of the advantages of using an AR-based application; users can quickly and easily view the objects as well as their attributes. Compared to the traditional method, this will significantly reduce time wasted in the field without requiring manually checking real-world objects with corresponding ones from the record (an easily confused process).

It is worth noting that in the dynamic visualization of the water flow in pipes, other dynamic input data (such as temperature or contamination degree) can be visualized in the same way if they have the same data structure. Moreover, mobile devices and PCs must share the same private network to transfer data. However, the status of sewer pipes obtained from models can be stored in CSV format and copied to a mobile hard drive, allowing the application to access and visualize it in the field.

The developed applications on mobile and HoloLens devices should be considered in the future to integrate more needed functions based on the visualization platform developed in this study. In addition, model optimization (in terms of capacity and detail level) for running on hardware-limited devices (for example, mobile devices or HoloLens) needs to be improved.

The experimental functions in this application were built and conducted based on the sewer network in a specific study area in Ålesund city. For other areas, the steps utilized for the study area will be replicated. This paper presented the fundamental steps for sewer network 3D visualization on handheld devices such as smartphones and HoloLens, and a circle interaction between this application and computer/server to transfer and process data for sewer management purposes was introduced. Customized functions can be developed for specific purposes. Technical and non-technical operators can consult this application as an assistant tool for collecting and visualizing data in the water sector.

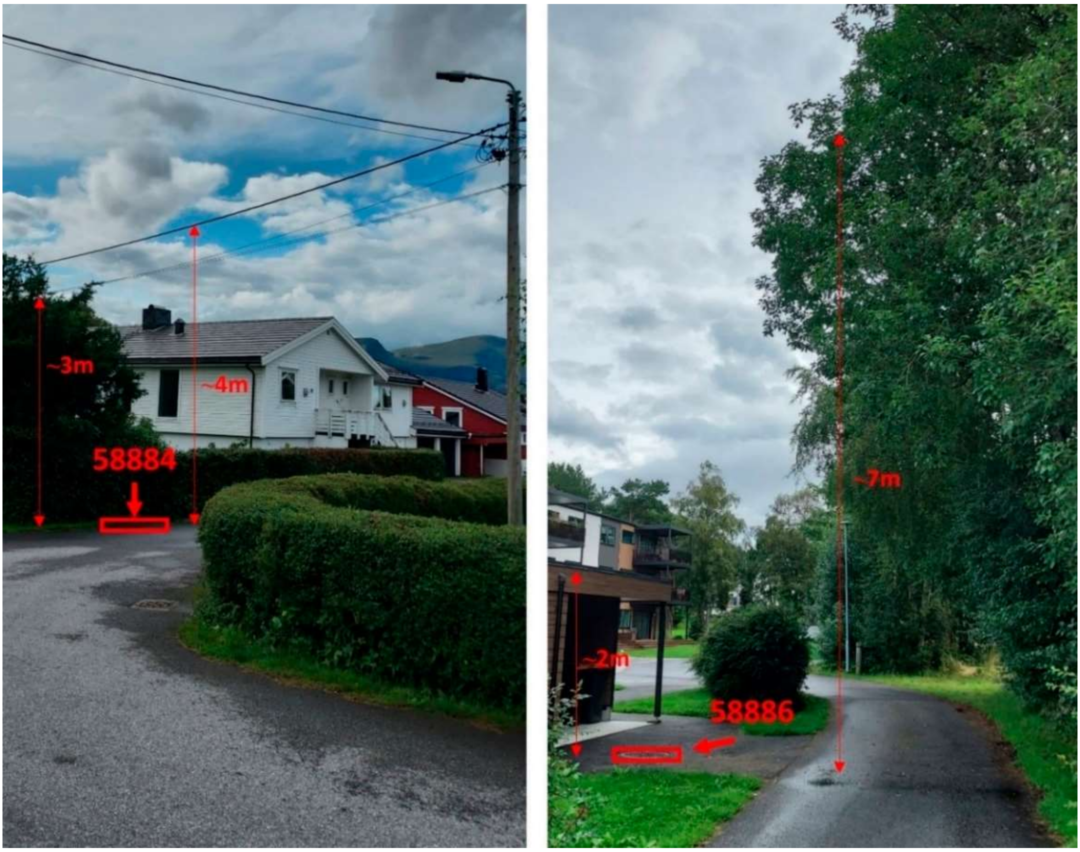


Figure 13. Noise background surrounding manholes.

4. Conclusions and Outlook

The outputs of this work reveal the potential integration of mobile devices, GIS, and AR techniques in the management of water infrastructure. The users' awareness of water infrastructure and the surrounding environment is enhanced in a way that allows the exploration of the relations between them.

The application is currently developed for mobile phones and has the following functionalities:

- Allows for the collection of data on pipe and environmental attributes, processing, condition assessment, and visualization of pipe status on a mobile device;
- Can be used by operators for the 3D visualization of buried sewer pipes, including their attributes and conditions;
- Allows for the real-time visualization of dynamic data (e.g., water flow, water temperature) of the pipes through integration of geo-pipe locations and sensor data;
- Can be used in the field for purposes of asset management.

A limitation of the application is the accuracy of the visualized and existing pipe infrastructure. Integrating handheld devices with external equipment such as orientation sensors or RTK technology will significantly enhance the positional accuracy of the system.

Author Contributions: Conceptualization, L.V.N. and R.S.; methodology, L.V.N. and R.S.; software, L.V.N.; validation, L.V.N.; formal analysis, L.V.N.; investigation, L.V.N.; writing—original draft preparation, L.V.N.; writing—review and editing, R.S. and D.T.B.; visualization, L.V.N.; supervision,

R.S. and D.T.B.; project administration, R.S.; funding acquisition, R.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Smart Water Project, grant number 90392200, financed by Ålesund Municipality and Norwegian University of Science and Technology (NTNU), Norway.

Data Availability Statement: Data will be made available on request.

Acknowledgments: The authors would like to thank the Norwegian Mapping Authority and Ålesund municipality for providing data for this research.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

GIS	Geographic Information System
AR	Augmented Reality
BIM	Building Information Modeling
GPR	Ground Penetrating Radar
GPS	Global Positioning System
GNSS	Global Navigation Satellite System
API	Application Programming Interface
ML	Machine Learning
ESRI	Environmental Systems Research Institute, Inc.
CSV	Comma-Separated Values
RMSE	Root Mean Square Error
PC	Personal Computer
WGS-84	World Geodetic System 1984
RTK	Real-Time Kinematic
RFID	Radio Frequency Identification
WCDMA	Wideband Code Division Multiple Access
LTE	Long-term Evolution
DGPS	Differential GPS

References

1. Statistics, N. Municipal Wastewater. Available online: https://www.ssb.no/en/natur-og-miljo/vann-og-avlop/statistikk/utslipp-og-rensing-av-kommunalt-avlop (accessed on 28 October 2023).
2. Fugledalen, T.; Rokstad, M.M.; Tscheikner-Gratl, F. On the influence of input data uncertainty on sewer deterioration models—A case study in Norway. *Struct. Infrastruct. Eng.* **2021**, *19*, 1064–1075. [CrossRef]
3. Nguyen, L.V.; Seidu, R. Application of Regression-Based Machine Learning Algorithms in Sewer Condition Assessment for Ålesund City, Norway. *Water* **2022**, *14*, 3993. [CrossRef]
4. Nguyen, L.V.; Bui, D.T.; Seidu, R. Comparison of Machine Learning Techniques for Condition Assessment of Sewer Network. *IEEE Access* **2022**, *10*, 124238–124258. [CrossRef]
5. Nguyen, L.V.; Seidu, R. Predicting sewer structural condition using hybrid machine learning algorithms. *Urban Water J.* **2023**, *20*, 882–896. [CrossRef]
6. Li, M.; Feng, X.; Han, Y.; Liu, X. Mobile augmented reality-based visualization framework for lifecycle O&M support of urban underground pipe networks. *Tunn. Undergr. Space Technol.* **2023**, *136*, 21. [CrossRef]
7. Nguyen, L.S.; Schaeli, B.; Sage, D.; Kayal, S.; Jeanbourquin, D.; Barry, D.A.; Rossi, L. Vision-based system for the control and measurement of wastewater flow rate in sewer systems. *Water Sci. Technol.* **2009**, *60*, 2281–2289. [CrossRef]
8. Bottani, E.; Vignali, G. Augmented reality technology in the manufacturing industry: A review of the last decade. *IISE Trans.* **2019**, *51*, 284–310. [CrossRef]
9. Chen, Y.; Wang, Q.; Chen, H.; Song, X.; Tang, H.; Tian, M. An overview of augmented reality technology. *J. Phys. Conf. Ser.* **2019**, *1237*, 6. [CrossRef]
10. Fite-Georgel, P. Is there a reality in Industrial Augmented Reality? In Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality, Basel, Switzerland, 26–29 October 2011; pp. 201–210.
11. Mirauda, D.; Erra, U.; Agatiello, R.; Cerverizzo, M. Applications of Mobile Augmented Reality to Water Resources Management. *Water* **2017**, *9*, 699. [CrossRef]

12. Centeno, J.A.S.; Kishi, R.T.; Mitishita, E.A. Three-dimensional Data Visualization in Water Quality Studies using Augmented Reality. In Proceedings of the 6th International Symposium on Mobile Mapping Technology, São Paulo, Brazil, 21–24 July 2009.
13. Schall, G.; Zollmann, S.; Reitmayr, G. Smart Vidente: Advances in mobile augmented reality for interactive visualization of underground infrastructure. *Pers. Ubiquitous Comput.* **2013**, *17*, 1533–1549. [CrossRef]
14. Haynes, P.; Hehl-Lange, S.; Lange, E. Mobile Augmented Reality for Flood Visualisation. *Environ. Model. Softw.* **2018**, *109*, 380–389. [CrossRef]
15. Hahmann, S.; Burghardt, D. How much information is geospatially referenced? Networks and cognition. *Int. J. Geogr. Inf. Sci.* **2013**, *27*, 1171–1189. [CrossRef]
16. Hunter, J.D. Matplotlib: A 2D Graphics Environment. *Comput. Sci. Eng.* **2007**, *9*, 90–95. [CrossRef]
17. Villanueva, R.A.M.; Chen, Z.J. ggplot2: Elegant Graphics for Data Analysis (2nd ed.). *Meas. Interdiscip. Res. Perspect.* **2019**, *17*, 160–167. [CrossRef]
18. Kjellin, A.; Pettersson, L.W.; Seipel, S.; Lind, M. Evaluating 2D and 3D visualizations of spatiotemporal information. *ACM Trans. Appl. Percept.* **2008**, *7*, 1–23. [CrossRef]
19. Dübel, S.; Röhlig, M.; Schumann, H.; Trapp, M. 2D and 3D presentation of spatial data: A systematic review. In Proceedings of the 2014 IEEE VIS International Workshop on 3DVis (3DVis), Paris, France, 9 November 2014; pp. 11–18.
20. Beha, F.; Göritz, A.; Schildhauer, T. Business model innovation: The role of different types of visualizations. In Proceedings of the ISPIM Conference Proceedings, Hamburg, Germany, 14–17 June 2015; p. 19.
21. Huston, D.; Xia, T.; Zhang, Y.; Fan, T.; Orfeo, D.; Razinger, J. Urban underground infrastructure mapping and assessment. In Proceedings of the Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems 2017, Portland, OR, USA, 25–29 March 2017; p. 11.
22. Fang, X.; Li, Q.; Zhu, J.; Chen, Z.; Zhang, D.; Wu, K.; Ding, K.; Li, Q. Sewer defect instance segmentation, localization, and 3D reconstruction for sewer floating capsule robots. *Autom. Constr.* **2022**, *142*, 104494. [CrossRef]
23. Soria, G.; Ortega Alvarado, L.M.; Feito, F.R. Augmented and Virtual Reality for Underground Facilities Management. *J. Comput. Inf. Sci. Eng.* **2018**, *18*, 9. [CrossRef]
24. Fenais, A.; Ariaratnam, S.T.; Ayer, S.K.; Smilovsky, N. Integrating Geographic Information Systems and Augmented Reality for Mapping Underground Utilities. *Infrastructures* **2019**, *4*, 60. [CrossRef]
25. Pereira, M.; Burns, D.; Orfeo, D.; Farrel, R.; Hutson, D.; Xia, T. New GPR System Integration with Augmented Reality Based Positioning. In Proceedings of the 2018 on Great Lakes Symposium on VLSI, Chicago, IL, USA, 23–25 May 2018; pp. 341–346.
26. Jimenez, R.J.P.; Becerril, E.M.D.; Nor, R.M.; Smagas, K.; Valari, E.; Stylianidis, E. Market potential for a location based and augmented reality system for utilities management. In Proceedings of the 2016 22nd International Conference on Virtual System & Multimedia (VSMM), Kuala Lumpur, Malaysia, 17–21 October 2016; pp. 1–4.
27. Kim, B.-h. Development of Augmented Reality Underground Facility Management System using Map Application Programming Interface and JavaScript Object Notation Communication. *Teh. Vjesn.* **2023**, *30*, 797–803. [CrossRef]
28. Tarek, H.; Marzouk, M. Integrated Augmented Reality and Cloud Computing Approach for Infrastructure Utilities Maintenance. *J. Pipeline Syst. Eng. Pract.* **2022**, *13*, 11. [CrossRef]
29. Côté, S.; Mercier, A. Augmentation of Road Surfaces with Subsurface Utility Model Projections. In Proceedings of the 2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Tuebingen/Reutlingen, Germany, 18–22 March 2018; pp. 535–536.
30. Rahman, A.; Xi, M.; Dabrowski, J.J.; McCulloch, J.; Arnold, S.; Rana, M.; George, A.; Adcock, M. An integrated framework of sensing, machine learning, and augmented reality for aquaculture prawn farm management. *Aquac. Eng.* **2021**, *95*, 102192. [CrossRef]
31. Haugen, H.J.; Viak, A. *Datafl yt—Klassifi Sering av Avløpsledninger*; Norwegian Water BA: Hamar, Norway, 2018.
32. Trimble. Trimble to Enhance its Office-to-Field Platform with the Acquisition of Google's SketchUp 3D Modeling Platform. Available online: https://investor.trimble.com/news-releases/news-release-details/trimble-enhance-its-office-field-platform-acquisition-googles?releaseid=667690 (accessed on 19 October 2022).
33. Bassett, T.; Lannon, S.C.; Waldron, D.; Jones, P.J. Calculating the solar potential of the urban fabric with SketchUp and HTB2. In Proceedings of the Solar Building Skins, Bressanone, Italy, 6–7 December 2012.
34. Lewis, G.M.; Hampton, S.J. Visualizing volcanic processes in SketchUp: An integrated geo-education tool. *Comput. Geosci.* **2015**, *81*, 93–100. [CrossRef]
35. Jusuf, S.K.; Ignatius, M.; Wong, N.H.; Tan, E. STEVE Tool Plug-in for SketchUp: A User-Friendly Microclimatic Mapping Tool for Estate Development. In *Sustainable Building and Built Environments to Mitigate Climate Change in the Tropics: Conceptual and Practical Approaches, Karyono, T.H., Vale, R., Vale, B., Eds.*; Springer International Publishing: Cham, Switzerland, 2017; pp. 113–130.
36. Burner, D.M.; Ashworth, A.J.; Laughlin, K.F.; Boyer, M.E. Using SketchUp to Simulate Tree Row Azimuth Effects on Alley Shading. *Agron. J.* **2018**, *110*, 425–430. [CrossRef]
37. Ma, Z.; Ren, Y. Integrated Application of BIM and GIS: An Overview. *Procedia Eng.* **2017**, *196*, 1072–1079. [CrossRef]
38. Kuok, K.K.; Kingston Tan, K.W.; Chiu, P.C.; Chin, M.Y.; Rahman, M.R.; Bin Bakri, M.K. *Application of Building Information Modelling (BIM) Technology in Drainage System Using Autodesk InfraWorks 360 Software*; Springer Nature: Singapore, 2022; pp. 209–224.
39. Barazzetti, L. *Integrated BIM-GIS Model Generation at the City Scale Using Geospatial Data*; SPIE: Bellingham, WA, USA, 2018; Volume 10773.

40. Hocking, J.; Schell, J. *Unity in Action: Multiplatform Game Development in C#*, 3rd ed.; Manning Publications Co.: Shelter Island, NY, USA, 2022; p. 416.
41. Juliani, A.; Berges, V.-P.; Teng, E.; Cohen, A.; Harper, J.; Elion, C.; Goy, C.; Gao, Y.; Henry, H.; Mattar, M. Unity: A General Platform for Intelligent Agents. *arXiv* **2018**, arXiv:1809.02627. [CrossRef]
42. Han, Y.-S.; Lee, J.; Lee, J.; Lee, W.; Lee, K. 3D CAD data extraction and conversion for application of augmented/virtual reality to the construction of ships and offshore structures. *Int. J. Comput. Integr. Manuf.* **2019**, *32*, 658–668. [CrossRef]
43. Samsung. Galaxy A42 5G. Available online: https://www.samsung.com/us/smartphones/galaxy-a42-5g/ (accessed on 31 October 2022).
44. Microsoft. HoloLens (1st gen) Hardware. Available online: https://learn.microsoft.com/en-us/hololens/hololens1-hardware (accessed on 1 November 2022).
45. Duque, N.; Duque, D.; Aguilar, A.; Saldarriaga, J. Sewer Network Layout Selection and Hydraulic Design Using a Mathematical Optimization Framework. *Water* **2020**, *12*, 3337. [CrossRef]
46. Zeiler, M. *Modeling Our World: The ESRI Guide to Geodatabase Design*; Environmental Systems Research Institute, Inc.: Redlands, CA, USA, 1999; Volume 40.
47. Wang, W.; Wu, X.; Chen, G.; Chen, Z. Holo3DGIS: Leveraging Microsoft HoloLens in 3D Geographic Information. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 60. [CrossRef]
48. Microsoft. Unity Development for HoloLens. Available online: https://learn.microsoft.com/en-us/windows/mixed-reality/develop/unity/unity-development-overview?tabs=arr%252CD365%252Chl2 (accessed on 1 November 2022).
49. Renfro, B.A.; Stein, M.; Boeker, N.; Terry, A. An Analysis of Global Positioning System (GPS) Standard Positioning Service (SPS) Performance for 2017. 2018. Available online: https://www.gps.gov/systems/gps/performance/2018-GPS-SPS-performance-analysis.pdf (accessed on 22 November 2023).
50. Jian, M.; Wang, Y.; Wu, B.; Cheng, Y. Hybrid cloud computing for user location-aware augmented reality construction. In Proceedings of the 2018 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon, Republic of Korea, 11–14 February 2018; pp. 190–194.
51. Chen, Y.; Zhao, S.; Farrell, J.A. Computationally efficient carrier integer ambiguity resolution in multiepoch GPS/INS: A common-position-shift approach. *IEEE Trans. Control Syst. Technol.* **2015**, *24*, 1541–1556. [CrossRef]
52. Blum, J.R.; Greencorn, D.G.; Cooperstock, J.R. Smartphone Sensor Reliability for Augmented Reality Applications. In *Mobile and Ubiquitous Systems: Computing, Networking, and Services*; MobiQuitous 2012; Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering; Springer: Berlin/Heidelberg, Germany, 2013; pp. 127–138.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

Appendix B

Application Guideline

NORWEGIAN UNIVERSITY OF SCIENCE AND TECHNOLOGY
Faculty of Engineering
Department of Ocean Operations and Civil Engineering

NTNU

Guideline for using Visualization Application

This guideline is a supporting document of the first author's PhD thesis:

Integrating Machine Learning and GIS for Sewer Condition Assessment and Visualization

Lam Van Nguyen

Water and Environmental Engineering Group
NTNU Ålesund

12/2022

Objectives of the guideline

The overall aim of this guideline is to provide a user with a step-by-step implementation for using a visualization application of sewer conditions on mobile phones and HoloLens devices. This platform is only used for a visualization based on the specific objects relating to the first author's PhD. thesis.

B.1. Mobile Application

B.1.1. Downloading and installing applications

There are two versions of the mobile application, named "*WaterNet on GoogleMap*" and "*WaterNet*". The downloaded links of these applications are provided in **Table B.1**. The basic difference between the two versions is described as follows:

- The "*WaterNet on GoogleMap*" application uses the Google map as a background image and the components of the sewer network (for example, manholes or pipes) are overlayed on this image. This 2D visualization application was designed for supporting the users to easily look for components of the sewer network on the field.
- The "*WaterNet*" application supports 2D and 3D visualizations of sewer components and presents predicted results of the sewer condition.

Table B.1. Downloaded links to the applications on android devices

Name	Download link
WaterNet_Google Map	https://www.mediafire.com/file/lw2gote7dwjmgl3/WaterNetOnGoogleMap.apk/file
WaterNet	https://www.mediafire.com/file/ua4qzwb2582bb6b/WaterNet.apk/file
Data sample	https://www.mediafire.com/file/okq6nn95frc41ic/Data_Sample.zip/file
Python code	https://www.mediafire.com/file/piwvybtyc2uy7b4/Unity_Python.py/file

These applications are distributed as an Android package with the file extension **APK**. Therefore, these applications are only compatible with android devices. After downloading applications from the above links, the users easily install these applications as any android application by clicking on them.

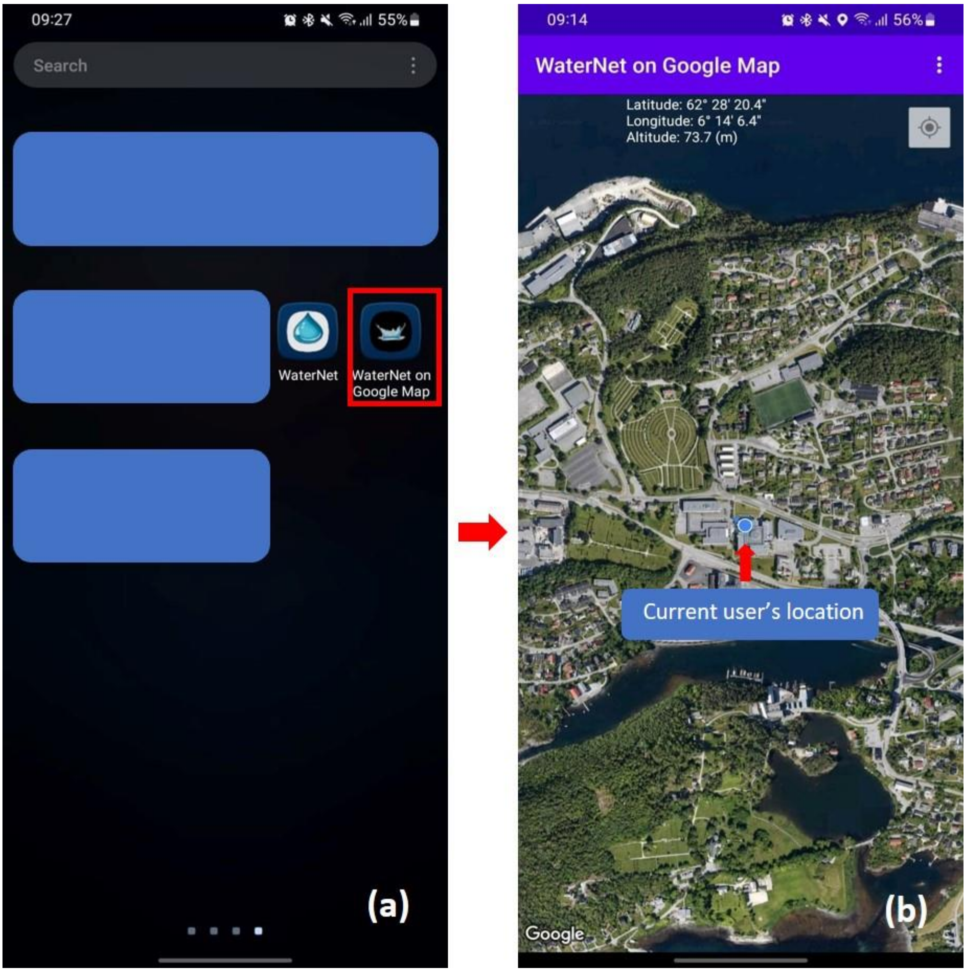
The interfaces of these applications are currently optimized for android phones with a screen resolution of 720 pixels x 1600 pixels. Other android devices with different screen sizes maybe have slightly different experiences.

B.1.2. The "WaterNet on GoogleMap" application

After installing the "*WaterNet on GoogleMap*" application on the android device, the symbol will appear on the menu of the device (**Figure B.1**a).

To get the geographical coordinates of the user's position, the GPS function on the device should be enabled before the application is opened. After the application is activated, the current position of the user is shown in the center of the screen (the blue dot) and the geographical coordinates (including latitude, longitude, and altitude) of the user also are shown on the screen (**Figure B.1**b).

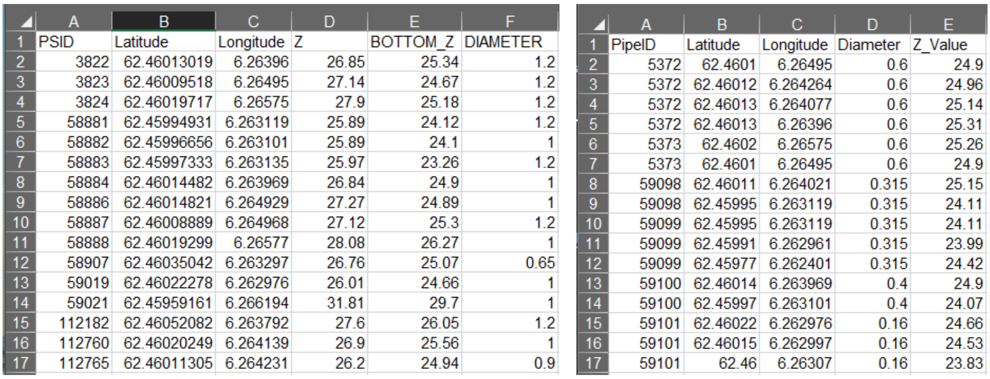
When the user moves, the user's location and geographical coordinates are updated in real-time. The user also moves the screen from any position to the current position by clicking on the symbol ⌖ in the top-right corner of the screen (**Figure B.1**b).



Latitude: 62° 28' 20.4"
Longitude: 6° 14' 6.4"
Altitude: 73.7 (m)

Figure B.1. The interface of the *"WaterNet on GoogleMap"* application

This application allows the user to import some components of the sewer network (such as manholes and pipes) from CSV format files. The structure of CSV files is shown in **Figure B.2**.



	A	B	C	D	E	F
1	PSID	Latitude	Longitude	Z	BOTTOM_Z	DIAMETER
2	3822	62.46013019	6.26396	26.85	25.34	1.2
3	3823	62.46009518	6.26495	27.14	24.67	1.2
4	3824	62.46019717	6.26575	27.9	25.18	1.2
5	58881	62.45994931	6.263119	25.89	24.12	1.2
6	58882	62.45996656	6.263101	25.89	24.1	1
7	58883	62.45997333	6.263135	25.97	23.26	1.2
8	58884	62.46014482	6.263969	26.84	24.9	1
9	58886	62.46014821	6.264929	27.27	24.89	1
10	58887	62.46008889	6.264968	27.12	25.3	1.2
11	58888	62.46019299	6.26577	28.08	26.27	1
12	58907	62.46035042	6.263297	26.76	25.07	0.65
13	59019	62.46022278	6.262976	26.01	24.66	1
14	59021	62.45959161	6.266194	31.81	29.7	1
15	112182	62.46052082	6.263792	27.6	26.05	1.2
16	112760	62.46020249	6.264139	26.9	25.56	1
17	112765	62.46011305	6.264231	26.2	24.94	0.9

(a) Sewer's manhole data format

	A	B	C	D	E
1	PipeID	Latitude	Longitude	Diameter	Z_Value
2	5372	62.4601	6.26495	0.6	24.9
3	5372	62.46012	6.264264	0.6	24.96
4	5372	62.46013	6.264077	0.6	25.14
5	5372	62.46013	6.26396	0.6	25.31
6	5373	62.4602	6.26575	0.6	25.26
7	5373	62.4601	6.26495	0.6	24.9
8	59098	62.46011	6.264021	0.315	25.15
9	59098	62.45995	6.263119	0.315	24.11
10	59099	62.45995	6.263119	0.315	24.11
11	59099	62.45991	6.262961	0.315	23.99
12	59099	62.45977	6.262401	0.315	24.42
13	59100	62.46014	6.263969	0.4	24.9
14	59100	62.45997	6.263101	0.4	24.07
15	59101	62.46022	6.262976	0.16	24.66
16	59101	62.46015	6.262997	0.16	24.53
17	59101	62.46	6.26307	0.16	23.83

(b) Sewer's pipe data format

Figure B.2. Example of the data format used in the "*WaterNet on GoogleMap*" application

Figure B.3 illustrates the steps for importing manholes and pipes from CSV files into the application. By coloring different characteristics (for example, material or network type) of each pipe, the user easily distinguishes sewer pipes in the field using this application.

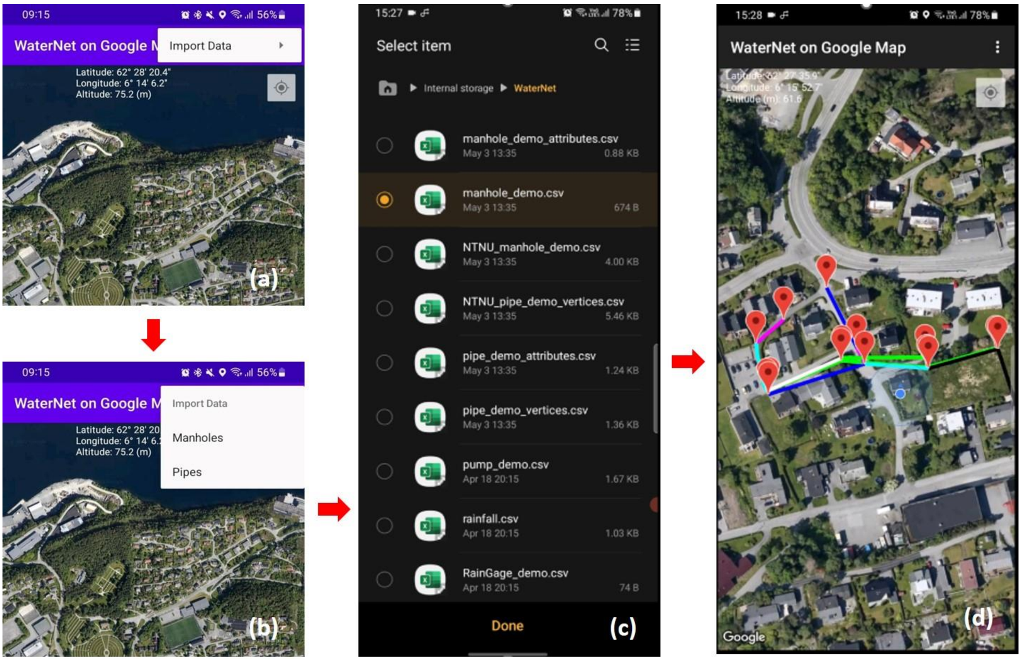


Figure B.3. Importing sewer components into the application

Figure B.4 shows the process of random checking of the "*WaterNet on GoogleMap*" application. The experimental results show that the difference between the actual sewer's manhole location and their visualized location on the "*WaterNet on GoogleMap*" application is smaller than 3 m. This difference satisfies for identifying sewer manholes in the field.

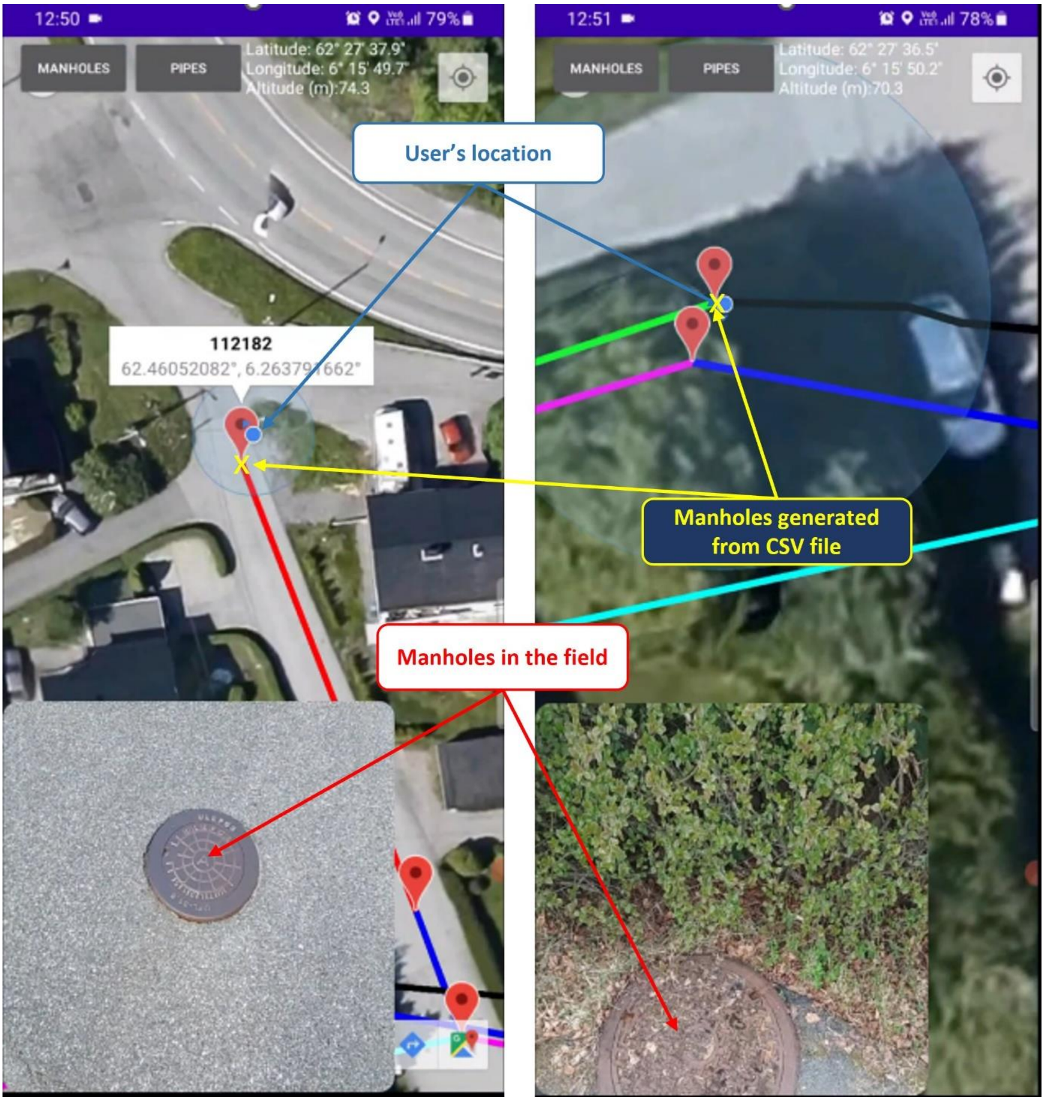


Figure B.4. Accuracy of the "*WaterNet on GoogleMap*" application

It is worth noting that other point-based objects (such as locations of pumps, outfalls, dividers, or storage units) can be imported using the "*Manholes*" function in this application.

B.1.3. The "WaterNet" application

The "*WaterNet*" application was designed for 3D visualization, integration of 3D models, and results predicted from machine learning models and real-time visualization. The interface of the "*WaterNet*" application is shown in **Figure B.5**.

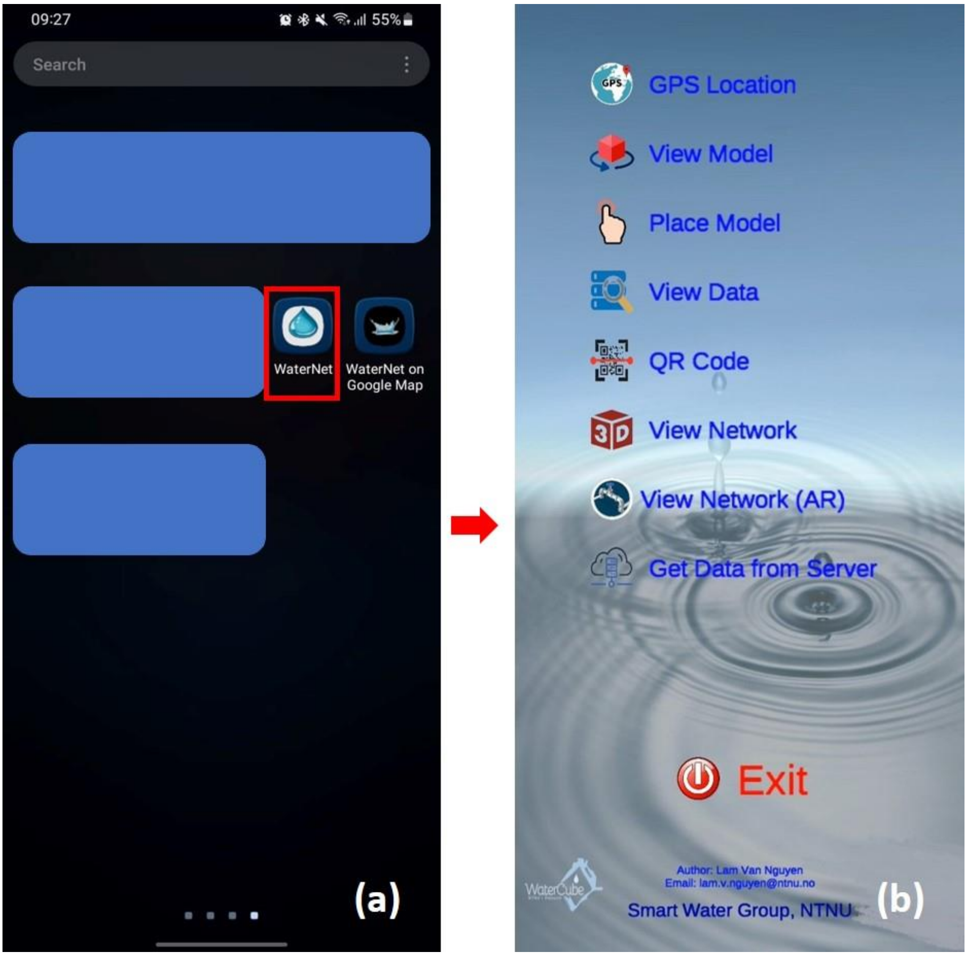
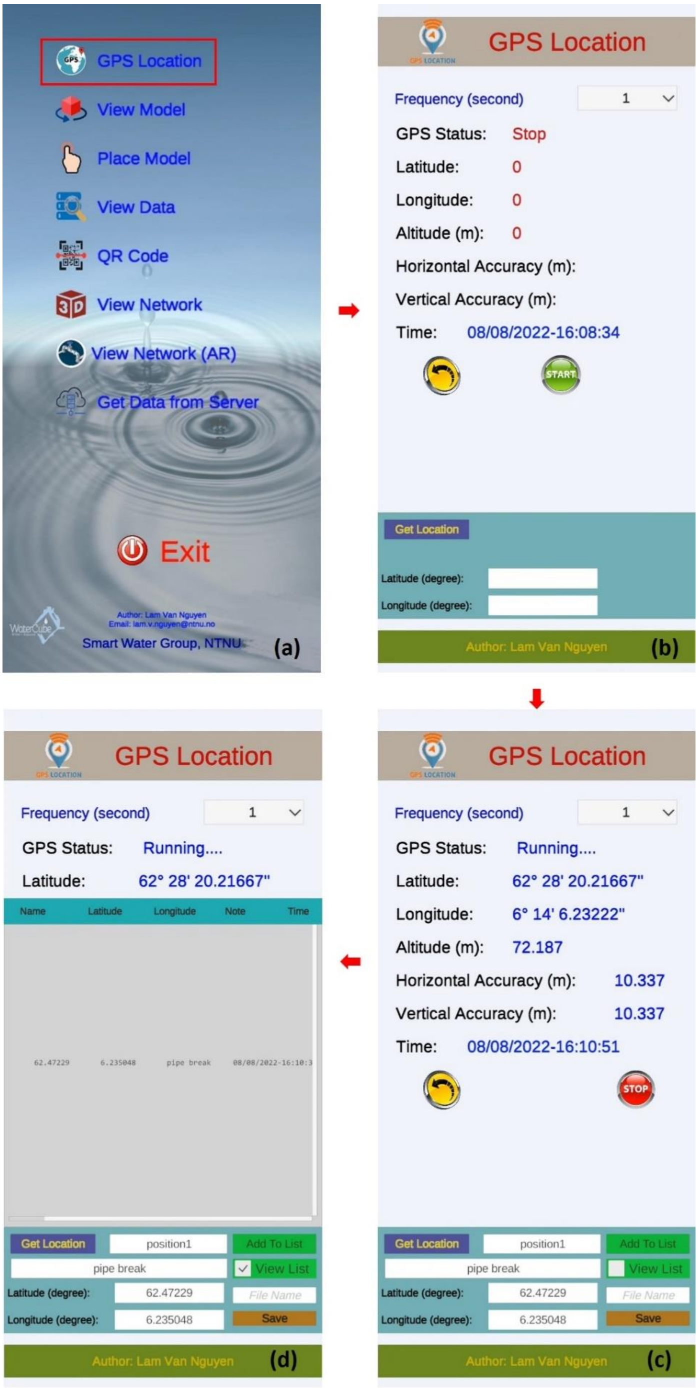


Figure B.5. The interface of the "*WaterNet*" application

The "*WaterNet*" application provides the user with some options to visualize data that are presented as follows:

B.1.3.1. GPS Location

This function allows the user to locate the device's location in the World Geodetic System 1984 (WGS84) coordinate reference from Global Positioning System (GPS) signal. This function is useful to pinpoint the problems in the field (such as cracks' locations or notice points) and the output can be saved in the Comma-Separated Values (CSV) type that is easily opened by a text editor such as Notepad or Microsoft Excel. The user can use this function to store notes in the field in CSV format on their device, they after that can transfer data from the android device to a personal computer (PC) and process them for different purposes. The basic steps for implementing this function are shown in **Figure B.6**.



(b)

Frequency (second): 1

GPS Status: Stop

Latitude: 0

Longitude: 0

Altitude (m): 0

Horizontal Accuracy (m):

Vertical Accuracy (m):

Time: 08/08/2022-16:08:34

Latitude (degree):

Longitude (degree):

(c)

Frequency (second): 1

GPS Status: Running....

Latitude: 62° 28' 20.21667"

Longitude: 6° 14' 6.23222"

Altitude (m): 72.187

Horizontal Accuracy (m): 10.337

Vertical Accuracy (m): 10.337

Time: 08/08/2022-16:10:51

position1

pipe break

Latitude (degree): 62.47229

Longitude (degree): 6.235048

(d)

Frequency (second): 1

GPS Status: Running....

Latitude: 62° 28' 20.21667"

Name	Latitude	Longitude	Note	Time
	62.47229	6.235048	pipe break	08/08/2022-16:10:3

position1

pipe break ☑

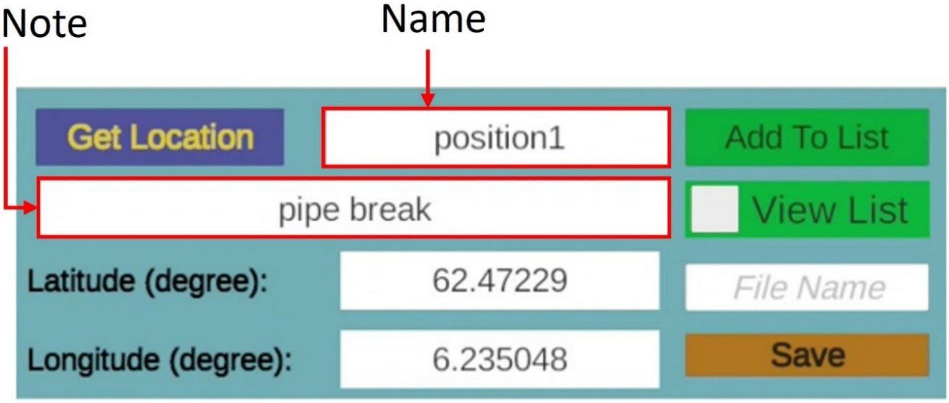
Latitude (degree): 62.47229

Longitude (degree): 6.235048

Figure B.6. Receiving GPS signal in the "*WaterNet*" application

The detailed steps for implementing this function are presented as follows:

- *Step 1:* Activate this function by selecting the symbol GPS Location on the device screen.
- *Step 2:* Define the frequency (in seconds) that the user wants to collect the GPS location.
- *Step 3:* Select the button START to begin receiving GPS signal for coordinate definition.
- *Step 4:* Click on the button "*Get Location*" at the wanted positions to get latitude and longitude. The user can set up names and notes for pinpointed locations (**Figure B.7**).



position1

pipe break

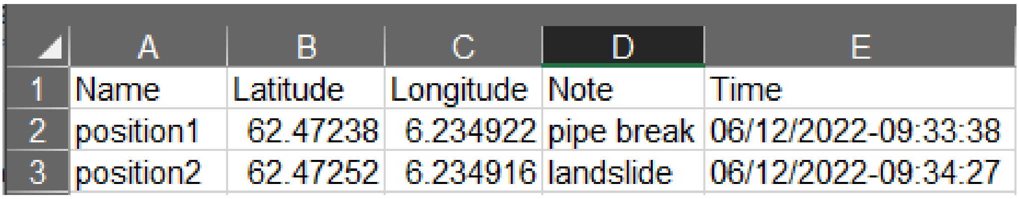
Latitude (degree): 62.47229

Longitude (degree): 6.235048

Figure B.7. Pinpointing in the "*WaterNet*" application

- *Step 5:* Select the button "*Add To List*" to add pinpointed location to the database. The user also views the points that are saved in the list by clicking on the checkbox "*View List*". The list of the point is directly shown on the device screen (**Figure B.6**d). The user can use the finger to move and see all information about the saved points. Based on this list, the user can check the information of the saved points or take more measurements on the field.
- *Step 6:* After the user finishes measuring in the field, the data file must be named in the area File Name (**Figure B.7**). To save data in the storage of the device, the user must click on the button "*Save*" in **Figure B.7**.
- *Step 7*: To finish this function, the user clicks on the button STOP. Selecting the button ↶ will quit this function and the user will be delivered to the main screen (**Figure B.5**).

The saved data from **Figure B.6**d is transferred to a PC, an example of the structure of this data is shown in **Figure B.8**.



	A	B	C	D	E
1	Name	Latitude	Longitude	Note	Time
2	position1	62.47238	6.234922	pipe break	06/12/2022-09:33:38
3	position2	62.47252	6.234916	landslide	06/12/2022-09:34:27

Figure B.8. An example of a record created by the "*WaterNet*" application

B.1.3.2. View Model

This function provides an option for the user to view a 3D model from assigned images. The steps for implementing this function are shown in **Figure B.9**. In this function, by using different assigned images, the user can view the different corresponding 3D models.

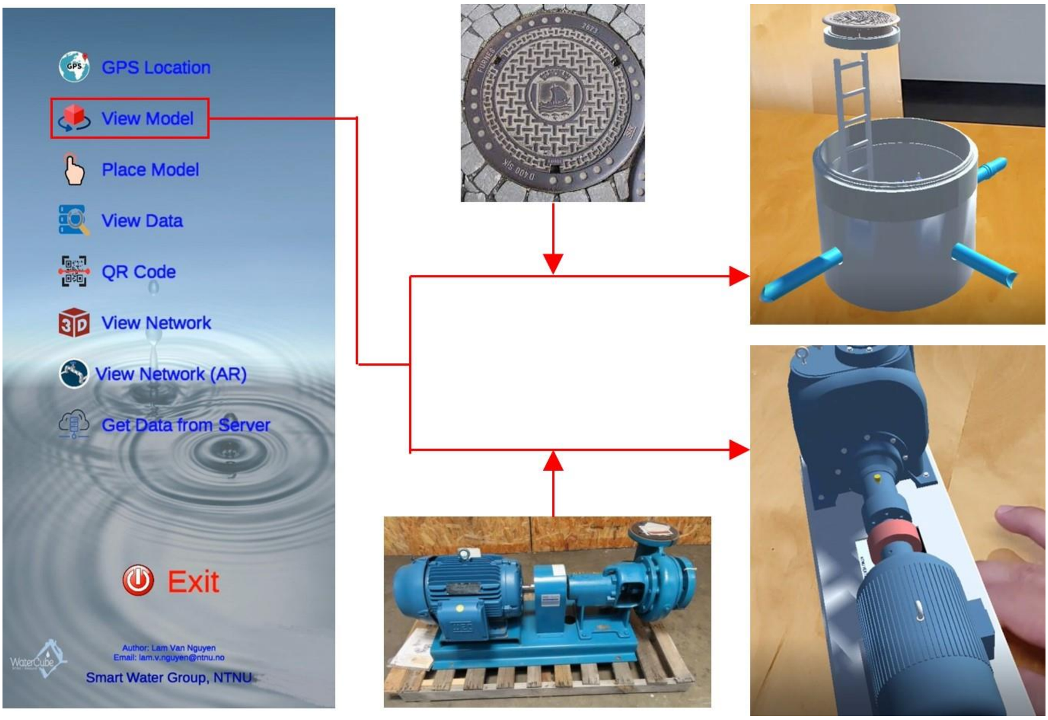


Figure B.9. Illustration of viewing the 3D model in the "*WaterNet*" application

B.1.3.3. Place Model

This function provides an option to view 3D objects on a real scale using an Augmented Reality perspective (**Figure B.10**).

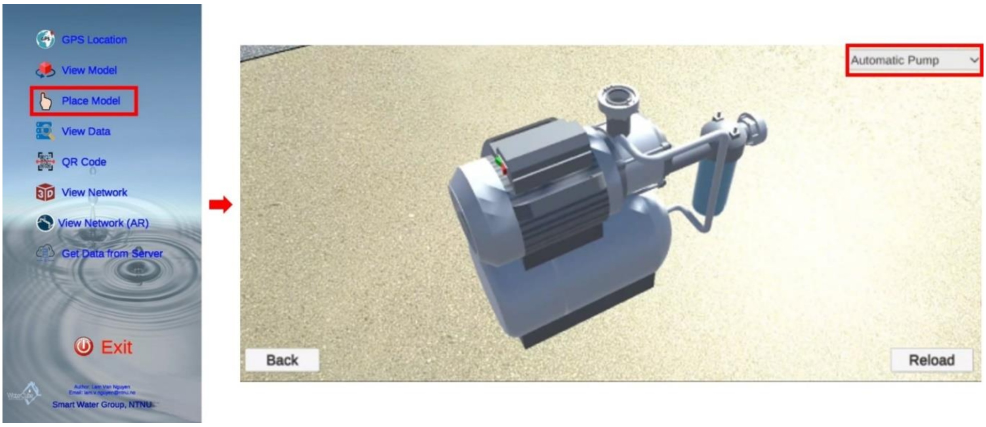


Figure B.10. Example of placing the 3D model in the "*WaterNet*" application

The user can view different objects by using the combo box in the top-right corner of the screen. This function provides two different 3D models to view: the pump (**Figure B.11**a) and the fire hydrant (**Figure B.11**b). It is worth noting that other 3D models can be further added depending on the specific purposes.

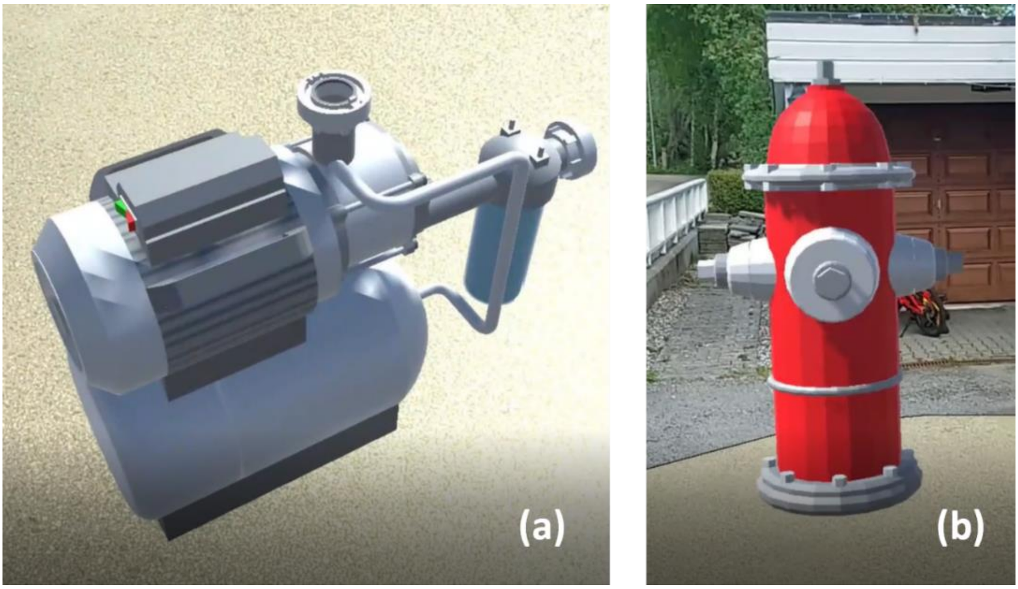
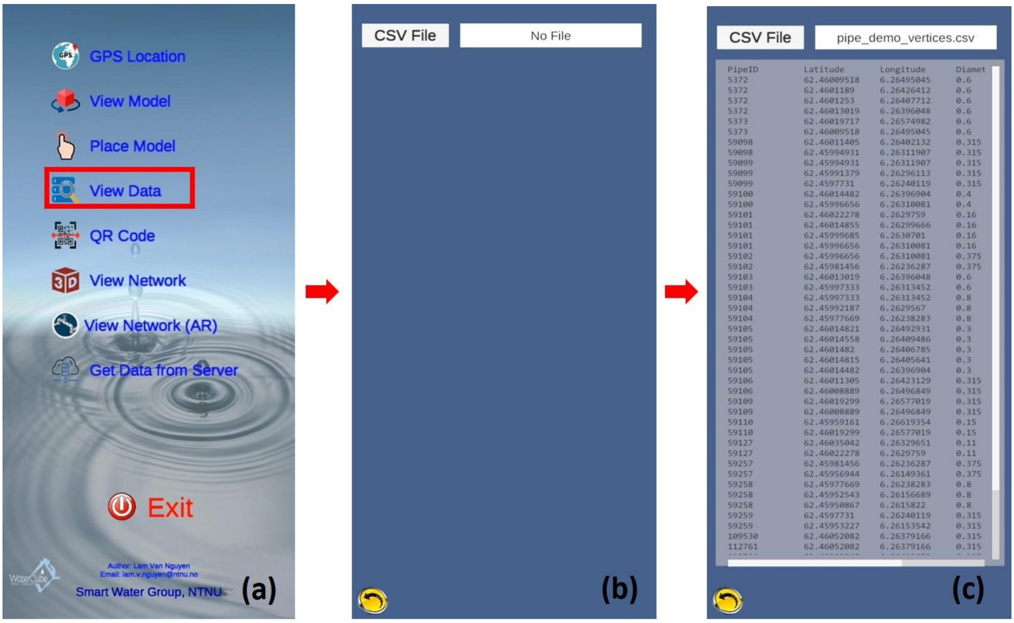


Figure B.11. Example of a 3D model in the "*WaterNet*" application

B.1.3.4. View Data

This function allows the user to directly view the CSV file using the "*WaterNet*" application (**Figure B.12**).



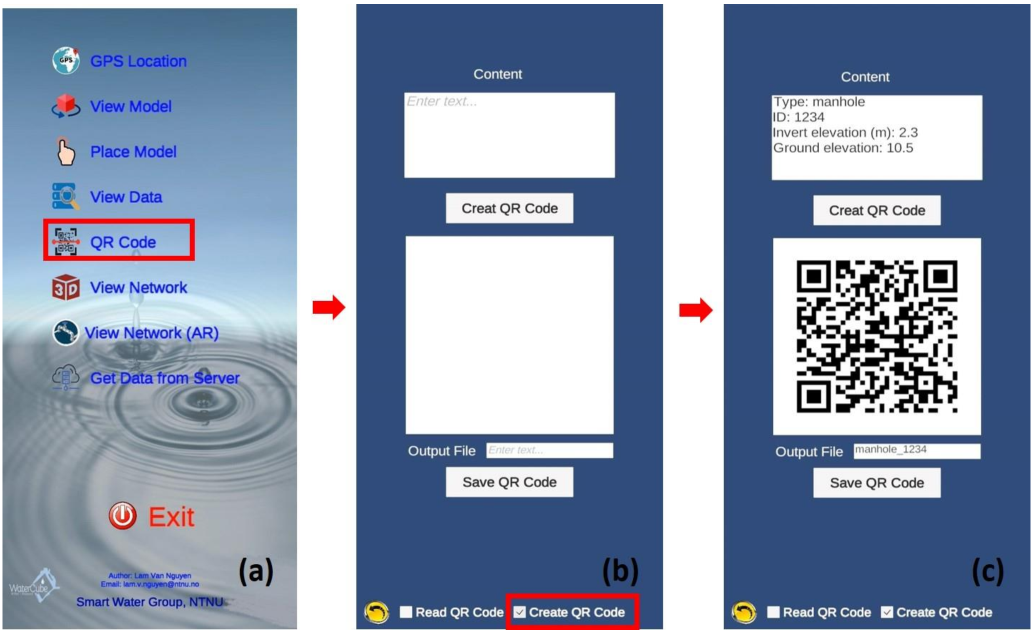
PipeID	Latitude	Longitude	Diamet
5372	62.46009518	6.26495045	0.6
5372	62.4601189	6.26426412	0.6
5372	62.4601253	6.26407712	0.6
5372	62.46013019	6.26396048	0.6
5373	62.46019717	6.26574982	0.6
5373	62.46009518	6.26495045	0.6
59098	62.46011405	6.26402132	0.315
59098	62.45994931	6.26311907	0.315
59099	62.45994931	6.26311907	0.315
59099	62.45991379	6.26296113	0.315
59099	62.4597731	6.26240119	0.315
59100	62.46014482	6.26396904	0.4
59100	62.45996656	6.26310081	0.4
59101	62.46022278	6.2629759	0.16
59101	62.46014855	6.26299666	0.16
59101	62.45999685	6.2630701	0.16
59101	62.45996656	6.26310081	0.16
59102	62.45996656	6.26310081	0.375
59102	62.45981456	6.26236287	0.375
59103	62.46013019	6.26396048	0.6
59103	62.45997333	6.26313452	0.6
59104	62.45997333	6.26313452	0.8
59104	62.45992187	6.2629567	0.8
59104	62.45977669	6.26238283	0.8
59105	62.46014821	6.26492931	0.3
59105	62.46014558	6.26409486	0.3
59105	62.4601482	6.26406785	0.3
59105	62.46014815	6.26405641	0.3
59105	62.46014482	6.26396904	0.3
59106	62.46011305	6.26423129	0.315
59106	62.46008889	6.26496849	0.315
59109	62.46019299	6.26577019	0.315
59109	62.46008889	6.26496849	0.315
59110	62.45959161	6.26619354	0.15
59110	62.46019299	6.26577019	0.15
59127	62.46035042	6.26329651	0.11
59127	62.46022278	6.2629759	0.11
59257	62.45981456	6.26236287	0.375
59257	62.45956944	6.26149361	0.375
59258	62.45977669	6.26238283	0.8
59258	62.45952543	6.26156689	0.8
59258	62.45950867	6.2615822	0.8
59259	62.4597731	6.26240119	0.315
59259	62.45953227	6.26153542	0.315
109530	62.46052082	6.26379166	0.315
112761	62.46052082	6.26379166	0.315

Figure B.12. Viewing CSV data in the "*WaterNet*" application

To use this function, the user first selects the symbol View Data on the menu screen of the application (**Figure B.12**a). Next, select the button CSV File on the next screen (**Figure B.12**b), and specify the location of the CSV file that the user wants to open. Finally, the entire CSV file is shown in a window at the center of the screen (**Figure B.12**c). To exit this function, the user clicks on the button ↶ in the bottom-left corner of the screen, and the user afterward is led to the main menu screen.

B.1.3.5. QR Code

This function allows the user to create and read QR codes. The attributes of sewer components can be coded and decoded using a QR code.

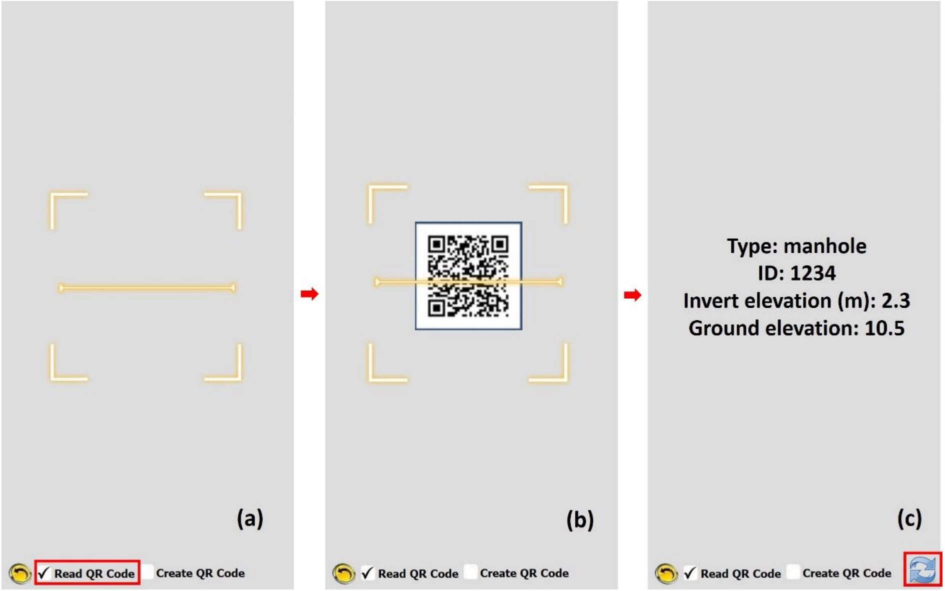


Type: manhole
ID: 1234
Invert elevation (m): 2.3
Ground elevation: 10.5

manhole_1234

Figure B.13. Creating a QR code in the "*WaterNet*" application

To create a QR code for assigning the attributes of the sewer component, the user first selects QR Code on the menu screen of the application (**Figure B.13**a). Next, select the check box "*Create QR Code*" in the bottom-right of the screen (**Figure B.13**b). Assign the wanted attributes for the object at the content box, select the button Creat QR Code to create a new QR code, and assign created attributes for the sewer component. A new QR code is generated at the center of the screen (**Figure B.13**c). To save the QR code, the user specifies the name of the QR code in the field "*Output File*", selects the button Save QR Code, then specifies the location to save the QR code on the device. Finally, this saved QR code can be transferred to a PC for storage and use for other purposes.



Type: manhole
ID: 1234
Invert elevation (m): 2.3
Ground elevation: 10.5

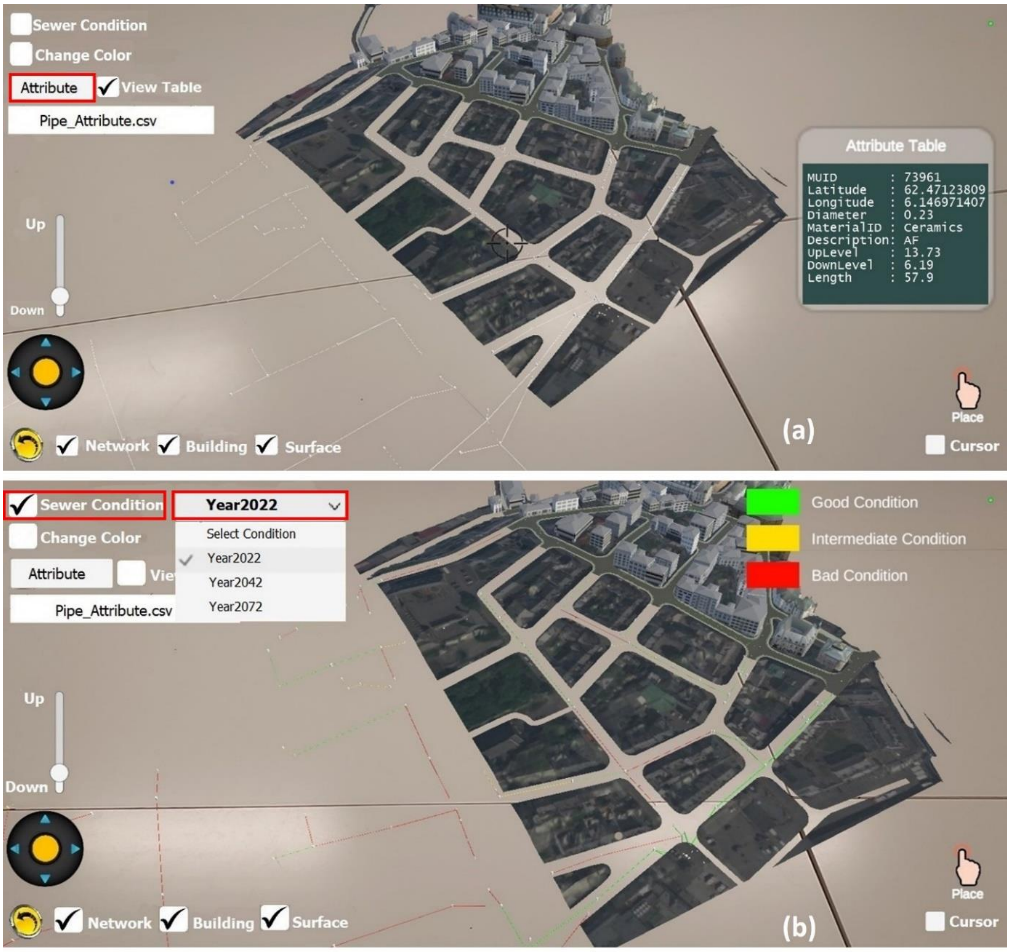
Figure B.14. Reading a QR code in the "*WaterNet*" application

To read a QR code, the user first selects the symbol QR Code on the menu screen of the application (**Figure B.14**a). Next, select the check box "*Read QR Code*" in the bottom-right of the screen (**Figure B.14**a). Move the screen to the QR code that the user wants to read, it makes sure that the yellow square on the center of the screen covers the QR code (**Figure B.14**b). The assigned attributes of the object will appear in the center of the screen (**Figure B.14**c). To delete current information on the screen and read a new one, the user selects the symbol 🔄 in the bottom-right corner of the screen (**Figure B.14**c). By clicking on the button ↶ in the bottom-left corner of the screen, the user is led to the main menu screen.

B.1.3.6. View Network

This function allows the user to import attributes of manholes or pipes from a CSV file and visualize them in real-time (**Figure B.15**a), the user also visualizes the sewer conditions using data from a CSV file (**Figure B.15**b). To activate this function, the user clicks on the symbol View Network on the main menu of the screen.

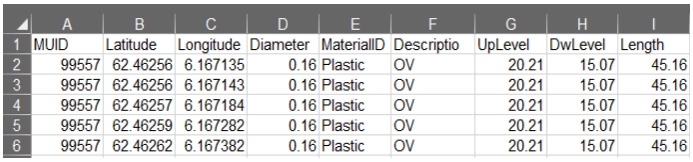
To view the attributes of the sewer components in real-time, the user first selects the button Attributes on the top-left corner of the screen (**Figure B.15**a). Next, specifies the location of the CSV file that contains the attributes of sewer components. An example of the attribute file is shown in **Figure B.16**. It is worth noting that the number of columns in this file is unlimited, the user can add as many columns as possible. Next, the user selects the check box ☑View Table, and an "*Attribute Table*" panel will appear on the right-hand side of the screen (**Figure B.15**a). By moving and rotating the device screen to the target marker ⌖ on the center of the screen hit the wanted object, the corresponding attributes of the object will appear in the "*Attribute Table*" panel (**Figure B.15**a).



Attribute Table

```
MUID        : 73961
Latitude    : 62.47123809
Longitude   : 6.146971407
Diameter    : 0.23
MaterialID  : Ceramics
Description: AF
UpLevel     : 13.73
DownLevel   : 6.19
Length      : 57.9
```

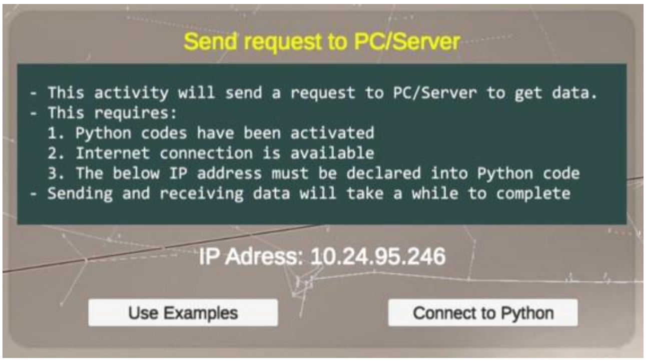
Figure B.15. Viewing the sewer network in the "*WaterNet*" application



	A	B	C	D	E	F	G	H	I
1	MUID	Latitude	Longitude	Diameter	MaterialID	Descriptio	UpLevel	DwLevel	Length
2	99557	62.46256	6.167135	0.16	Plastic	OV	20.21	15.07	45.16
3	99557	62.46256	6.167143	0.16	Plastic	OV	20.21	15.07	45.16
4	99557	62.46257	6.167184	0.16	Plastic	OV	20.21	15.07	45.16
5	99557	62.46259	6.167282	0.16	Plastic	OV	20.21	15.07	45.16
6	99557	62.46262	6.167382	0.16	Plastic	OV	20.21	15.07	45.16

Figure B.16. An example of the attribute file

To view the sewer conditions, the user first selects the check box ☑ Sewer Condition in the top-left corner of the screen (**Figure B.15**b).



```
- This activity will send a request to PC/Server to get data.
- This requires:
  1. Python codes have been activated
  2. Internet connection is available
  3. The below IP address must be declared into Python code
- Sending and receiving data will take a while to complete
```

IP Adress: 10.24.95.246

Figure B.17. Option for viewing sewer conditions

A window will appear on the center of the screen that requires the user to either use an example data or connect to a PC to transform the predictive conditions (**Figure B.17**).

- ❖ If the user selects the button Use Examples, the condition of sewer pipes in the year 2022 is showed up. This data was predicted using the machine learning model from one of the first author's works and it was set up as default values for example visualization.

- ❖ If the user selects the button Connect to Python, the application will send a request to a PC that is running a python code which is provided in **Table B.1**. After getting the request from the "*WaterNet*" application, this python code will run a function to get the conditions of sewer pipes from a pre-registered location on the PC and send the data back to the android devices for visualization. This approach will be efficient in the case the user wants to modify the sewer conditions file for visualization for different scenarios. To transfer data from PC to android devices by using this approach, the user needs to do below steps:

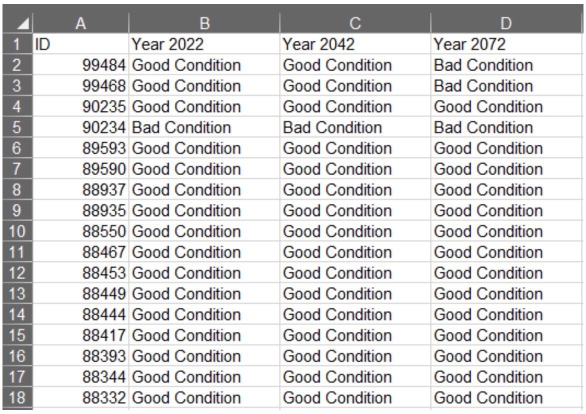
  - ➢ *Step 1:* Ensure the PC for running the python codes and android devices are connected to the same internet network.

  - ➢ *Step 2:* Run the python codes provide in **Table B.1** by using any Integrated Development Environment used for programming in Python (such as PyCharm, Spyder, etc.,). In this tutorial, we use Spyder to run python codes.

  - ➢ *Step 3:* Change the host address `host, port = "10.24.95.9", 65432` in the python code by the IP address created by the "*WaterNet*" application (**Figure A.17**). Please notice that this IP address will be different depending on each specific android device.

- *Step 4:* It takes several seconds to read and transfer data from a PC to an android device depending on the capacity of the data and internet connection.
- *Step 5:* If the data is loaded successfully, the combo box in the top-left corner of the screen will contain the years of sewer conditions (**Figure B.15**b).

The data structure of the sewer conditions is partly shown in **Figure B.18**. The first column contains the name/ID of sewer pipes in this data format. From the second column, the condition of the sewer pipes each year is stored in each column. From the second row, each sewer pipe's name and corresponding state are stored in each row.

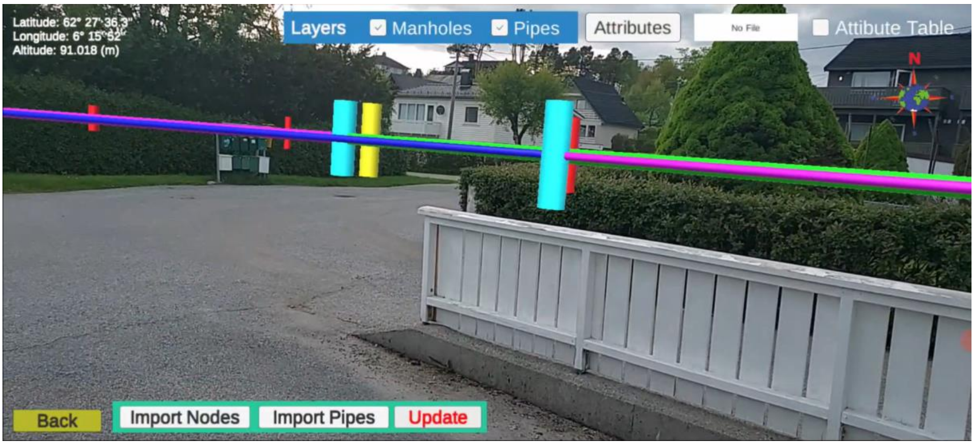


	A	B	C	D
1	ID	Year 2022	Year 2042	Year 2072
2	99484	Good Condition	Good Condition	Bad Condition
3	99468	Good Condition	Good Condition	Bad Condition
4	90235	Good Condition	Good Condition	Good Condition
5	90234	Bad Condition	Bad Condition	Bad Condition
6	89593	Good Condition	Good Condition	Good Condition
7	89590	Good Condition	Good Condition	Good Condition
8	88937	Good Condition	Good Condition	Good Condition
9	88935	Good Condition	Good Condition	Good Condition
10	88550	Good Condition	Good Condition	Good Condition
11	88467	Good Condition	Good Condition	Good Condition
12	88453	Good Condition	Good Condition	Good Condition
13	88449	Good Condition	Good Condition	Good Condition
14	88444	Good Condition	Good Condition	Good Condition
15	88417	Good Condition	Good Condition	Good Condition
16	88393	Good Condition	Good Condition	Good Condition
17	88344	Good Condition	Good Condition	Good Condition
18	88332	Good Condition	Good Condition	Good Condition

Figure B.18. An example of the sewer conditions

B.1.3.7. View Network (AR)

This function allows the user to view the sewer network using the Augmented Reality technique. Because of the limitation of the hardware of Android devices, this function is only applied to a small area of the study area. This function will define the relative relationship between the user's location and objects of the sewer network using a GPS signal. Therefore, this function does not work on HoloLens devices that do not have GPS receivers. Additionally, the accuracy of the objects detected using this function mainly depends on the accuracy of the GPS receiver on the android device. To activate this function, the user clicks on the symbol View Network (AR) on the main menu of the screen. The interface of this function is shown in **Figure B.19**.

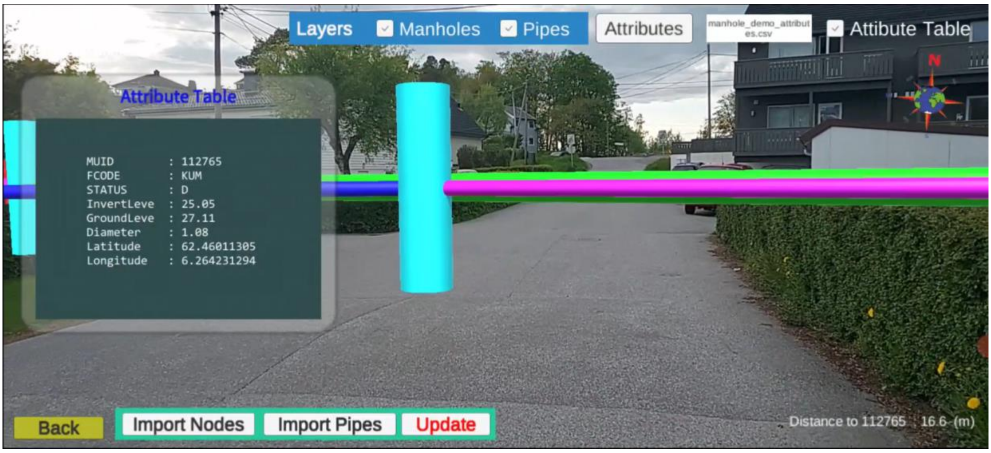


Latitude: 62° 27' 36.3"
Longitude: 6° 15' 52"
Altitude: 91.018 (m)

Figure B.19. The interface of the "*View Network (AR)*" function

The below steps describe how to use this function:

- *Step 1:* Ensure that the location function on the android devices is enabled before activating this function.
- *Step 2:* Click on the button Import Node and/or button Import Pipe to import manholes and/or pipes from the CSV file. The structures of the CSV file containing the manholes and pipes are shown in **Figure A.2**a and **Figure A.2**b, respectively.
- *Step 3:* The symbol N shows the different angles in degrees relative to the geographic North pole and the android device's compass. It is recommended to wait for the geographical locations at the top-left corner of the screen to be shown up and this angle is approximately zero before clicking on the button Update to put the sewer network on the screen.
- *Step 4:* To turn display/hide the manhole layer or pipe layer on the screen, the user can check/uncheck the corresponding check boxes ☑ Manholes or ☑ Pipes, respectively.
- *Step 5:* To view the attributes of manholes or pipes, the user must import the CSV file that contains the attributes of manholes or pipes by clicking on the button Attributes and checking on the check box ☑ Attribute Table at the top-right corner of the screen. By moving the screen to the point at the center of the screen hit the objects, the attributes of corresponding objects will appear on the small panel at the left-hand side of the screen (**Figure B.20**).



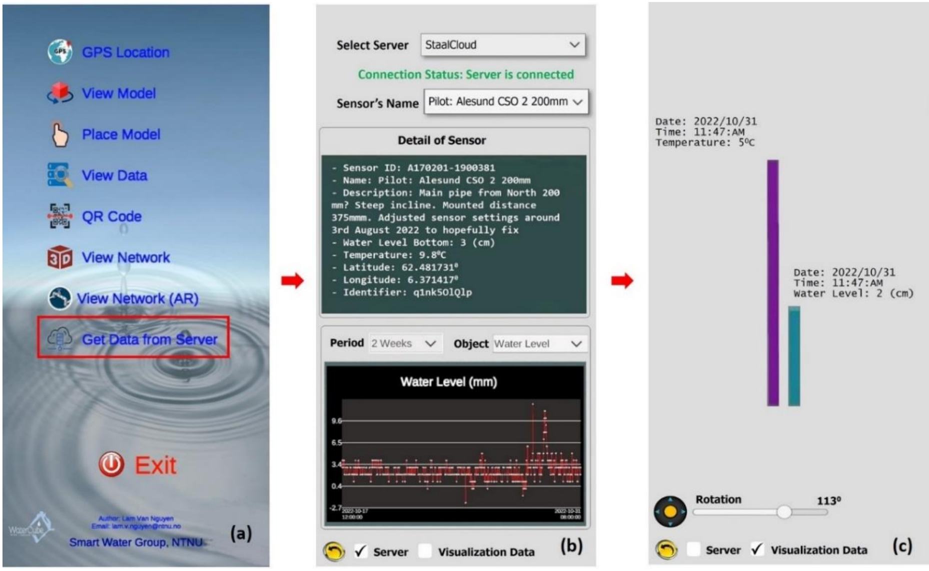
Attribute Table

```
MUID        : 112765
FCODE       : KUM
STATUS      : D
InvertLeve  : 25.05
GroundLeve  : 27.11
Diameter    : 1.08
Latitude    : 62.46011305
Longitude   : 6.264231294
```

Figure B.20. Viewing the object's attributes in the "*View Network (AR)*" function

B.1.3.8. Get Data from Server

This function allows the user to access and visualize data from the provided server in real-time. In general, accessing the server to get data requires authorization (for example, username and password), we only, therefore, illustrate how to access and get data from the "*StaalCloud*" server controlled by Ålesund city in this example.



Select Server: StaalCloud

Connection Status: Server is connected

Sensor's Name: Pilot: Alesund CSO 2 200mm

Detail of Sensor

```
- Sensor ID: A170201-1900381
- Name: Pilot: Alesund CSO 2 200mm
- Description: Main pipe from North 200
mm? Steep incline. Mounted distance
375mmm. Adjusted sensor settings around
3rd August 2022 to hopefully fix
- Water Level Bottom: 3 (cm)
- Temperature: 9.8°C
- Latitude: 62.481731°
- Longitude: 6.371417°
- Identifier: q1nk5O1Q1p
```

Period: 2 Weeks

Object: Water Level

```
Date: 2022/10/31
Time: 11:47:AM
Temperature: 5ºC
```

```
Date: 2022/10/31
Time: 11:47:AM
Water Level: 2 (cm)
```

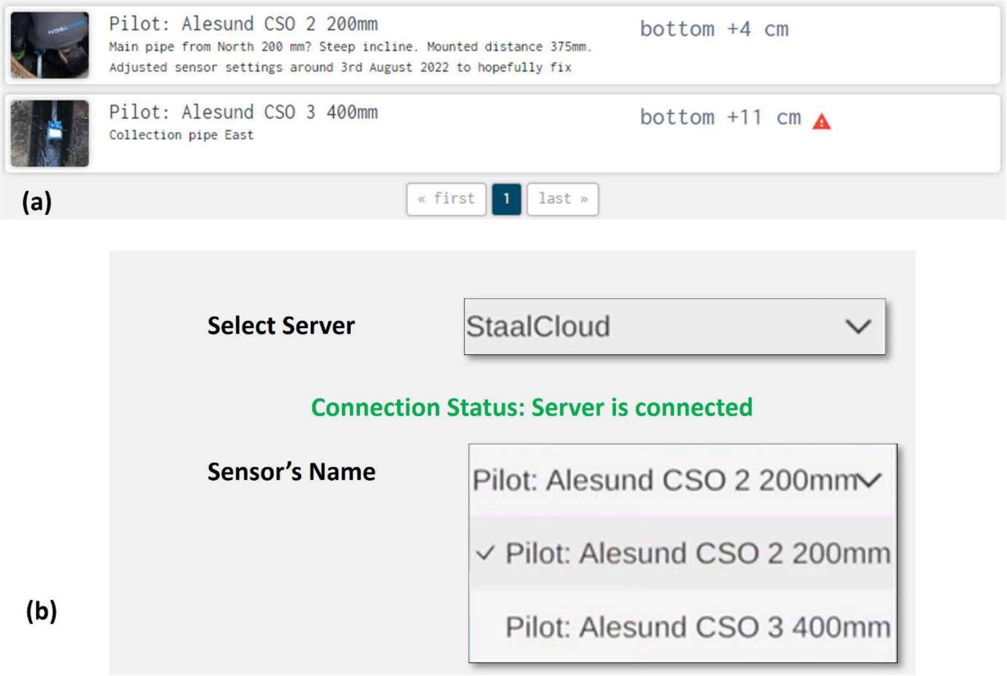
Rotation: 113º

Figure B.21. The interface of the "*Get Data from Server*" function

The below steps describe how to use this function:

- *Step 1:* Ensure that the internet connection is connected to the android devices. Select the symbol Get Data from Server from the main menu of the application (**Figure B.21**a).

- *Step 2:* Select the combo box "*Select Server*" to connect to the server. If the application connects to the server successfully, the red text "Connection Status: Not connected" will change to the green text "Connection Status: Server is connected" (**Figure B.21**b).
- *Step 3:* The combo box "*Sensor's Name*" lists all sensors on the server. A comparison of the number of sensors on the server and the number of sensors obtained by this function is shown in **Figure B.22**.



Pilot: Alesund CSO 2 200mm
Main pipe from North 200 mm? Steep incline. Mounted distance 375mm.
Adjusted sensor settings around 3rd August 2022 to hopefully fix
bottom +4 cm

Pilot: Alesund CSO 3 400mm
Collection pipe East
bottom +11 cm

Select Server: StaalCloud

Connection Status: Server is connected

Sensor's Name: Pilot: Alesund CSO 2 200mm
✓ Pilot: Alesund CSO 2 200mm
Pilot: Alesund CSO 3 400mm

Figure B.22. Sensors on the server **(a)** and the result obtained by the "*Get Data from Server*" function **(b)**

- *Step 4a:* To view data in real-time, the user selects the checkbox ✓ Server at the bottom of the screen (**Figure B.21**b). This function allows the user to view the water level and water temperature received from sensors by changing the values of the combo box "*Object*" Object Water Level in **Figure B.21**b. Additionally, this function allows the user to view these aforementioned values in 1 day, 2 days, 1 week, 2 weeks, and 1 month by changing the values of the combo box "*Period*" Period 2 Weeks in **Figure B.21**b. **Figure B.23** illustrates an example of real-time data received from the server using the

"*Get Data from Server*" function.

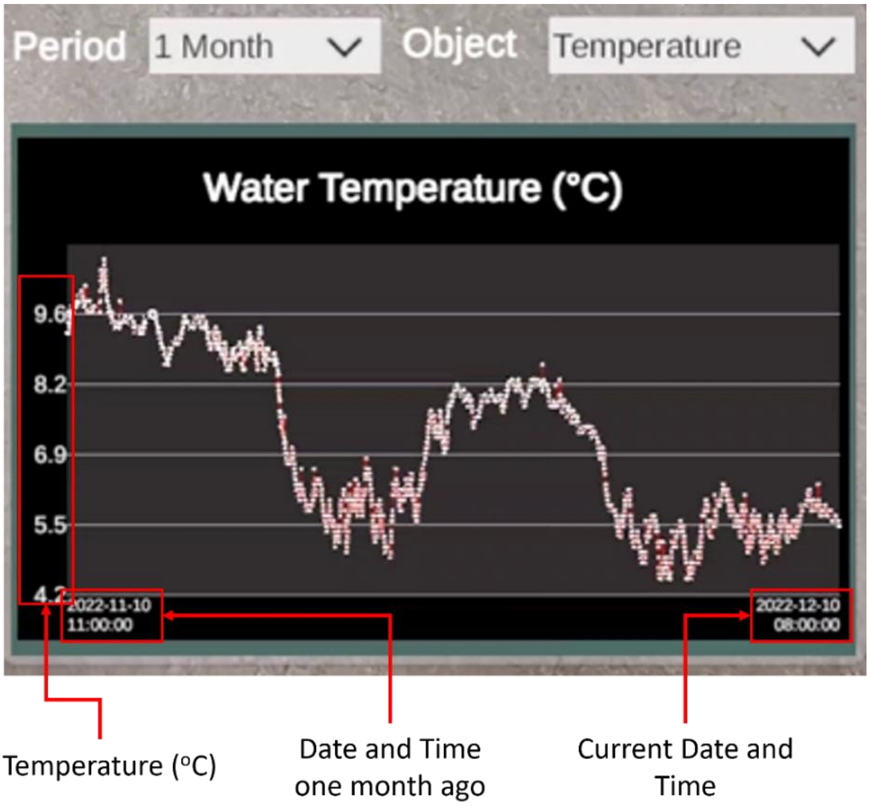


Figure B.23. Real-time data visualization obtained from the server

- *Step 4b:* To visualize data in 3D, the user selects the checkbox ✓ **Visualization Data** at the bottom of the screen. The water level and water temperature will be visualized simultaneously (**Figure B.21**c).

B.2. HoloLens Application

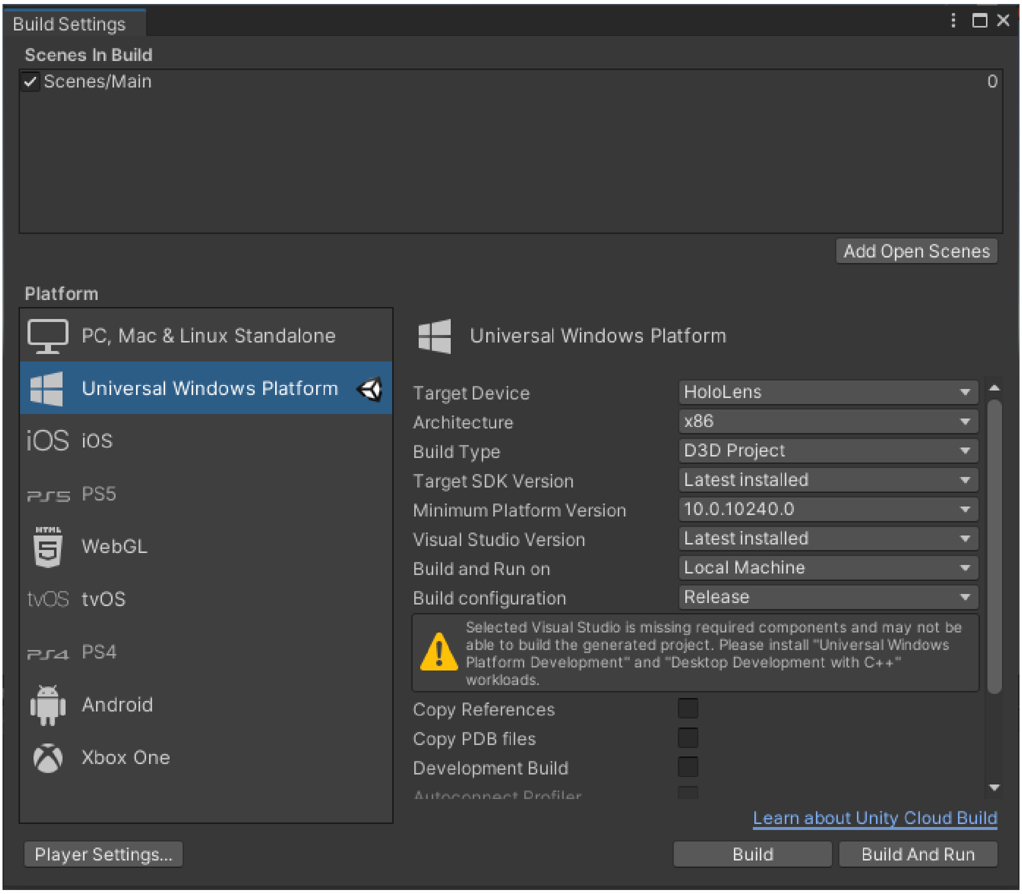
The "*WaterNet*" application used on the HoloLens device is distributed by a project that can be implemented via Unity. To download the Unity project, the user accesses the link provided in **Table B.2**.

Table B.2. Downloaded link of the "WaterNet" application on the HoloLens device

Name	Download link
WaterNet_HoloLens	https://www.mediafire.com/file/lgqg6k2kxdi54bk/WaterNet_HoloLens.zip/file

The steps for implementing and running this application are presented as follows:

- *Step 1:* Download and extract the file from the link that is provided in **Table B.2**.
- *Step 2:* Download and install the Unity software (https://unity.com/download). The Unity version 2020.3.32f1 is recommended for the implementation of this application.
- *Step 3:* Open Unity and import this project. Make sure that Unity software is optimized for HoloLens devices by selecting *File/Build Setting…* In the *Platform* section, select "*Universal Windows Platform*" and *HoloLens* is selected in the section *Target Device* (**Figure B.24**).



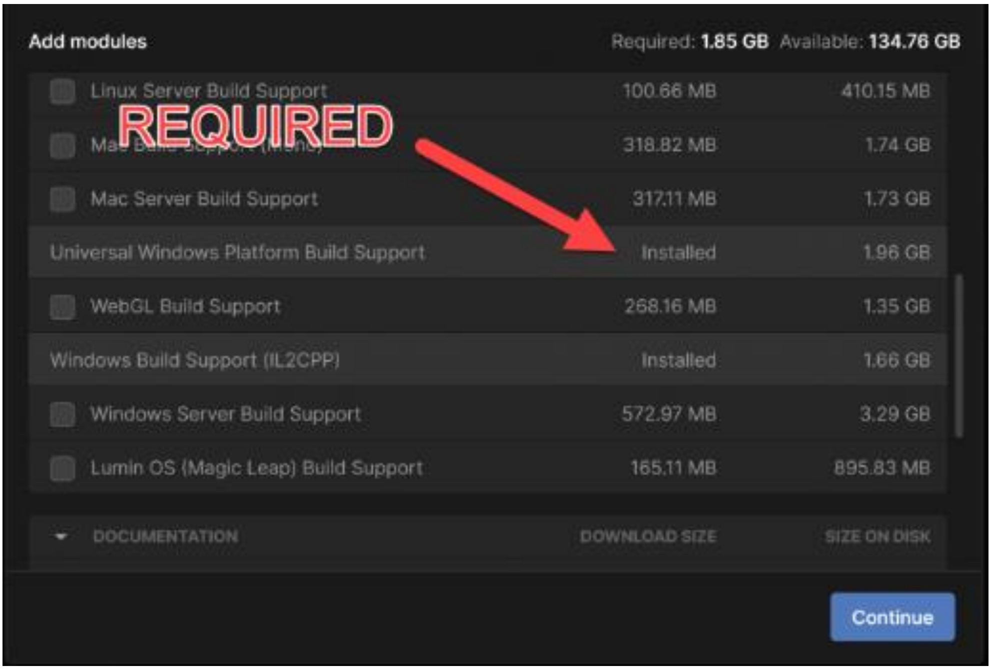
Target Device: HoloLens
Architecture: x86
Build Type: D3D Project
Target SDK Version: Latest installed
Minimum Platform Version: 10.0.10240.0
Visual Studio Version: Latest installed
Build and Run on: Local Machine
Build configuration: Release

Selected Visual Studio is missing required components and may not be able to build the generated project. Please install "Universal Windows Platform Development" and "Desktop Development with C++" workloads.

Copy References □
Copy PDB files □
Development Build □
Autoconnect Profiler □

Figure B.24. Unity configuration for HoloLens device

If the "*Universal Windows Platform*" option is not installed, the user can install this package via Unity Hub (**Figure B.25**).



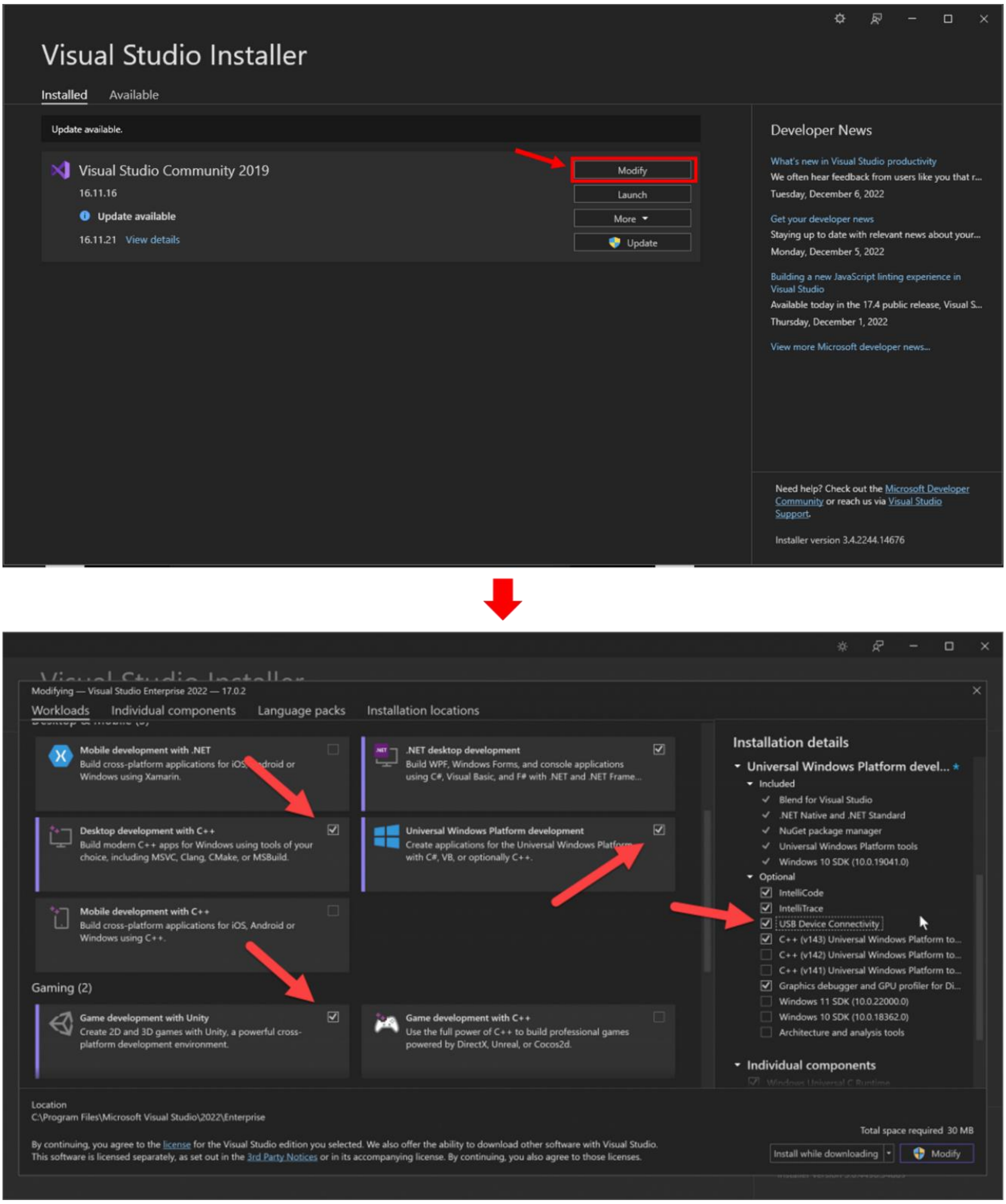
Add modules — Required: **1.85 GB** Available: **134.76 GB**

□	Linux Server Build Support	100.66 MB	410.15 MB
□	Mac Build Support (Mono)	318.82 MB	1.74 GB
□	Mac Server Build Support	317.11 MB	1.73 GB
	Universal Windows Platform Build Support	Installed	1.96 GB
□	WebGL Build Support	268.16 MB	1.35 GB
	Windows Build Support (IL2CPP)	Installed	1.66 GB
□	Windows Server Build Support	572.97 MB	3.29 GB
□	Lumin OS (Magic Leap) Build Support	165.11 MB	895.83 MB
▾	DOCUMENTATION	DOWNLOAD SIZE	SIZE ON DISK

REQUIRED

Figure B.25. Installing the "*Universal Windows Platform*" package

If the user sees the notice reminding to install "*Universal Windows Platform Development*" and "*Desktop Development with C++*" workloads as in **Figure B.24**, please open "*Visual Studio Installer*" from the ***Start*** menu and install them as in **Figure B.26**.



Visual Studio Installer

Installed Available

Update available.

Visual Studio Community 2019
16.11.16
Update available
16.11.21 View details

Developer News

What's new in Visual Studio productivity
We often hear feedback from users like you that r...
Tuesday, December 6, 2022

Get your developer news
Staying up to date with relevant news about your...
Monday, December 5, 2022

Building a new JavaScript linting experience in Visual Studio
Available today in the 17.4 public release, Visual S...
Thursday, December 1, 2022

View more Microsoft developer news...

Need help? Check out the Microsoft Developer Community or reach us via Visual Studio Support.

Installer version 3.4.2244.14676

Modifying — Visual Studio Enterprise 2022 — 17.0.2

Workloads Individual components Language packs Installation locations

- □ Mobile development with .NET — Build cross-platform applications for iOS, Android or Windows using Xamarin.
- ☑ .NET desktop development — Build WPF, Windows Forms, and console applications using C#, Visual Basic, and F# with .NET and .NET Frame...
- ☑ Desktop development with C++ — Build modern C++ apps for Windows using tools of your choice, including MSVC, Clang, CMake, or MSBuild.
- ☑ Universal Windows Platform development — Create applications for the Universal Windows Platform with C#, VB, or optionally C++.
- □ Mobile development with C++ — Build cross-platform applications for iOS, Android or Windows using C++.

Gaming (2)

- ☑ Game development with Unity — Create 2D and 3D games with Unity, a powerful cross-platform development environment.
- □ Game development with C++ — Use the full power of C++ to build professional games powered by DirectX, Unreal, or Cocos2d.

Installation details

▾ Universal Windows Platform devel... *
▾ Included
✓ Blend for Visual Studio
✓ .NET Native and .NET Standard
✓ NuGet package manager
✓ Universal Windows Platform tools
✓ Windows 10 SDK (10.0.19041.0)
▾ Optional
☑ IntelliCode
☑ IntelliTrace
☑ USB Device Connectivity
☑ C++ (v143) Universal Windows Platform to...
□ C++ (v142) Universal Windows Platform to...
□ C++ (v141) Universal Windows Platform to...
☑ Graphics debugger and GPU profiler for Di...
□ Windows 11 SDK (10.0.22000.0)
□ Windows 10 SDK (10.0.18362.0)
□ Architecture and analysis tools
▾ Individual components

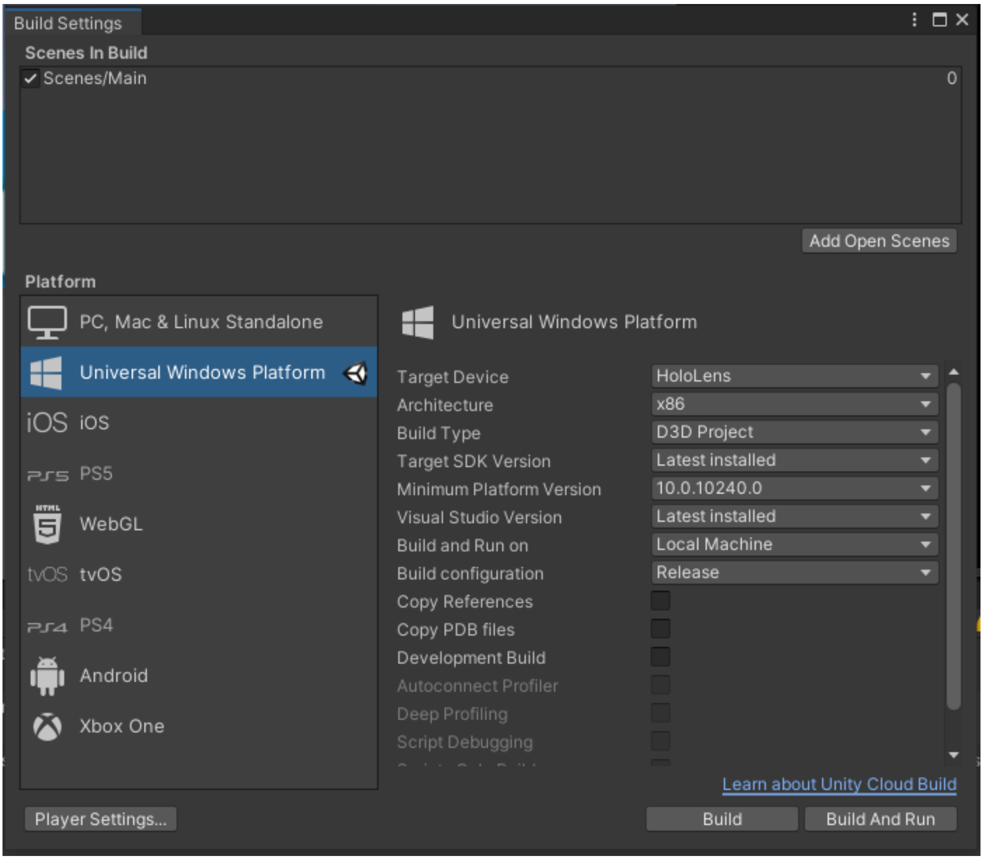
Location
C:\Program Files\Microsoft Visual Studio\2022\Enterprise

By continuing, you agree to the license for the Visual Studio edition you selected. We also offer the ability to download other software with Visual Studio. This software is licensed separately, as set out in the 3rd Party Notices or in its accompanying license. By continuing, you also agree to those licenses.

Total space required 30 MB

Figure B.26. Installing the "*Universal Windows Platform Development*" and "*Desktop Development with C++*" workloads

After installing the packages and workloads successfully, the configuration window for HoloLens looks as in **Figure B.27**. Before compiling this application on a HoloLens device, the user must add the "*Main*" scene in the section "*Scenes In Build*" (**Figure B.27**) and activate the object "*SceneDescriptionPanelRev*" as in **Figure B.28**.



Build Settings

Scenes In Build

☑ Scenes/Main 0

Add Open Scenes

Platform

- PC, Mac & Linux Standalone
- Universal Windows Platform
- iOS
- PS5
- WebGL
- tvOS
- PS4
- Android
- Xbox One

Universal Windows Platform

Target Device	HoloLens
Architecture	x86
Build Type	D3D Project
Target SDK Version	Latest installed
Minimum Platform Version	10.0.10240.0
Visual Studio Version	Latest installed
Build and Run on	Local Machine
Build configuration	Release
Copy References	□
Copy PDB files	□
Development Build	□
Autoconnect Profiler	□
Deep Profiling	□
Script Debugging	□

Learn about Unity Cloud Build

Player Settings... Build Build And Run

Figure B.27. Successful configuration for HoloLens device in Unity

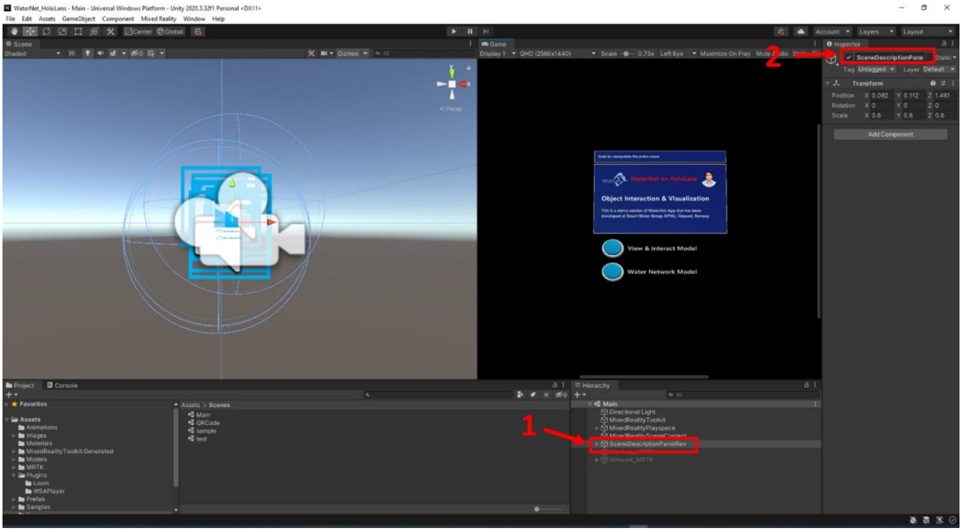


Figure B.28. The "*WaterNet*" project in Unity

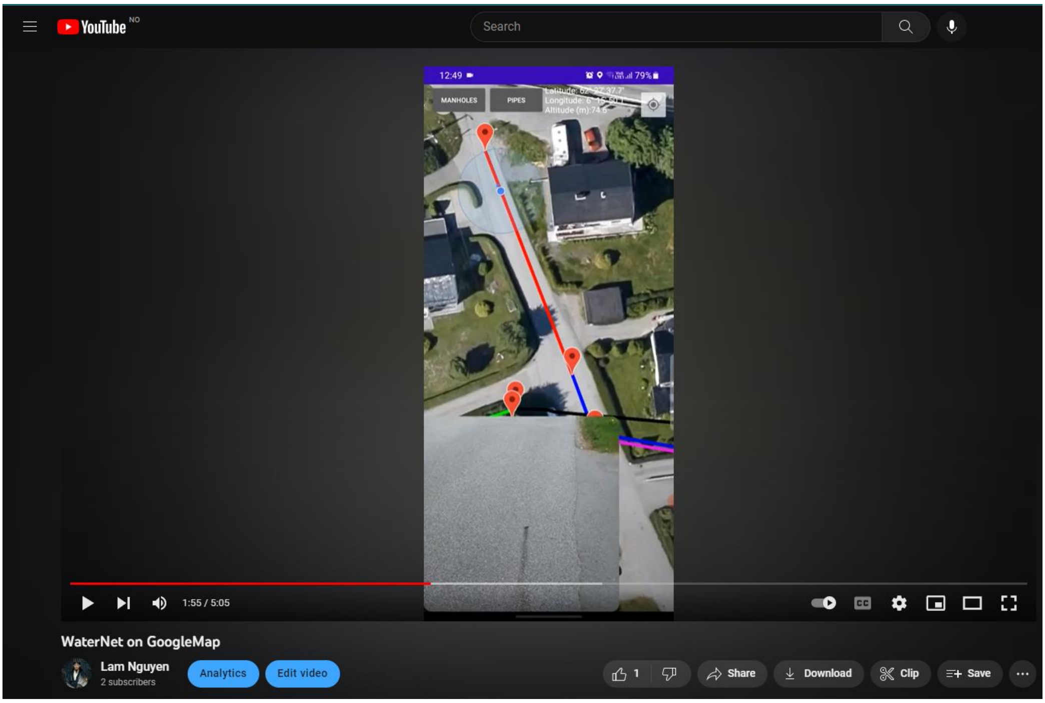
➢ *Step 4*: Connect the PC with the HoloLens device. Select *File/Build Setting…/Build and Run* to compile the application on the HoloLens device.

B.3. Video examples

This section provides video examples of using the above functions.

B.3.1. Video example of the "WaterNet on GoogleMap" application

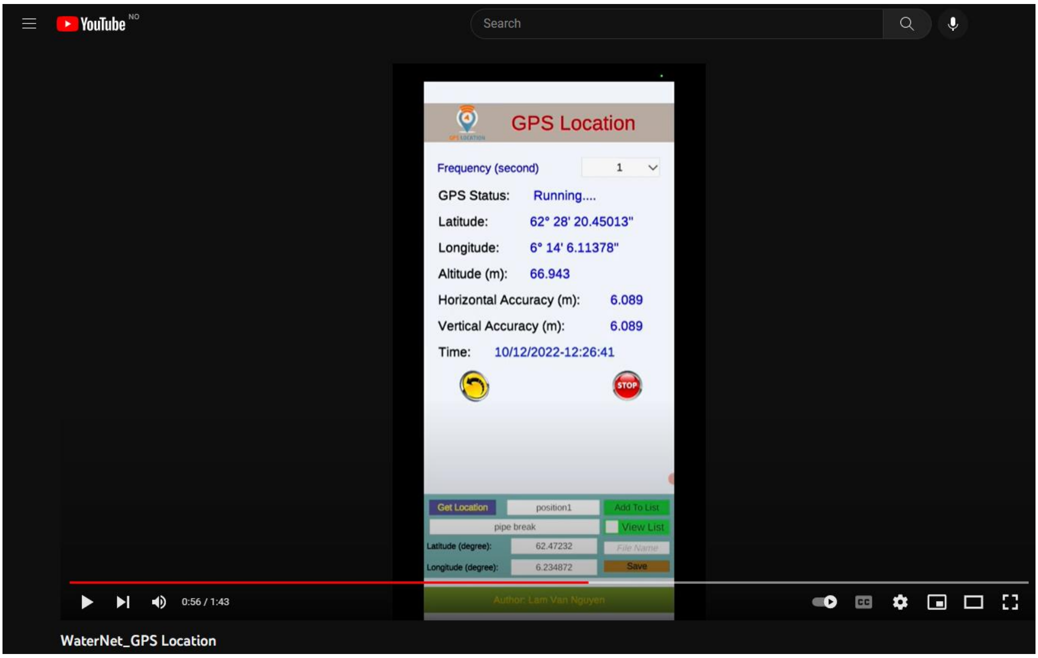
Video link: https://www.youtube.com/watch?v=3jfjDVP6Cag



B.3.2. Video examples of the "WaterNet" Application on an Android device

B.3.2.1. GPS Location

Video link: https://www.youtube.com/watch?v=u_gcxmjHVmk



Frequency (second) 1

GPS Status: Running....

Latitude: 62° 28' 20.45013"

Longitude: 6° 14' 6.11378"

Altitude (m): 66.943

Horizontal Accuracy (m): 6.089

Vertical Accuracy (m): 6.089

Time: 10/12/2022-12:26:41

position1

pipe break

Latitude (degree): 62.47232

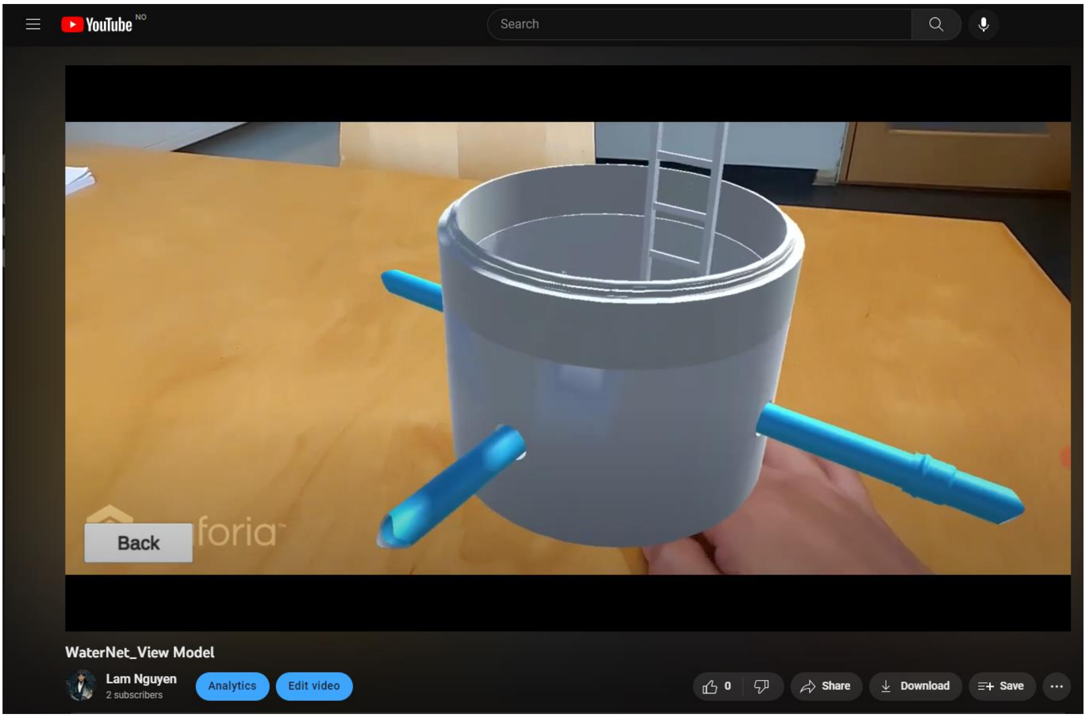
Longitude (degree): 6.234872

Author: Lam Van Nguyen

WaterNet_GPS Location

B.3.2.2. View Model

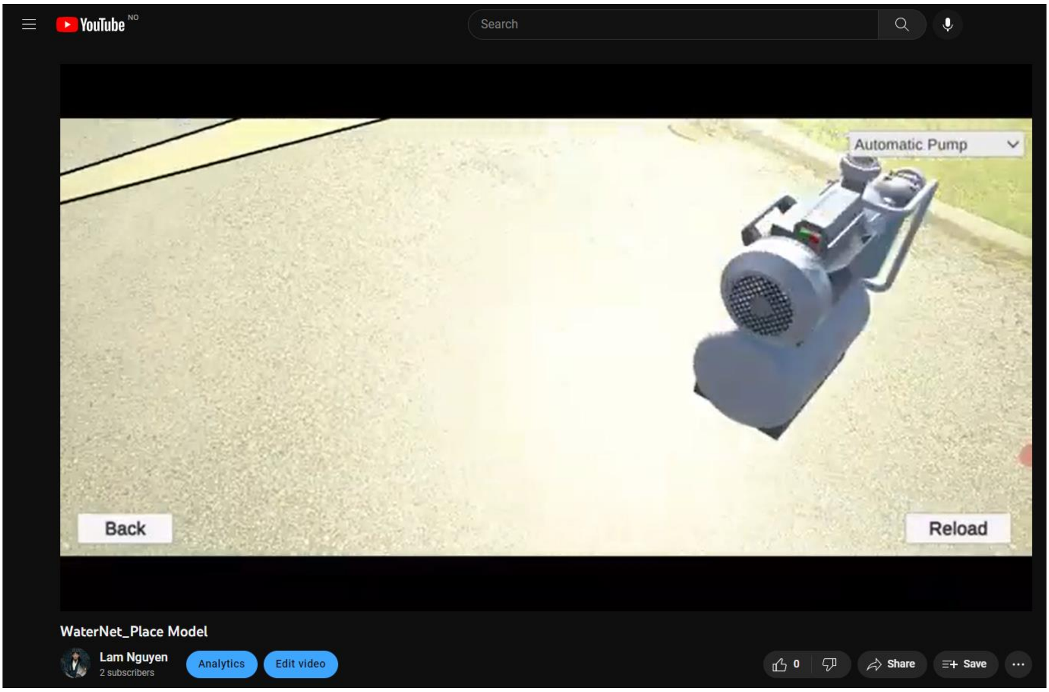
Video link: https://www.youtube.com/watch?v=pBOkH39HMbw



WaterNet_View Model

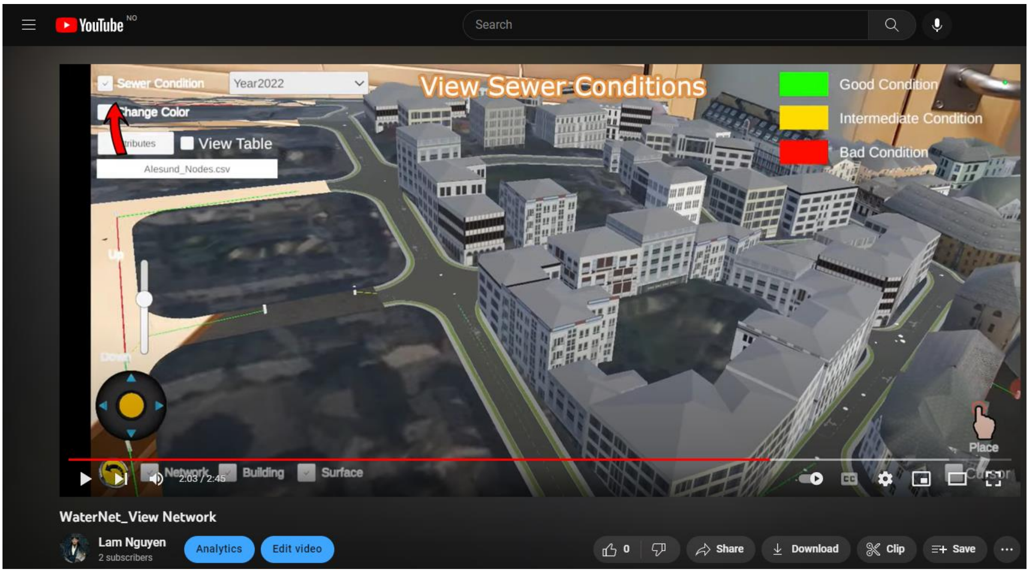
B.3.2.3. Place Model

Video link: https://www.youtube.com/watch?v=WpE02c0Lo5U



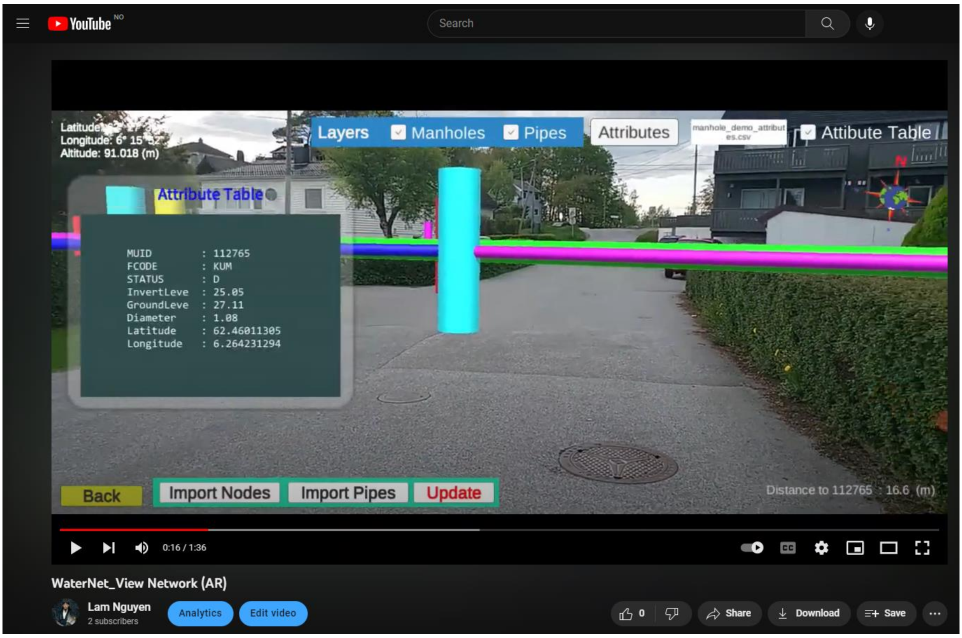
B.3.2.4. View Network

Video link: https://www.youtube.com/watch?v=sK93RZY0Lsg



B.3.2.5. View Network (AR)

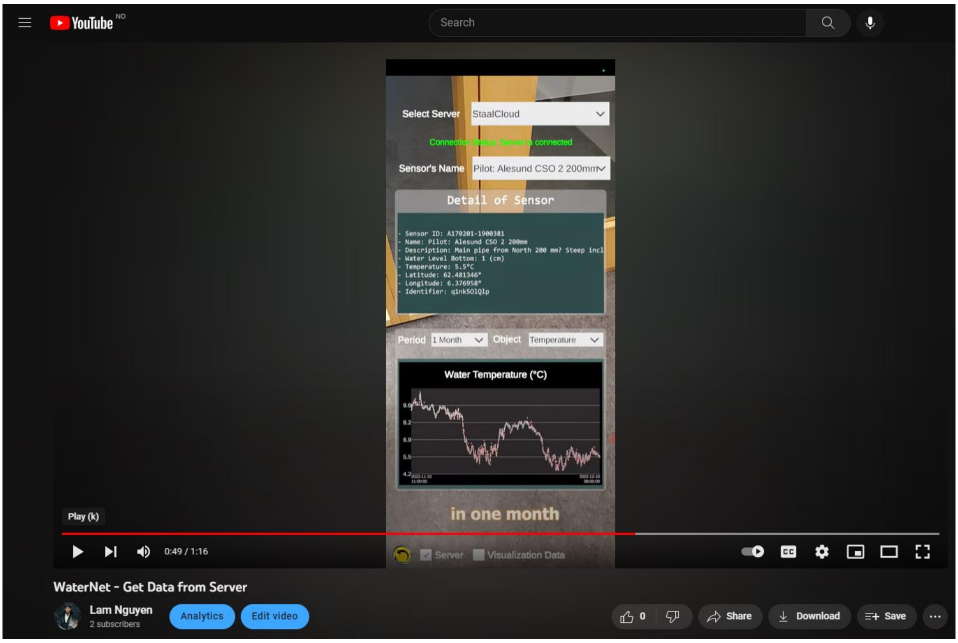
Video link: https://www.youtube.com/watch?v=aOKMS2HQd8A



```
MUID         : 112765
FCODE        : KUM
STATUS       : D
InvertLeve   : 25.05
GroundLeve   : 27.11
Diameter     : 1.08
Latitude     : 62.46011305
Longitude    : 6.264231294
```

B.3.2.6. Get Data from Server

Video link: https://www.youtube.com/watch?v=WZU25vCQ4FA



Select Server: StaalCloud

Connection Status: Server is connected

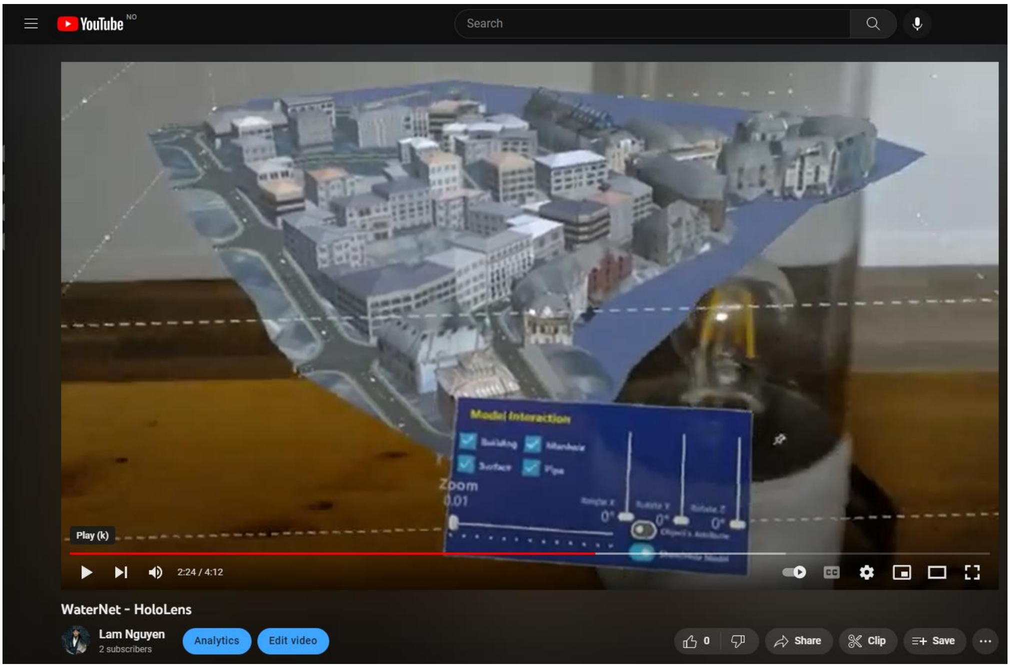
Sensor's Name: Pilot: Alesund CSO 2 200mm

```
- Sensor ID: A170201-1900381
- Name: Pilot: Alesund CSO 2 200mm
- Description: Main pipe from North 200 mm? Steep incl
- Water Level Bottom: 1 (cm)
- Temperature: 5.5°C
- Latitude: 62.481346°
- Longitude: 6.376958°
- Identifier: q1nk5O1Q1p
```

Period: 1 Month   Object: Temperature

B.3.3. Video example of the "WaterNet" application on HoloLens

Video link: https://www.youtube.com/watch?v=bjv-whMe4RA



ISBN 978-82-326-8306-2 (printed ver.)
ISBN 978-82-326-8305-5 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU
Norwegian University of
Science and Technology