IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 June 2023, accepted 4 July 2023, date of publication 7 July 2023, date of current version 13 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3293089

RESEARCH ARTICLE

Machine Learning-Based Predictive Model of Ground Subsidence Risk Using Characteristics of Underground Pipelines in Urban Areas

SUNGYEOL LEE, JAEMO KANG, AND JINYOUNG KIM
Department of Geotechnical Engineering Research, Korea Institute of Civil Engineering and Building Technology (KICT), Goyang 10223, South Korea
Corresponding author: Sungyeol Lee (leesy@kict.re.kr)

Research for this paper was carried out under the KICT Research Program (project no. 20230116-001, Underground Utilities Diagnosis and Assessment Technology (4/4)) funded by the Ministry of Science and ICT.

ABSTRACT In this study, a machine learning-based prediction model was developed using the attribute information of underground pipelines and the history information of ground subsidence in order to predict the risk level of ground subsidence in urban areas. The target area was divided into a grid with sizes of 100m×100m, 300m×300m, and 500m×500m, and the attribute information of underground pipelines in the grid and ground subsidence data were utilized to build a dataset. For input data, the pipeline's diameter, the number of years used, and density were selected based on the pipeline's length as the basic unit. Additionally, the risk level of ground subsidence was determined as the output data using historical information. A total of 36 datasets were built according to the conditions, and factors with significant correlation were selected through a correlation analysis of the datasets. The developed datasets were divided into training data and evaluation data. The synthetic minority oversampling technique was used to resolve the data imbalance. The model performance evaluation indexes used in this study were F1-score and AUC(Area Under the Curve). The performance of each model was compared, and the comparison results showed that a model that applied a preprocessed dataset with 500m×500m grid size, 10 years in use, 100mm pipeline diameter, and 1–2 ground sinks in Level 1 risk range to the LGBM(Light Gradient Boosting Model) classifier derived the best evaluation indexes(F1-Score:0.750, AUC:0.840). The map was found to be effective for predicting the risk level of ground subsidence in urban areas.

INDEX TERMS Ground subsidence, machine learning, prediction model, risk map, underground pipeline.

I. INTODUCTION

As road subsidence(Ground Subsidence) frequently occur, in particular around urban areas where the population density is high, it is necessary to manage underground pipelines and various facilities distributed underground. The road subsidence phenomenon has increased around urban areas of metropolitan cities. As the population inflow into cities has increased, facilities of traffic projects such as subways and underground passes and utility pipelines such as water supply and sewerage, telecommunications, and electric power lines are constructed without systematic plans, which increases the risk of road subsidence. In particular, as the aging of these facilities accelerates, the frequency of road subsidences and the risk have also increased [1], [2].

Ground subsidence (sink) is an academic term that expresses a sudden collapse of the ground surface locally and vertically [3]. It is different from a sinkhole, which is a phenomenon in which limestone in the ground is dissolved in groundwater and collapses to the surface layer, a distinction that needs to be made. The causes of road subsidences are the inflow of upper earth and sand to the defective part of aged pipelines or the outflow of earth and sand along with groundwater due to the leakage of water from the damaged part, which create an empty hole. Because of this, the shear strength of soil is reduced, thereby making a road

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

subsidence [4]. As another cause, changes in groundwater in the surrounding ground may occur due to the indiscriminate discharge of groundwater during foundation excavation construction, resulting in a significant ground subsidence. In addition, when backfilling is not properly done during the underground excavation, a hollow space is created underground, through which earth and sand are discharged, thereby making a ground subsidence [5], [6].

Currently, ground subsidences are managed in many different ways. The government of the state of Florida (USA) requires homeowners to purchase sinkhole coverage as part of the standard homeowners insurance policy, and provides location information of underground facilities and a 'one-call' system for underground facility managers, excavators, and demolition workers to avoid and prevent damage to underground facilities during excavation or demolition [7], [8]. In addition, Tokyo (Japan) has conducted regular mandatory investigations of hollow holes underneath its roads, evaluating the risk level and determining the priority of repair through ongoing research [9], [10]. In Singapore, the Land Transportation Authority inspects road subsidences periodically and has implemented a repair system within 24 hours with a 24-hour reporting window, as well as operating a website for road maintenance [11]. As such, efforts to manage the risk of ground subsidence are being made by the management authority, and various studies are being conducted to predict the risk level of ground subsidence.

More recently, research has been conducted on facility risk monitoring and underground pipeline management using advanced machine learning and deep learning technologies. Advanced machine learning and Internet of Things (IoT) techniques have been used to monitor the condition of motors and detect vulnerabilities to cyber attacks [12], [13]. In addition, the Alternative Transients Program has been used to calculate the induced voltage of normal and abnormal gas pipes buried near overhead transmission lines (OHTLs) [14].

Study findings have been published that used the analytic hierarchy process (AHP) and a decision tree, which is a machine learning algorithm, to produce the importance and weights of influencing factors in urban ground subsidence that occur due to various causes [15]. A study that proposes a regression equation to calculate the risk level of ground subsidence in urban areas in Korea through logistic regression analysis was also published [16]. Furthermore, a model that can predict the risk level of ground subsidence was published using the number of years used and the diameters of pipelines among the attribute information of underground pipelines [17], [18].

However, research on determining the appropriate grid partition size for predicting the risk of ground subsidence is currently limited, and there is a lack of developed machine learning models with high reliability trained specifically for regions where ground subsidence occurs often. Thus, in this study, the regions in a metropolitan area of South Korea (referred to as Region A) where ground subsidence frequently occurs were divided into grid cells based on specific conditions. The study selected the number of years that underground utilities had been in use, their diameter, length, and the density of pipelines with a high correlation to ground subsidence as influencing factors. Based on this, a machine learning prediction model was proposed for assessing the risk of ground subsidence. To do this, the results of machine learning models in which datasets with various conditions were applied were compared to select a model with the optimum performance. And the importance of the influencing factors used in the classification of ground subsidence risk level by the machine learning model was proposed through the selected model. In addition, a prediction map of ground subsidence risk level in the target area was prepared through the model.

II. STUDY METHOD

This study aimed to construct a ground subsidence prediction model using machine learning based on the attribute information of underground pipelines and the history of ground subsidence, and to select a dataset that produces the optimal performance. To do this, Region A was divided into grids of 100m x 100m, 300m x 300m, and 500m x 500m. The study extracted the attribute information, density, and ground subsidence history of underground pipelines included in the divided grid, and built 36 datasets by dividing them into categories according to the conditions. The dataset was divided into 80% training dataset and 20% test dataset for model evaluation. The developed data had an imbalance in the risk of ground subsidence. Thus, the synthetic minority oversampling technique (SMOTE) was used to balance the data by increasing the minority data in the training data. For the development of the relevant prediction models three machine algorithms were investigated, namely random forests (RF), XGBoost and LightGBM.

The dataset was tuned with hyperparameters that can optimize each model to check the evaluation index of the model. The model that shows the best evaluation index and dataset division conditions was selected, and the importance of the influence factors, which were used to classify the risk level of ground subsidence by the model, was verified. In addition, the risk map of ground subsidence in the target region was prepared using the selected best model. Figure 1 shows a flow chart of the study.

A. TARGET REGION

Region A was selected as the target region to develop a prediction model of the ground subsidence risk level, as it had the most ground subsidence (around 29%) in the metropolitan area in Korea. The target region was a stream area, which was characterized by the sand and soil particles introduced by the change in the river flow. This alluvial deposit is characterized by soft ground due to the layered accumulation of sand and earth for a long period and the smooth flow of groundwater. Once the groundwater level changes, particles such as sand

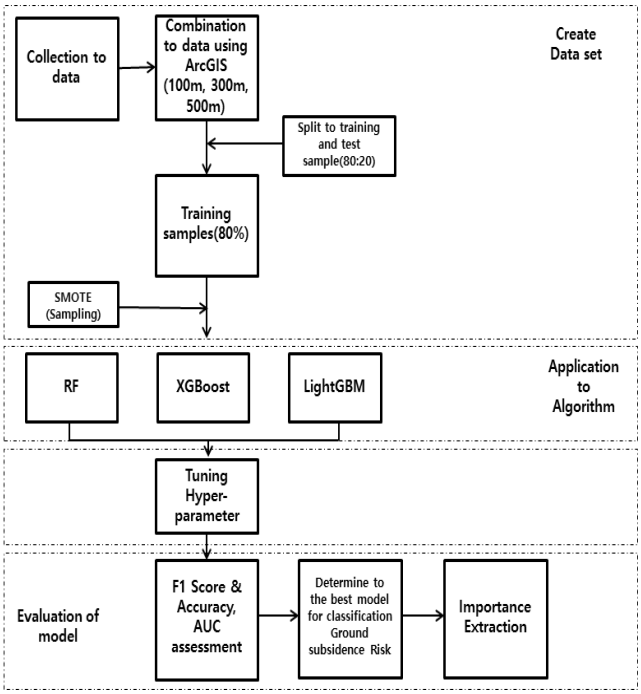

FIGURE 1. Flow Chart of this study.

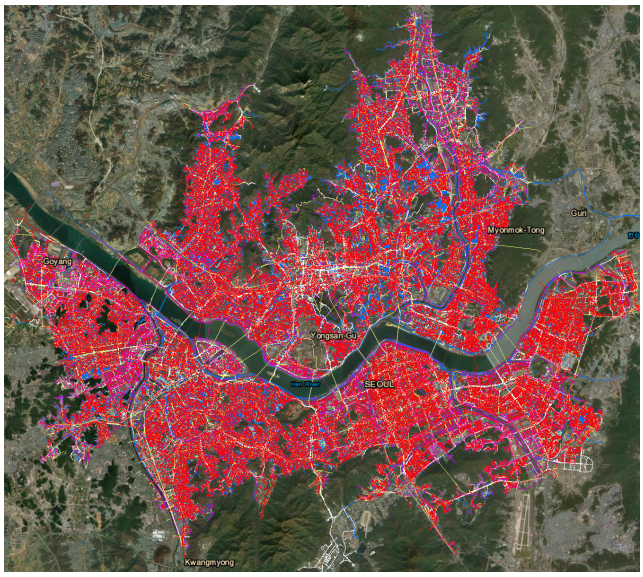

FIGURE 2. Space information on six types of pipelines in seoul.

and soil are discharged through the flow of groundwater, which enlarges the empty space and leads to road subsidences. Geologically, riverbeds or riverbed sediments require more careful and safe design and countermeasures because the ground is soft and the depth of the soft ground is deep, while the flow of groundwater is fast. Thus, underground pipelines are easily damaged due to ground subsidence at riverbed sediment regions, which are historically riverbeds, regions where the soft ground depth was deep, or sandy soil regions where the groundwater flow is fast. Accordingly, empty space may be expanded due to the damage to the connecting part of underground pipelines, which requires management.

III. DATA

A. CHARACTERISTICS OF THE RAW DATA

There are typically six types of underground pipelines buried in urban areas. These typical six types are water pipes, sewer pipes, communication pipes, power cables, gas pipes, and heat pipes, and the mapping using the data of the six types of utility pipelines throughout the urban area is shown in Figure 2.

The numbers and lengths of each pipeline in the entire urban area where the study target belongs are 1,048,566/10,283,352m of water pipes, 394,958/10,827,968m of sewer pipes, 162,735/9,431,645m of communication lines, 168,431/3,125,244m of power lines, 666,820/10,584,726m of gas lines, and 61,939/2,434,560m of heating pipes. The longest total length of pipeline was for sewer pipes, followed by gas line, water pipe, and communication line, and these four types of pipelines were highly dense.

More than 60% of road sinks that occurred in the entire urban area of the study target are caused by damage in sewer pipes [1], [2].

In Japan, the situation is similar, as around 30% to 50% of the road sinks that occurred in the 10-year period from 1999 to 2009 were caused by damage to sewer pipes. However, road sinks do not occur as the result of a single cause. Rather, they occur due to a combination of causes, such as the ground condition around the pipeline and the impact of groundwater caused by civil works or underground structures, as well as aged underground pipelines. In this study, factors that affected road sinks were selected in two specific districts where many road sinks occur, and a correlation analysis was conducted.

The influence level of the factors was analyzed through statistical analysis and machine learning using the selected factors, and the risk level of road sink occurrence in the selected region was predicted. Data used included the total length of six pipelines distributed over two districts located in the target region in this study, which was around 4,958,953m, and the number of road sinks that occurred in the target region from 2008 to 2016, which was 1,061.

B. DATA PREPROCESSING

To build a prediction model of risk levels of ground subsidences using machine learning, preprocessing of the raw data was conducted. The ArcGIS program was used to divide the target region into grids with square sizes of 100m×100m, 300m×300m, and 500m×500m, and the attribute information of underground pipelines included in the grid and history information of ground subsidences were used to extract data that were applied to machine learning.

In this study, attribute information of six types of underground pipelines was obtained, which was then integrated

into a single type of attribute information to extract data. The attribute information of underground pipelines used as the input data included various factors of data such as the pipe's materials, length, diameter, year of burial, depth, and density. However, available data after omitted and error values were excluded were the length of the pipe, the burial year, diameter, and density. Thus, to build a dataset in this study, the number of years used was selected, which was calculated through the diameter of the entire pipelines and the burial years, the basic unit was set to the length of the pipeline, and the density was obtained by the method of calculating the length of the pipeline that corresponds to the unit area through the linear density analysis.

To select a ground subsidence risk prediction model that exhibits the optimum performance, attribute information of underground pipelines and the risk level of ground subsidence was divided by a certain section. The diameters of the pipelines, which were input data, were divided into 50 mm and 100 mm, and the numbers of years used was divided into five and 10 years.

There are no quantified criteria to define the risk level of ground subsidence, which is the output data. Thus, this study aimed to check the definition of the risk level that showed the optimal classification performance by changing the criteria based on which the risk level grade was calculated. To do this, the total number of ground subsidence occurrences in the grid was divided into a total of three stages by summing them. Then, a dataset was constructed by varying the number of ground subsidence occurrences within the grid according to certain conditions for defining Level '1'. This can be explained in more detail as follows: As presented in Table 1, the risk levels were divided into '0', '1', and '2.' Level '0' represents zero occurrences of ground collapses within the grid. Level '1' is defined based on the conditions of having one occurrence, 1-2 occurrences, or 1-3 occurrences. Level '2' is defined with 2 occurrences, 3 occurrences, or more than 4 occurrences (depending on the conditions of Level '1'). With these three conditions, a dataset was constructed. Furthermore, 3(1) represents the condition in which the number of ground subsidence occurrences within the grid is set to 1 for determining Grade '1'. 3(1-2) indicates the risk level grade set with 1-2 occurrences, and 3(1-3) represents the risk level grade set with 1-3 occurrences. Table 2 presents the classification units and categories of the number of years used and diameter.

Table 3 presents the dataset conditions used in this study. In this table, the column labeled "System" represents the size classification of the grid. The columns labeled "Year" and "Diameter" indicate the number of years used and the diameter classification unit, respectively. The dataset that was built according to each condition was applied to the machine learning algorithm to check the model's performance, enabling the most suitable data division condition and model to be selected. Table 2 presents the classification units and categories of the number of years used and diameter.

TABLE 1. Calculation of risk level according to the number of ground subsidence occurrence.

Risk Level / Risk Grade (Level 1's range)	Level 0	Level 1	Level 2
3(1)	0	1	2 or more
3(1-2)	0	1.2	3 or more
3(1-3)	0	1, 2, and 3	4 or more

TABLE 2. Category of factors.

Factors	Unit	Category
Year (year)	5	1~5, 6~10, 11~15, 16~20, 21~25, 26~30, 31~35, 36~40, 41~45, 46~50
	10	1~10, 11~20, 21~30, 31~40, 41~50
Diameter(mm)	50	1~50, 51~100, 101~150, 151~200, 201~250, 251~300, 301~350, 351~400, 401~450, 451~500, 501~550, 551~600, 651~700, 751~800, 851~900, 951~1000
	100	1~100, 101~200, 201~300, 301~400, 401~500, 501~600, 601~700, 701~800, 801~900, 901~1000

C. NUMBER OF DATA RECORDS ACCORDING TO CONDITIONS

When the target area was divided into a grid with a square size of 100m×100m, a total of 6,315 grid squares were generated. When the target area was divided into a grid with a square size of 300m×300m, a total of 826 grid squares were generated. Finally, when the target area was divided into a grid with a square size of 500m×500m, a total of 325 grid squares were generated. Furthermore, the number of data records in each level varies according to the risk level determination range. Table 4 presents the number of data records according to grid square size and ground subsidence risk level.

IV. DATA CORRELATION ANALYSIS

To develop an effective prediction model for ground subsidence risk, a dataset was built according to the divisions

TABLE 3. Condition of datasets.

Number	System	Year	Diameter	Risk Grade (Grade 1's range)
1	100m×100m	5		3(1)
2			50	3(1-2)
3				3(1-3)
4				3(1)
5			100	3(1-2)
6				3(1-3)
7		10		3(1)
8			50	3(1-2)
9				3(1-3)
10				3(1)
11			100	3(1-2)
12				3(1-3)
13	300m×300m	5		3(1)
14			50	3(1-2)
15				3(1-3)
16				3(1)
17			100	3(1-2)
18				3(1-3)
19		10		3(1)
20			50	3(1-2)
21				3(1-3)
22				3(1)
23			100	3(1-2)
24				3(1-3)
25	500m×500m	5		3(1)
26			50	3(1-2)
27				3(1-3)
28				3(1)
29			100	3(1-2)
30				3(1-3)
31		10		3(1)
32			50	3(1-2)
33				3(1-3)
34				3(1)
35			100	3(1-2)
36				3(1-3)

of attribute information of underground pipelines and the risk level of ground subsidences. The input data in the dataset (attribute information of underground pipelines) were set to the independent variables, and the output data (ground subsidence risk level) were set to the independent variable to conduct a correlation analysis. Significant input data were selected and applied to the model.

Correlation analysis is an analysis method that identifies whether there is a linear relationship between independent and dependent variables, and the size of the relationship between variables. The size of the correlation is calculated as presented in (1). The correlation coefficient has a range from -1 to +1. The larger the correlation between two variables is, the closer the coefficient is to +1 or -1 [20], [21].

TABLE 4. Condition of datasets.

Grid	Grade 1's range	0	1	2
100m×100m	1	5,671	443	201
	1–2	5,671	571	73
	1–3	5,671	618	26
300m×300m	1	554	105	167
	1–2	554	154	118
	1–3	554	182	90
500m×500m	1	158	48	119
	1–2	158	78	89
	1–3	158	96	71

$$\mathrm{Corr}(X, Y) = \rho(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_x \sigma_y} \tag{1}$$

To determine whether the correlation coefficient between two variables is significant after calculating the coefficient, a hypothesis of the correlation ρ of the population is tested. Here, the hypothesis and test statistic are presented in (2) [22].

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{2}$$

A correlation analysis of 36 datasets was conducted to select statistically significant ($p<0.05$) factors regardless of the correlation coefficient size between variables. The selected factors were applied to the algorithm. The density of the pipelines was a statistically significant factor in all conditions.

V. MACHINE LEARNING MODELS FOR SUBSIDENCE RISK PREDICTION

A. PRELIMINARIES ON MACHINE LEARNING ALGORITHMS

1) RANDOM FOREST (RF)

RF algorithm is an ensemble model based on a regression and classification tree proposed by Breiman et al. [23], [24]. The ensemble model derives optimum results by repeating a single algorithm or learning multiple algorithms, producing better performance than that of learning a single model once. RF creates multiple tree algorithms and selects the best result based on the results derived from each of the trees. As such, the RF algorithm, which is composed of classification and regression trees, has the following characteristic advantages: a small risk of overfitting and unrestricted selection of variables, as well as excellent model performance to derive results

even if the correlation between data is not close [25], [26]. Thus, RF has been widely used to solve regression and classification problems when applying a machine learning technique in various fields [27], [28].

RF predicts the outcome as a binary value of 0 or 1 as presented in (3) after extracting an arbitrary number of input data from a number of single-algorithm predictors and performing a final decision by majority vote on the results derived from each predictor, where, $y_i = f_i(X)$, and w_i refers to the weight. If the calculated value is larger than the threshold value, the predicted value is 1, otherwise it is 0 [29].

$$F(X) = \sum w_i y_i \tag{3}$$

2) XGBOOST (EXTREME GRADIENT BOOSTING, XGB)

The XGBoost(XGB) algorithm is a model proposed to solve the overfitting problem found in linear or tree-based models. It was developed to improve the large scale of data processing and learning speed [30]. In XGBoost, multiple classifiers are created to learn in sequence, and the results derived in each model are reflected in the next model to solve a problem, which is a boosting technique. Its main hyperparameters are the number of trees and the depth, etc [31]. The calculation equation for the decision-making of XGBoost is presented in (4), where $\hat{y_i}$ refers to the i-th sample's prediction value and f_k refers to the prediction value where the k-th tree's sigmoid function is applied. The output is derived by summing all prediction values. The prediction value can be calculated using (5).

$$\hat{y}_i = \sum_{K=1}^{K} f_i(x_i) \tag{4}$$

$$\hat{y}_i = \frac{1}{1 + e^{-f(x_i)}} \tag{5}$$

The error is calculated using the difference between the prediction and real values in the tree, and the weight is calculated to reduce the error as presented in (6). $\hat{y_i}^{(t-1)}$ refers to the prediction value of the previous model, $h_t(x_i)$ refers to the tree trained by the current model, and η refers to the learning rate, which is the percentage of reflections from the prior model. The model's error is reduced by iterating this method [32].

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta h_t(x_i) \tag{6}$$

3) LIGHTGBM

LightGBM(LGBM) is a high-performance algorithm based on a tree algorithm. It is used to select a priority rank of influence factors and solve a regression and classification problem. A boosting technique similar to XGBoost is applied here. It is characterized by its fast operation using partial data and reduction of features to shorten the operation time. Thus, LightGBM processes a large volume of data at a fast rate with a high degree of accuracy. It also derives the importance between influencing factors used. Thus, it is widely used [33], [34]. LightGBM calculates the loss function using cross-entropy. The equation for calculating the cross entropy is presented in (7), where N is the number of samples, K is the number of classes, $y_{i,j}$ refers to the binary variable indicating whether the i-th sample belongs to the j-th class, and $p_{i,j}$ refers to the probability that the i-th sample belongs to the j-th class. LightGBM derives its results by learning to update the model while minimizing the CE received from the previous model [35].

$$CE = \frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} y_{i,j} \log(p_{i,j}) \tag{7}$$

TABLE 5. Model evaluation according to AUC.

AUC	Evaluation
AUC≧0.9	Excellent
0.8≦AUC<0.9	Good
0.7≦AUC<0.8	Fair
AUC<0.7	Poor

B. MODEL EVALUATION INDEXES

In this study, a dataset was composed according to conditions and this dataset is applied to machine learning models. Then, the results are compared. The selected evaluation indexes for the comparison of the model's performance were accuracy, F1-score, AUC, which are normally used as the evaluation indexes in the classification model.

Accuracy is an index that can intuitively evaluate the reliability of the model. However, if it is used in a dataset that exhibits unbalanced data features, it is difficult to clearly evaluate a model. Thus, in this study, accuracy was used to check whether the model was overfitted by comparing the score (accuracy) of train and test data. The smaller the score difference was, the lower the overfitting risk.

F1-score is mainly used as an objective evaluation index for classification models where unbalanced data are applied. It is an index that exhibits the harmonic mean of precision (the number of actual true cases out of the predicted true cases by the model) and recall (the number of predicted true cases out of actual true cases in the data) [18]. Using this, it can evaluate whether a prediction model properly classifies each of the classes [36], [37]

AUC is an index that can evaluate the model's performance through the area of the receiver operating characteristic (ROC) curve. An ROC curve is displayed using recall and specificity. Table 5 presents the criteria that show the model's performance according to the AUC values proposed by Fawcett [38]. If the AUC is larger than 0.8, the model's performance is evaluated as good.

(8)–(12) present the methods to calculate the evaluation indexes of the model.

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

$$\text{Recall(Sensitivity)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

$$\text{F1Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{11}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \tag{12}$$

C. RESULTS

To build a ground subsidence prediction model, the scikit-learn library, which included Python 3.8 and machine learning packages, was used, and the algorithms used were RF, XGB, and LGBM. The tuning was conducted with hyperparameters, which derived the optimum result through the trial-and-error method after applying 36 datasets to the algorithms, and the results were compared. Then, the model that exhibited the optimum performance was selected.

Table 6 presents the derived results of the prediction model for ground subsidence risk level using the test data, which was 20% of the total data. The model's accuracy was verified through the test score, and overfitting of the model was determined based on the difference from the training score [39], [40]. The model's performance was evaluated using the F1-score and AUC, which refers to the area of the ROC curve, due to the characteristics of unbalanced data. The model's results where each of the datasets was applied were compared to select the optimum model.

The overall comparison results of the model's evaluation indexes revealed that when the dataset had 10 years of use, 100mm pipeline diameter, and the range of Step 1 was set to 1–2 ground subsidences in the data with a 500m×500m grid square size, the model where this dataset was applied to the LGBM algorithm derived the highest F1-score (0.750) and 0.80 or higher AUC. In addition, the comparison of the difference between the train and test scores showed that overfitting was avoided.

The model results according to the grid size showed that the model produced the highest F1-score (0.640) in the data divided with a 300m×300m grid square size when the dataset had 10-year use, 50mm pipeline diameter, and the range of Step 1 risk level was one to three ground subsidences and this data set was applied to RF. In addition, the best model (0.430) in the data divided with 100m×100m grid square size was revealed when the dataset had five-year use, 100mm pipeline diameter, and the range of Step 1 risk level was set to one to two ground subsidences, and this dataset was applied to the RF algorithm. As such, the best index was derived when the grid division for building a prediction model of the ground subsidence risk level was 500m×500m.

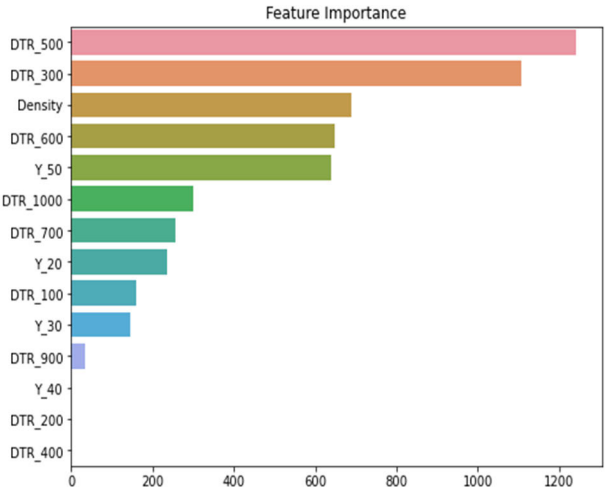

FIGURE 3. Importance of the factors used to model.

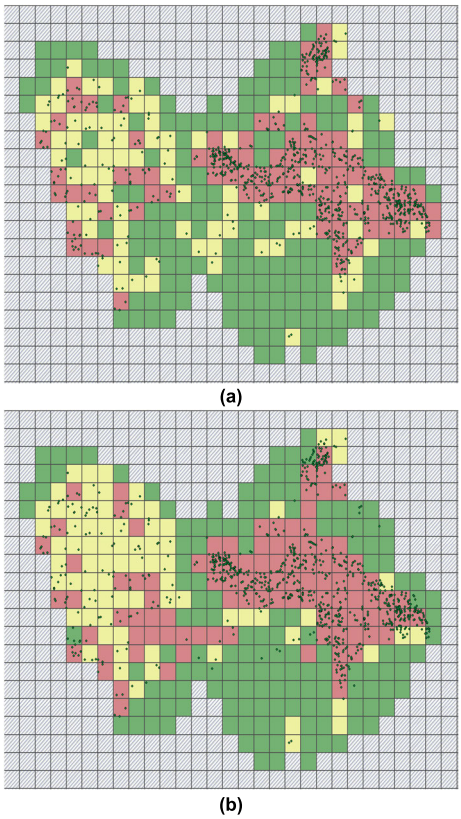

FIGURE 4. Map of ground subsidence risk from (a) Map of ground subsidence risk using real data, (b) Prediction map of ground subsidence risk level.

The averages of the evaluation indexes of each dataset were 0.36, 0.59, and 0.57 at the condition of 100m×100m, 300m×300m, and 500m×500m grid square size, respectively. Thus, the dataset divided with 100m×100m grid square size had the worst performance. This was due to the decrease in the number of ground subsidence occurrences inside the grid, as the target region was finely divided, which

TABLE 6. Results of the machine learning model.

	RF				XGB				LGBM			
	Train Score	Test Score	F1-Score (macro)	AUC (macro)	Train Score	Test Score	F1-Score (macro)	AUC (macro)	Train Score	Test Score	F1-Score (macro)	AUC (macro)
1	0.542	0.413	0.260	0.590	0.554	0.421	0.270	0.590	0.554	0.421	0.270	0.590
2	0.807	0.727	0.410	0.810	0.813	0.731	0.410	0.800	0.837	0.751	0.410	0.800
3	0.875	0.739	0.400	0.810	0.842	0.732	0.400	0.780	0.838	0.721	0.410	0.790
4	0.414	0.386	0.260	0.590	0.475	0.363	0.240	0.550	0.443	0.379	0.250	0.560
5	0.801	0.743	0.430	0.820	0.808	0.734	0.410	0.810	0.874	0.762	0.420	0.820
6	0.796	0.686	0.390	0.800	0.827	0.725	0.400	0.800	0.794	0.704	0.380	0.780
7	0.536	0.468	0.280	0.590	0.524	0.411	0.270	0.600	0.522	0.447	0.270	0.610
8	0.807	0.730	0.420	0.810	0.822	0.721	0.390	0.800	0.840	0.741	0.400	0.820
9	0.793	0.706	0.400	0.790	0.829	0.730	0.400	0.780	0.804	0.722	0.390	0.780
10	0.615	0.510	0.300	0.590	0.547	0.477	0.290	0.590	0.580	0.509	0.300	0.590
11	0.750	0.700	0.420	0.810	0.803	0.713	0.410	0.810	0.856	0.747	0.420	0.830
12	0.810	0.700	0.380	0.810	0.847	0.747	0.400	0.790	0.826	0.735	0.390	0.800
13	0.763	0.669	0.610	0.840	0.750	0.651	0.590	0.850	0.758	0.681	0.620	0.840
14	0.772	0.675	0.630	0.850	0.728	0.645	0.590	0.830	0.702	0.614	0.540	0.800
15	0.775	0.681	0.620	0.840	0.744	0.651	0.580	0.820	0.719	0.663	0.610	0.820
16	0.756	0.669	0.610	0.850	0.744	0.651	0.580	0.840	0.767	0.687	0.620	0.840
17	0.714	0.633	0.570	0.830	0.697	0.627	0.560	0.830	0.684	0.620	0.570	0.800
18	0.758	0.663	0.600	0.840	0.775	0.663	0.610	0.830	0.722	0.639	0.570	0.810
19	0.764	0.675	0.620	0.840	0.743	0.663	0.610	0.820	0.766	0.669	0.610	0.830
20	0.707	0.651	0.580	0.830	0.728	0.602	0.520	0.800	0.728	0.627	0.540	0.810
21	0.781	0.687	0.640	0.840	0.753	0.663	0.600	0.820	0.689	0.633	0.560	0.800
22	0.766	0.663	0.610	0.840	0.752	0.675	0.620	0.840	0.753	0.669	0.610	0.840
23	0.783	0.675	0.620	0.830	0.744	0.657	0.570	0.810	0.766	0.663	0.590	0.820
24	0.707	0.639	0.570	0.830	0.744	0.663	0.600	0.820	0.783	0.657	0.590	0.830
25	0.631	0.615	0.490	0.760	0.664	0.615	0.520	0.690	0.783	0.631	0.580	0.760
26	0.721	0.615	0.560	0.840	0.615	0.554	0.440	0.700	0.618	0.523	0.420	0.700
27	0.718	0.615	0.580	0.840	0.751	0.631	0.580	0.840	0.737	0.662	0.620	0.790
28	0.748	0.646	0.620	0.830	0.683	0.554	0.500	0.690	0.637	0.523	0.480	0.710
29	0.767	0.646	0.610	0.860	0.637	0.585	0.520	0.740	0.626	0.538	0.500	0.740
30	0.740	0.692	0.670	0.840	0.745	0.631	0.580	0.810	0.724	0.615	0.560	0.810

TABLE 6. *(Continued.)* **Results of the machine learning model.**

31	0.775	0.600	0.540	0.820	0.675	0.569	0.450	0.690	0.669	0.569	0.460	0.710
32	0.770	0.631	0.570	0.820	0.734	0.662	0.600	0.820	0.737	0.723	0.670	0.840
33	0.691	0.662	0.630	0.840	0.789	0.677	0.630	0.850	0.780	0.662	0.640	0.810
34	0.745	0.662	0.620	0.810	0.688	0.585	0.510	0.710	0.678	0.523	0.450	0.700
35	0.751	0.631	0.580	0.840	0.707	0.662	0.610	0.770	0.790	0.783	0.750	0.830
36	0.710	0.615	0.590	0.840	0.751	0.692	0.660	0.850	0.762	0.677	0.650	0.850

TABLE 7. **Hyperparameters of the excellent model according to grid size.**

Grid	Model	Hyperparameter
100m×100m	RF	n_estimators=500, max_depth=7
300m×300m	RF	n_estimators=500, max_depth=4
500m×500m	LGBM	n_estimators=300, max_depth=3, learning_rate=0.001

increased the data that did not have ground subsidences. Table 7 presents the hyperparameters of excellent models according to the grid size.

D. IMPORTANCE ANALYSIS

The LGBM classifier, which was selected as the fittest model to predict the risk level of ground subsidences in urban areas, includes a function that can produce the importance of the input data used for the prediction of the ground subsidence risk level. Thus, the importance of the factors used in the prediction of the ground subsidence risk level was verified using this function (Figure 3). The results showed that DTR_500 (401mm–500mm pipeline diameter) had the highest importance followed by DTR_300, density, DTR_600, and Y_50 (41–50 years of use). On the other hand, the importance of the number of years used was relatively lower than that of pipeline diameter.

VI. MAP OF GROUND SUBSIDENCE RISK

Figure 4 shows (a) the map of the ground subsidence risk level using real data and (b) the map that predicts the ground subsidence risk level in the target area through the derived optimum classifier. The areas marked green in the risk map are Class ''0,'' which is relatively safe in terms of ground subsidences, while the yellow are Class ''1'' and the red color are Class ''2,'' a high-risk area. The points on the map indicate real ground subsidences that occurred in the past.

As shown in the prediction risk map, the center area where ground subsidences were concentrated was well predicted, but the area where fewer ground subsidences occurred was not well predicted. Moreover, some high-risk area predictions by the classifier actually had no ground subsidences in the past. Although this may be regarded as a kind of prediction error of the classifier, it can also be viewed as a ground subsidence-prone area in the future, which requires preparation against ground subsidence-related accidents.

VII. CONCLUSION

The main cause of ground subsidence in urban areas was found to be damage to underground pipes. Thus, in this study, the target area was divided into a certain size of grid, and the underground pipeline attribute information and ground subsidence history information contained in the grid were applied to the machine learning classifier to select the optimal ground subsidence risk prediction model. The data applied to the classifier were selected through correlation analysis to create a dataset, and it was found that the density of underground pipelines showed a significant correlation in all datasets. Applying the datasets of a total of 36 cases to the classifiers showed that when the grid size was 500m×500m, the number of years used was 10 years, the pipeline diameter was 100 mm, and classification was done using the density, the output data, obtained by applying the dataset where the risk level was set to '1' with a range of 1-2 occurrences of ground subsidence to an LGBM classifier model, showed the best evaluation index, achieving an F1-Score of 0.750 and an AUC of 0.830.

In addition, the evaluation indexes of the classifiers according to the grid size were compared and the results exhibited that the indexes of classifiers where the dataset was divided into a 100m×100m grid showed a relatively low performance (F1-Score average: 0.36). This was due to the data imbalance as the number of ground subsidence occurrences '0' rapidly increased because of the narrowed range with the decrease in the grid size. The number of ground subsidence occurrences in the grid increased as the grid size increased, minimizing the data imbalance and improving the model's performance.

Furthermore, the risk level of ground subsidence in the target area was displayed on a map using the best performance

model and compared with the history information of real ground subsidence. The results of this comparison showed that places where many ground subsidences occurred in the past were relatively well-predicted, whereas places where ground subsidences occurred sporadically were not well-predicted in terms of accuracy.

Although it may be challenging to pinpoint the exact locations of ground subsidence areas based on a grid size of 500m×500m, it is possible to proactively address potential ground collapse occurrences in high-risk areas by predicting the risk levels at the grid size unit, and then employing techniques such as ground penetrating radar surveys. Ground subsidence has complex causes. For this reason, additional studies are needed to collect data such as attribute information and geotechnical information of subways and underground tunnels to add available factors, and studies to improve model performance and reliability by expanding a target region will be needed in the future.

REFERENCES

[1] *Cause Analysis of Cavity at Seokchon Underground Roadway and Road Cavity*, Seokchon-Dong Cavity Cause Invest. Committee, Seoul City, South Korea, 2014.

[2] *The Road Subsidence Conditions and Safety Improvement Plans in Seoul*, Seoul Inst., 2016.

[3] T. Mukunoki, J. Otani, S. Nonaka, T. Horii, and R. Kuwano, "Evaluation of cavity generation in soils subjected to sewerage defects using X-ray CT," in *Proc. Int. Workshop X-Ray CT Geomaterials*, 2006, pp. 365–371.

[4] T. Mukunoki, N. Kumano, J. Otani, and R. Kuwano, "Visualization of three dimensional failure in sand due to water inflow and soil drainage from defective underground pipe using X-ray CT," *Soils Found.*, vol. 49, no. 6, pp. 959–968, Dec. 2009.

[5] R. Kuwano, T. Horii, H. Kohashi, and K. Yamauchi, "Defects of sewer pipes causing cave-in's in the road," in *Proc. 5th Int. Symp. New Technol. Urban Saf. Mega Cities Asia*, 2006, pp. 347–353.

[6] J. Y. Kim, J. M. Kang, C. H. Choi, and D. H. Park, "Correlation analysis of sewer integrity and ground subsidence," *J. Korean Geo-Environ. Soc.*, vol. 18, no. 6, pp. 31–37, 2017.

[7] Y. S. Han, "Proposal of the development direction on the special act on underground safety management for preparation of the proactive underground safety management system," *J. Korean Geotechnical Soc.*, vol. 34, no. 7, pp. 17–27, 2018.

[8] E. D. Zisman, "A standard method for sinkhole detection in the Tampa, Florida, area," *Environ. Eng. Geosci.*, vol. 7, no. 1, pp. 31–50, Feb. 2001.

[9] M. Sato, Y. Uno, and R. Ito, "Conditions of the cavity formation and sinkholes in the practical ground," in *Proc. 7th Asia–Pacific Conf. Unsaturated Soils*, 2019, vol. 7, no. 2, pp. 493–500.

[10] M. Sato, Y. Uno, and R. Ito, "Evaluation method for determining potential risks of detected underground cavities by using soil properties," *J. Jpn. Soc. Civil Eng.*, vol. 78, no. 1, pp. 70–85, 2022.

[11] S. M. Lee and H. M. Yoon, "A study for improvement of policy on ground subsidence prevention in urban areas," *Seoul Stud.*, vol. 18, no. 1, pp. 27–42, 2017.

[12] M. Tran, M. Elsisi, K. Mahmoud, M. Liu, M. Lehtonen, and M. M. F. Darwish, "Experimental setup for online fault diagnosis of induction machines via promising IoT and machine learning: Towards Industry 4.0 empowerment," *IEEE Access*, vol. 9, pp. 115429–115441, 2021.

[13] M. Tran, M. Elsisi, M. Liu, V. Q. Vu, K. Mahmoud, M. M. F. Darwish, A. Y. Abdelaziz, and M. Lehtonen, "Reliable deep learning and IoT-based monitoring system for secure computer numerical control machines against cyber-attacks with experimental verification," *IEEE Access*, vol. 10, pp. 23186–23197, 2022.

[14] N. M. K. Abdel-Gawad, A. Z. El Dein, and M. Magdy, "Mitigation of induced voltages and AC corrosion effects on buried gas pipeline near to OHTL under normal and fault conditions," *Electric Power Syst. Res.*, vol. 127, pp. 297–306, Oct. 2015.

[15] Y. S. Jin, "The analysis on correlation of precipitation and risk factors to the soil subsidence," Ph.D. dissertation, Chonnam Nat. Univ., Gwangju, South Korea, 2018, pp. 104–105.

[16] K. Y. Kim, "Susceptibility model for sinkholes caused by damaged sewer pipes based on logistic regression," M.S. thesis, Seoul Nat. Univ., Seoul, South Korea, 2018.

[17] M. S. Han, "A risk assessment of ground subsidence by GPR and CCTV investigation," M.S. thesis, Seoul National Univ. Sci. Technol., Seoul, South Korea, 2017.

[18] S. Lee, J. Kim, J. Kang, and W. Baek, "Comparison of machine learning models to predict the occurrence of ground subsidence according to the characteristics of sewer," *J. Korean Geo-Environ. Soc.*, vol. 23, no. 4, pp. 5–10, 2022.

[19] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[20] A. G. Asuero, A. Sayago, and A. González, "The correlation coefficient: An overview," *Crit. Rev. Anal. Chem.*, vol. 36, no. 1, pp. 41–59, 2006.

[21] H. Xu and Y. Deng, "Dependent evidence combination based on Shearman coefficient and Pearson coefficient," *IEEE Access*, vol. 6, pp. 11634–11640, 2018.

[22] U. Wählby, E. N. Jonsson, and M. O. Karlsson, "Comparison of stepwise covariate model building strategies in population pharmacokinetic-pharmacodynamic analysis," *AAPS PharmSci*, vol. 4, no. 4, pp. 68–79, Dec. 2002.

[23] L. Breiman, "Random forests," *Mach. Learn.*, vol. 5, pp. 5–32, Oct. 2001.

[24] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, Jan. 2005.

[25] G. Louppe, "Understanding random forests," Univ. Liege, Leige, Belgium, Tech. Rep., 2014, p. 211.

[26] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. Berlin, Germany: Springer, 2009, p. 745.

[27] E. J. Park, J. H. Park, and H. H. Kim, "Mapping species-specific optimal plantation sites using random forest in Gyeongsangnam-do province, South Korea," *J. Agricult. Life Sci.*, vol. 53, no. 6, pp. 65–74, 2019.

[28] S. H. Lee, Y. A. Yoon, J. H. Jung, H. S. Sim, T. W. Chang, and Y. S. Kim, "A machine learning model for predicting silica concentrations through time series analysis of mining data," *J. Korean Soc. Quality Manag.*, vol. 48, no. 3, pp. 511–520, 2020.

[29] M. Masud, A. K. Bairagi, A. A. Nahid, N. Sikder, S. Rubaiee, A. Ahmed, and D. Anand, "A pneumonia diagnosis scheme based on hybrid features extracted from chest radiographs using an ensemble learning algorithm," *J. Healthcare Eng.*, vol. 2021, p. 11, Feb. 2021.

[30] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.

[31] Y. Zhang and A. Haghani, "A gradient boosting method to improve travel time prediction," *Transp. Res. C, Emerg. Technol.*, vol. 58, pp. 308–324, Sep. 2015.

[32] D. Zhang, H. D. Chen, H. Zulfiqar, S. S. Yuan, Q. L. Huang, Z. Y. Zhangand, and K. J. Deng, "iBLP: An XGBoost-based predictor for identifying bioluminescent proteins," *Comput. Math. Methods Med.*, vol. 2021, p. 15, Jan. 2021.

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

[34] K. Guolin, M. Qi, F. Thomas, W. Taifeng, C. Wei, M. Weidong, Y. Qiwei, and L. Tie-Yan, "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 3149–3157.

[35] J. Lv, C. Wang, W. Gao, and Q. Zhao, "An economic forecasting method based on the LightGBM-optimized LSTM and time-series model," *Comput. Intell. Neurosci.*, vol. 2021, p. 10, Sep. 2021.

[36] M. Sokolova, N. Japkowicz, and S. Szpakowicz, "Beyond accuracy, F-score and ROC: A family of discriminant measures for performance evaluation," in *Advances in Artificial Intelligence* (Lecture Notes in Computer Science), vol. 4304. Heidelberg, Germany: Springer, 2006, pp. 1015–1021.

[37] L. Wang, F. Chu, and W. Xie, "Accurate cancer classification using expressions of very few genes," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 4, no. 1, pp. 40–53, Jan. 2007.

[38] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.
[39] Q. Gu, L. Zhu, and Z. Cai, "Evaluation measures of the classification performance of imbalanced data sets," in *Proc. 4th Int. Symp. Comput. Intell. Intell. Syst., Commun. Comput. Inf. Sci.*, vol. 51. Heidelberg, Germany: Springer, 2009, pp. 461–471.
[40] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation measures for models assessment over imbalanced data sets," *J. Inf. Eng. Appl.*, vol. 3, no. 10, pp. 27–38, 2013.

SUNGYEOL LEE was born in Incheon, South Korea, in 1992. He received the Ph.D. degree in civil engineering from Chonnam National University, Gwangju, South Korea, in 2022. Since 2022, he has been a Postdoctoral Researcher with the Korea Institute of Civil Engineering and Building Technology. His research interests include ground subsidence risk analysis using machine learning and statistical.

JAEMO KANG was born in Busan, South Korea, in 1974. He received the Ph.D. degree in civil engineering from Hanyang University, Seoul, South Korea, in 2017. Since 2002, he has been a Senior Researcher with the Korea Institute of Civil Engineering and Building Technology. His research interests include underground spatial information, underground utilities diagnosis, and ground subsidence risk analysis.

JINYOUNG KIM was born in Gwangju, South Korea, in 1987. She received the Ph.D. degree in civil engineering from Chonnam National University, Gwangju, in 2014. From 2015 to 2017, she was a Postdoctoral Researcher with the Korea Institute of Civil Engineering and Building Technology. Since 2018, she has been a Senior Researcher with the Korea Institute of Civil Engineering and Building Technology. Her research interests include construction automation and ground subsidence risk analysis.