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,

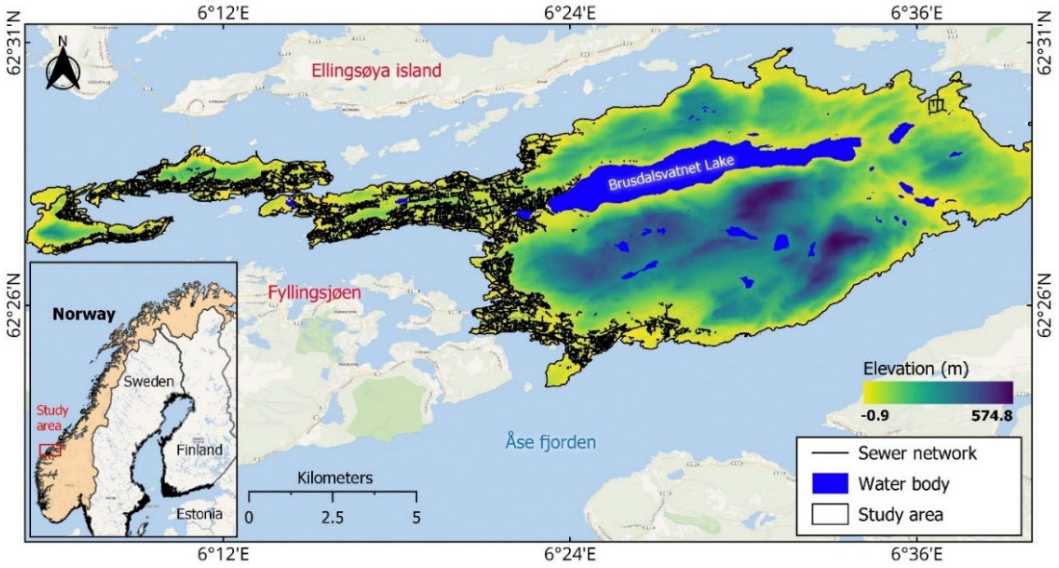

FIGURE 1. The sewer network in the study area.

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 $\times$ 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)

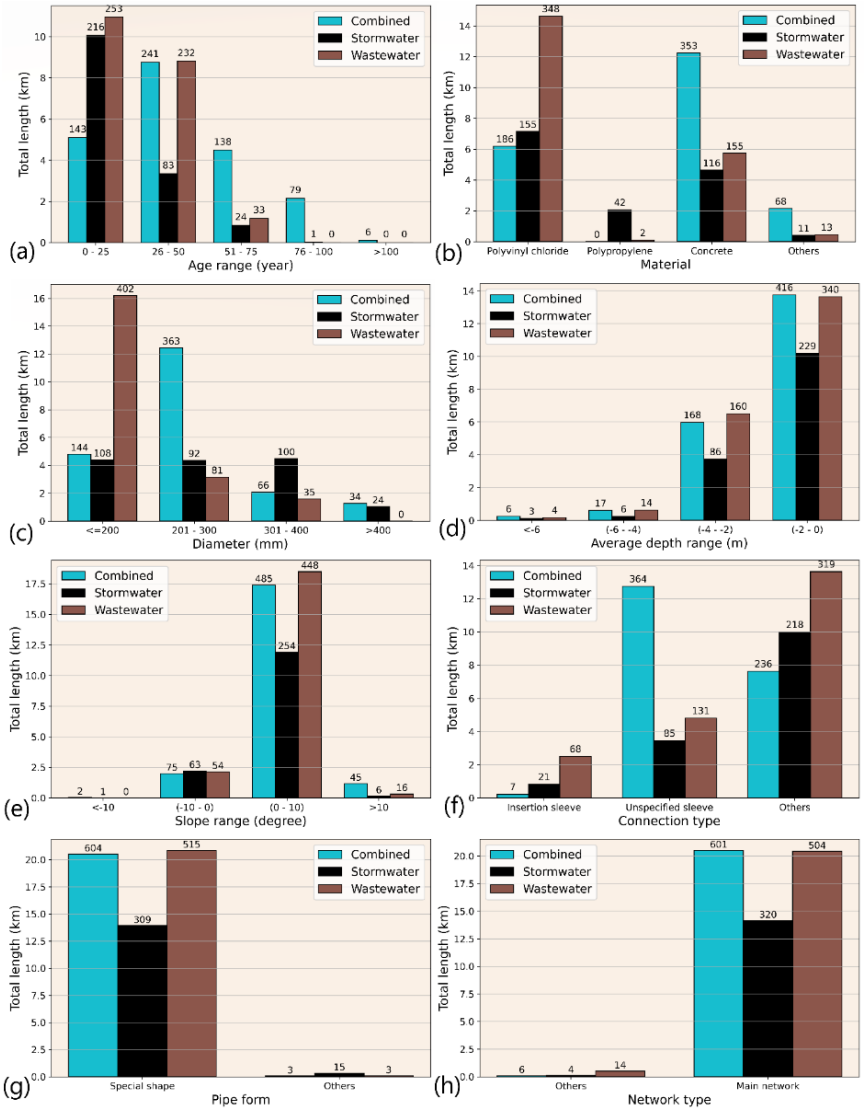

FIGURE 2. The physical characteristics of the sewage network.

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 $\times$ 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

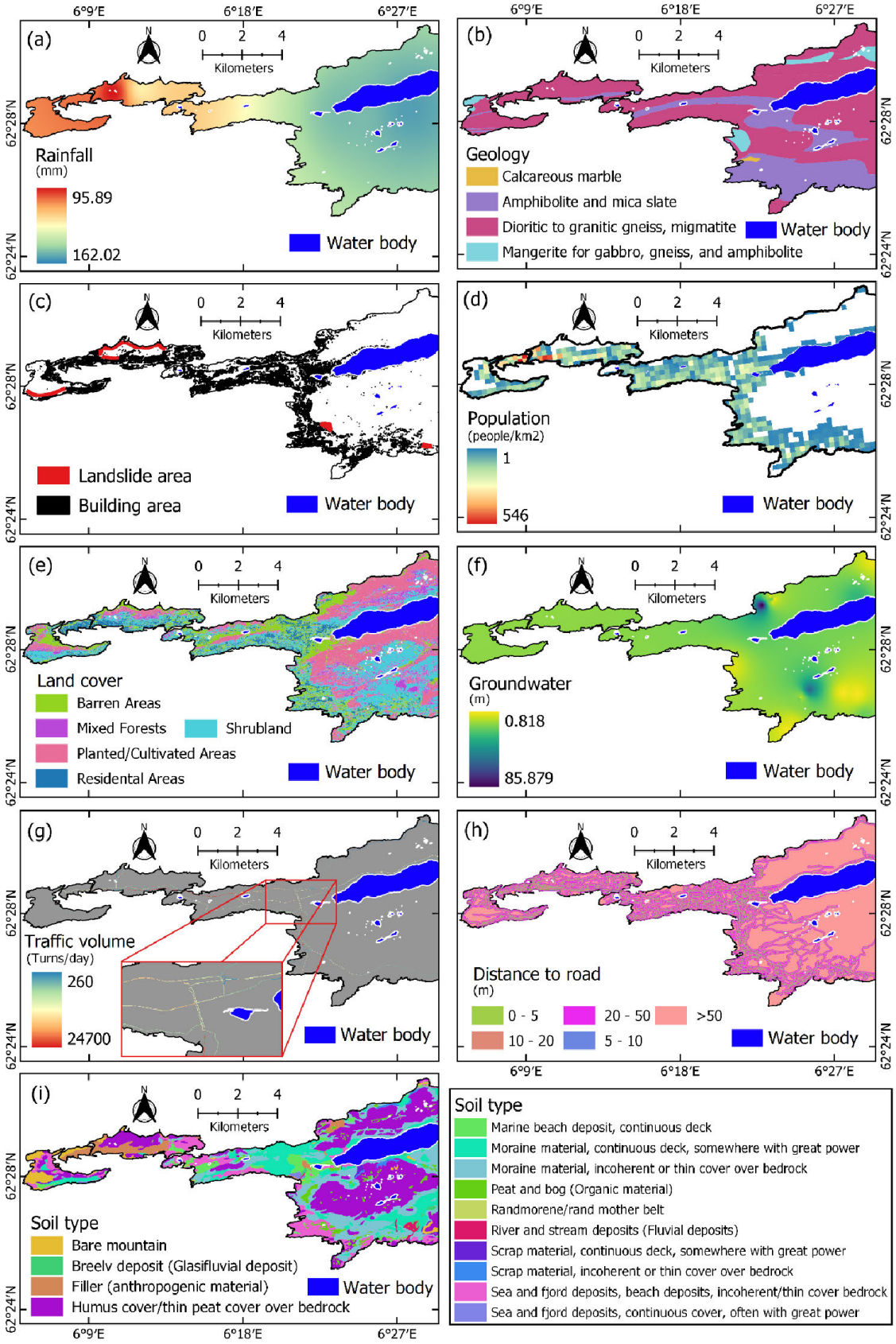

FIGURE 3. The environmental factors used in this study.

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]:

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

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]:

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

where auxiliary variables α_i are Lagrange multipliers, C is the regularization parameter, and $K(x, x_i)$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

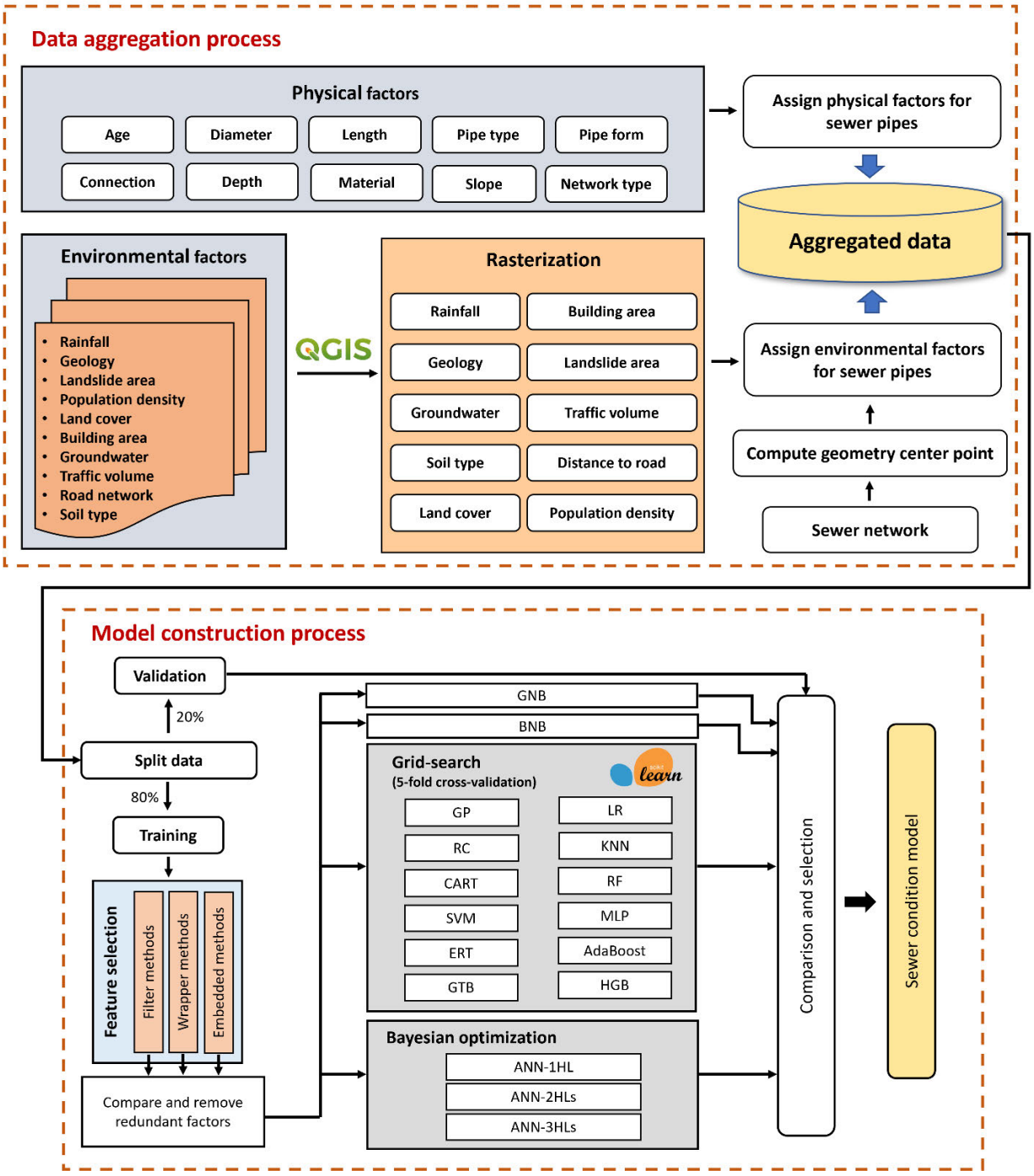

FIGURE 4. The framework for modeling the condition of a sewerage system.

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].

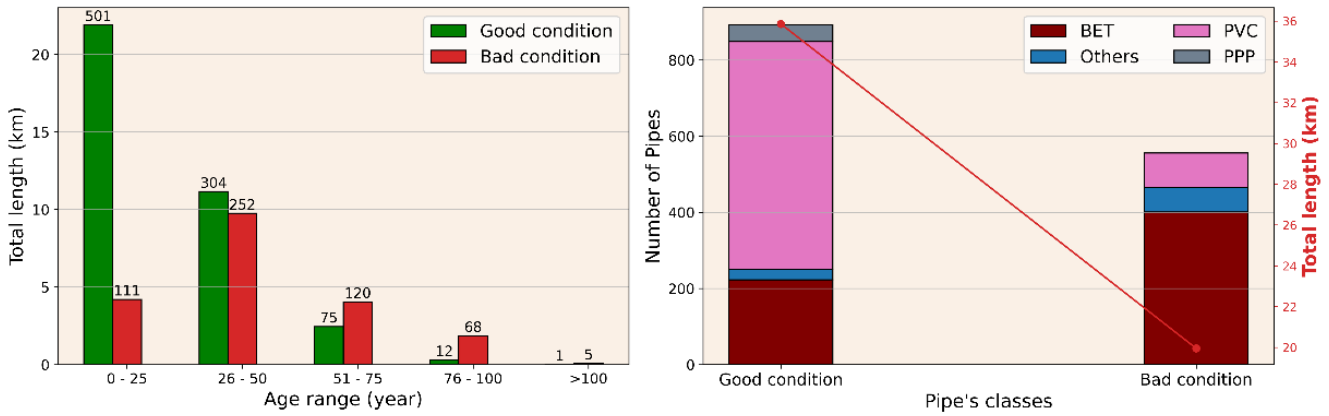

FIGURE 5. Different classes of sewer pipes in the study area.

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	- $C = 2^{-10}, 2^{-9}, \ldots, 2^{9}, 2^{10}$ - Penalization: l_1, l_2 - Solver: *newton-cg, lbfgs, liblinear, sag, saga*	- $C = 2^{-5}$ - Penalization: l_2 - Solver: *newton-cg*	76.96
RC	- $\alpha = 0.1,0.2,\ldots,0.9,1.0$	- $\alpha = 0.9$	77.05
GNB	-	-	73.69
BNB	-	-	74.55
KNN	- $K_{neighbor} = 1,2,\ldots,99,100$ - 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* - $n_{feature} = 1,2,\ldots,18,19$	- Quality of a split: *gini* - $n_{feature} = 17$	73.86
RF	- Quality of a split: *gini, entropy* - $n_{feature} = 1,2,\ldots,18,19$ - $n_{tree} = 10,20,\ldots,990,1000$	- Quality of a split: *entropy* - $n_{feature} = 19$ - $n_{tree} = 260$	78.08
SVM	- Kernel function: *linear, rbf, poly, sigmoid* - $C = 2^{-15}, 2^{-14}, \ldots, 2^{4}, 2^{5}$ - $d = 1,2,\ldots,9,10$ - $\gamma = 2^{-10}, 2^{-9}, \ldots, 2^{2}, 2^{3}$	- Kernel function: *poly* - $C = 2^{3}$ - $d = 5$ - $\gamma = 2^{-3}$	77.74
MLP	- Activation function: *logistic, tanh, relu* - Solver: *lbfgs, sgd, adam* - $n_{neuron} = 1,2,\ldots,199,200$	- Activation function: *tanh* - Solver: *adam* - $n_{neuron} = 99$	78.26
ERT	- Quality of a split: *gini, entropy* - $n_{feature} = 1,2,\ldots,18,19$ - $n_{tree} = 10,20,\ldots,990,1000$	- Quality of a split: *entropy* - $n_{feature} = 2$ - $n_{tree} = 660$	77.39
AdaBoost	- $n_{boosting} = 10,20,\ldots,990,1000$ - $Learning\ rate = 0.1,0.2,\ldots,0.9,1.0$	- $n_{boosting} = 40$ - $Learning\ rate = 0.5$	77.39
GTB	- $n_{estimator} = 10,20,\ldots,990,1000$ - $Learning\ rate = 0.1,0.2,\ldots,0.9,1.0$	- $n_{estimator} = 20$ - $Learning\ rate = 0.1$	78.52
HGB	- $n_{iteration} = 1,2,\ldots,29,30$ - $Learning\ rate = 0.1,0.2,\ldots,0.9,1.0$	- $n_{iteration} = 7$ - $Learning\ rate = 0.4$	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* - $n_{neuron} = 1,2,\ldots,99,100$	- 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

RF model	Predicted Class: Good condition	Predicted Class: Bad condition
Actual Class: Good condition	141	33
Actual Class: Bad condition	30	86

AdaBoost model	Predicted Class: Good condition	Predicted Class: Bad condition
Actual Class: Good condition	144	30
Actual Class: Bad condition	36	80

GTB model	Predicted Class: Good condition	Predicted Class: Bad condition
Actual Class: Good condition	133	41
Actual Class: Bad condition	29	87

SVM model	Predicted Class: Good condition	Predicted Class: Bad condition
Actual Class: Good condition	129	45
Actual Class: Bad condition	30	86

HGB model	Predicted Class: Good condition	Predicted Class: Bad condition
Actual Class: Good condition	142	32
Actual Class: Bad condition	40	76

GP model	Predicted Class: Good condition	Predicted Class: Bad condition
Actual Class: Good condition	138	36
Actual Class: Bad condition	37	79

ERT model	Predicted Class: Good condition	Predicted Class: Bad condition
Actual Class: Good condition	138	36
Actual Class: Bad condition	40	76

RC model	Predicted Class: Good condition	Predicted Class: Bad condition
Actual Class: Good condition	134	40
Actual Class: Bad condition	38	78

KNN model	Predicted Class: Good condition	Predicted Class: Bad condition
Actual Class: Good condition	139	35
Actual Class: Bad condition	42	74

LG model	Predicted Class: Good condition	Predicted Class: Bad condition
Actual Class: Good condition	141	33
Actual Class: Bad condition	46	70

ANN-2HLs model	Predicted Class: Good condition	Predicted Class: Bad condition
Actual Class: Good condition	142	32
Actual Class: Bad condition	43	73

MLP model	Predicted Class: Good condition	Predicted Class: Bad condition
Actual Class: Good condition	145	29
Actual Class: Bad condition	54	62

DT model	Predicted Class: Good condition	Predicted Class: Bad condition
Actual Class: Good condition	138	36
Actual Class: Bad condition	49	67

ANN-1HL model	Predicted Class: Good condition	Predicted Class: Bad condition
Actual Class: Good condition	127	47
Actual Class: Bad condition	42	74

GNB model	Predicted Class: Good condition	Predicted Class: Bad condition
Actual Class: Good condition	152	22
Actual Class: Bad condition	63	53

BNB model	Predicted Class: Good condition	Predicted Class: Bad condition
Actual Class: Good condition	146	28
Actual Class: Bad condition	59	57

ANN-3HLs model	Predicted Class: Good condition	Predicted Class: Bad condition
Actual Class: Good condition	143	31
Actual Class: Bad condition	45	71

FIGURE 7. Confusion matrix of each machine learning model.

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.

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

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

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

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

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

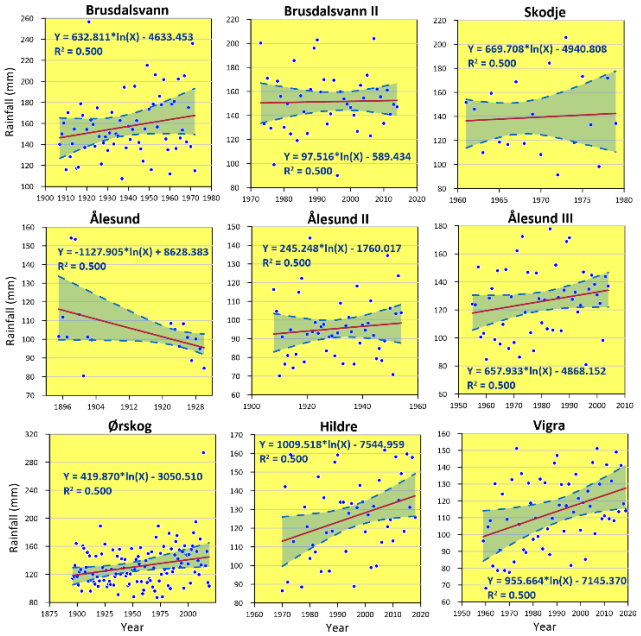

FIGURE 8. Fitting functions for rainfall interpolation.

$t(t = 2022, 2042, 2072)$. 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

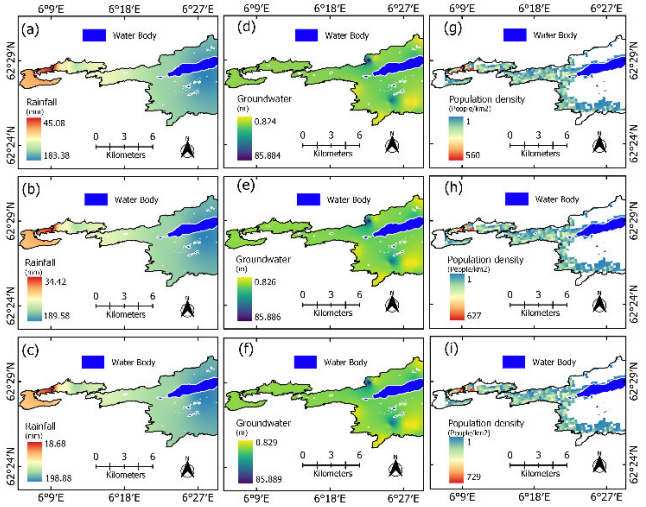

FIGURE 9. Maps of interpolated rainfall, groundwater, and population in the years 2022, 2042, and 2072.

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.

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

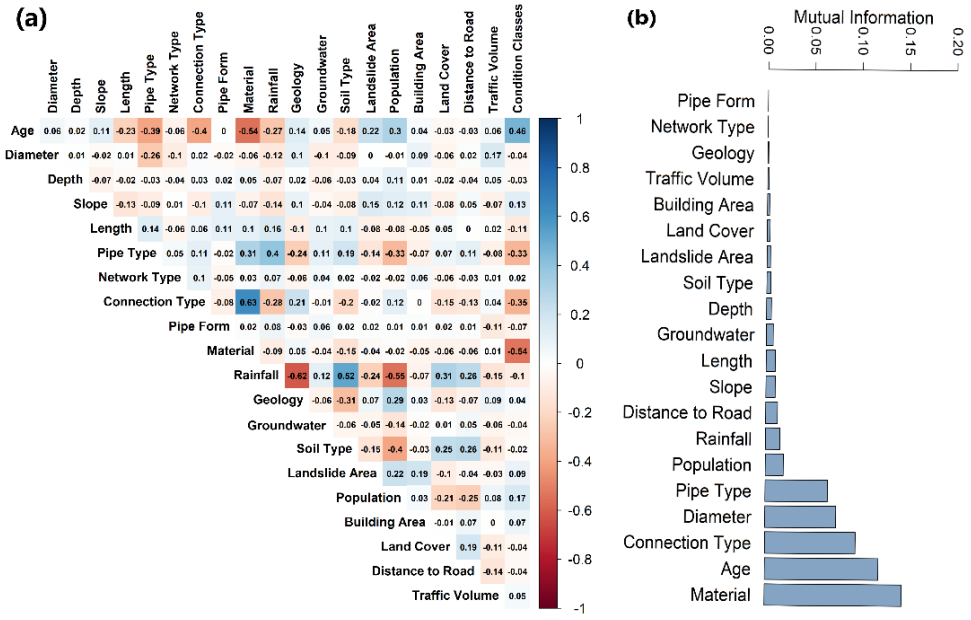

FIGURE 10. The filter feature selection methods: (a) PR correlation and (b) MI.

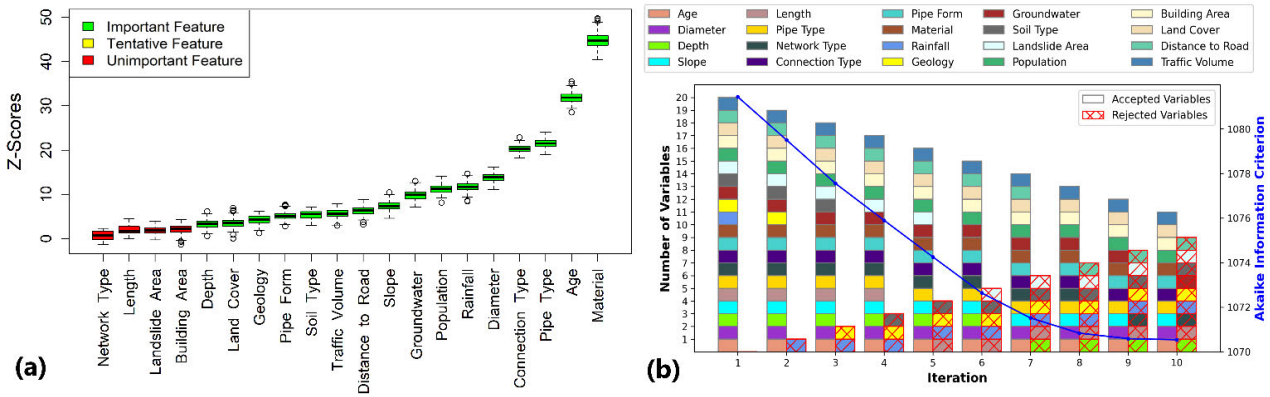

FIGURE 11. The wrapper feature selection methods: (a) Boruta and (b) SFS.

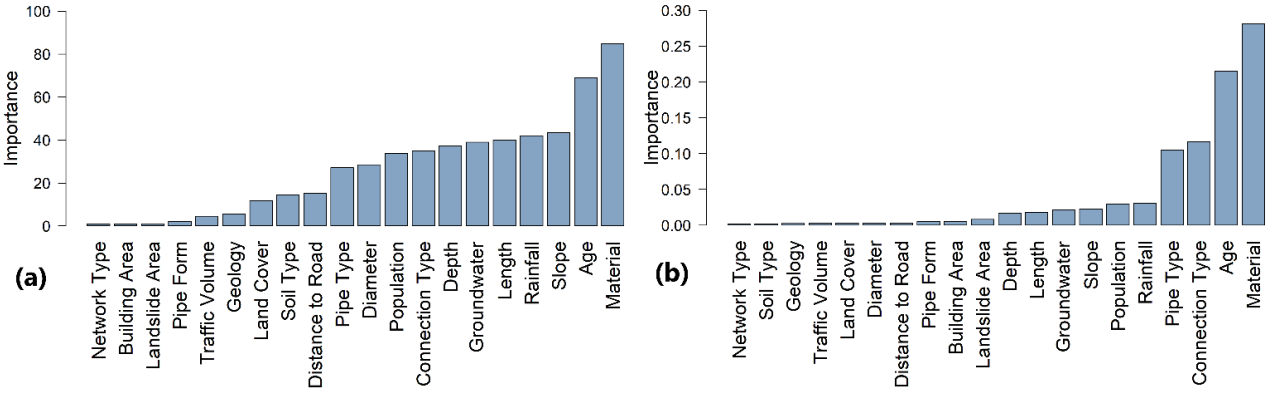

FIGURE 12. The embedded feature selection methods: (a) RF and (b) RFE.

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

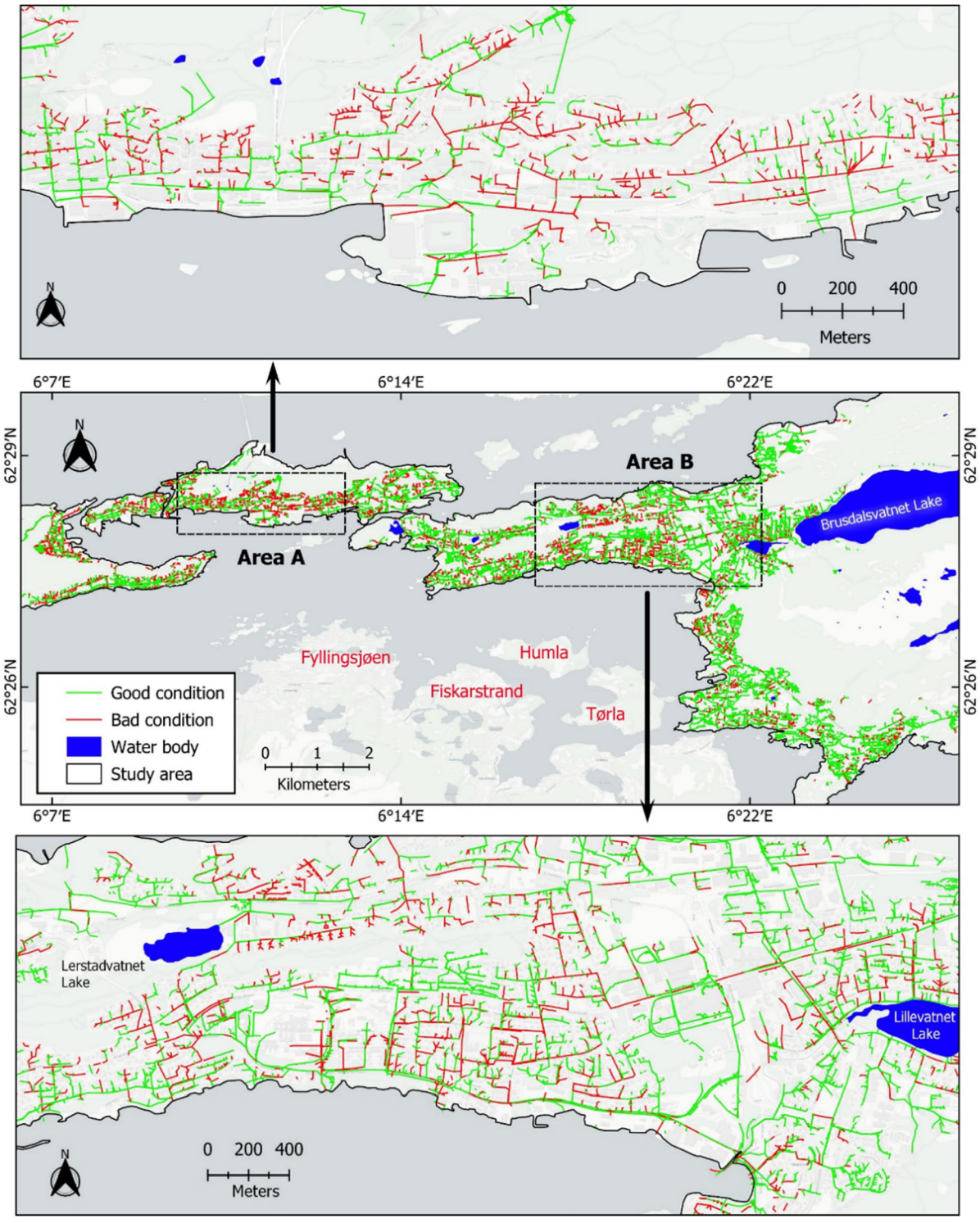

FIGURE 13. The sewer condition map in 2022 in the study area.

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,

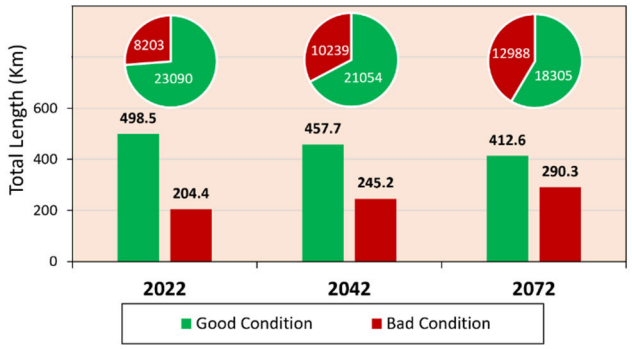

FIGURE 14. Summary of predicted classes in the study area.

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). *Dataflyt—Klassifisering 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.