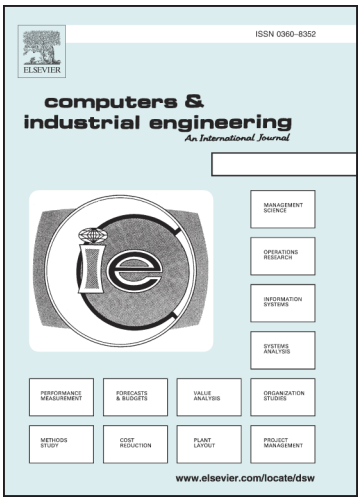
Journal Pre-proofs



Deep Learning-based Sewer Defect Classification for Highly Imbalanced Dataset

L. Minh Dang, SeonJae Kyeong, Yanfen Li, Hanxiang Wang, Tan N. Nguyen, Hyeonjoon Moon

PII: S0360-8352(21)00534-9
DOI: https://doi.org/10.1016/j.cie.2021.107630
Reference: CAIE 107630

To appear in: *Computers & Industrial Engineering*

Received Date: 23 January 2021
Revised Date: 5 June 2021
Accepted Date: 18 August 2021

Please cite this article as: Minh Dang, L., Kyeong, S., Li, Y., Wang, H., Nguyen, T.N., Moon, H., Deep Learning-based Sewer Defect Classification for Highly Imbalanced Dataset, *Computers & Industrial Engineering* (2021), doi: https://doi.org/10.1016/j.cie.2021.107630

Deep Learning-based Sewer Defect Classification for Highly Imbalanced Dataset

Deep Learning-based Sewer Defect Classification for Highly Imbalanced Dataset

L. Minh Dang[1], SeonJae Kyeong[1], Yanfen Li[1], Hanxiang Wang[1], Tan N. Nguyen[2], Hyeonjoon Moon[1]*

[1] Department of Computer Science and Engineering, Sejong University, Seoul, Republic of Korea

[2] Department of Architectural Engineering, Sejong University, 209 Neungdong-ro, Gwangjin-gu, Seoul 05006, Republic of Korea

Corresponding Author

Hyeonjoon Moon, Professor

Department of Computer Science and Engineering, Sejong University,

Seoul, Republic of Korea

Tel: +82-2-3408-4068

Email: hmoon@sejong.ac.kr

Credit author statement

L. Minh Dang: Conceptualization, Writing - Original Draft. SeonJae Kyeong: Data curation, Visualization. Yanfen Li: Data Curation, Writing - Review & Editing. Hanxiang Wang: Software. Tan N. Nguyen: Validation. Hyeonjoon Moon: Supervision, Funding acquisition.

Highlights

- A manually collected sewer defect dataset that contains over 38,000 images.

- An efficient deep learning-based sewer defect classification framework.
- Effectively deal with the imbalanced data problem using various approaches.
- Subtitle recognition that gives more information about detected defects.
- A novel frame reduction algorithm that significantly reduces the computational time.

Abstract

Sanitary sewer systems play a fundamental role in protecting water quality and the public well-being. Structural, civil, and functional operations of any sewer network can deteriorate at accelerated levels due to harsh environments inside the sewer pipes. The existing maintenance procedures are usually deemed inefficient in terms of the assessment accuracy, reliability, safety, and the cost due to the difficulty of detecting and diagnosing defects inside the sewer network. As a result, this paper proposes a robust and efficient deep learning-based framework that can detect and evaluate the defects automatically with high accuracy. The main contributions of the work include (1) a fine-tuned deep learning-based sewer defect detection framework that is based on the block-based architecture, which contains a series of convolutional layers that can efficiently extract the abstract features from the defective regions, (2) hybrid extensions of the proposed model that apply the ensemble-based approach and the cost-sensitive learning-based method in order to cope with the imbalanced data problem (IDP) efficiently, and (3) a novel frame reduction algorithm that is based on analyzing the contextual information of the closed-circuit television (CCTV) videos. The experimental results indicated that the proposed framework obtained a state-of-the-art performance compared to the previous sewer defect detection systems, and it was robust against the IDP. The benefits of the proposed defect detection framework are that it motivates more efficient defect analysis algorithms and promotes a complete integration of deep learning-based approaches in real-world sewer defect analysis applications.

Keywords: Sewer network; crack classification; deep learning; CCTV; text recognition; imbalanced data;

1. Introduction

Infrastructure is a fundamental factor that can stimulate the economic development of every community because it connects supply chains, brings new opportunities to struggling communities, and defends the nation against an increasingly unpredictable natural environment. Moreover, it stimulates the economy by providing millions of jobs in construction and maintenance yearly [1]. Modern concrete structures, such as sewer pipelines, require a substantial financial investment, a carefully planned blueprint, lengthy construction time, and operational issues [2]. Even though these structures can be used for prolonged periods if they are adequately maintained, manual maintenance procedures are often considered ineffective in terms of cost, safety, assessment accuracy, and reliability [3]. In addition, the structural, civil, and functional systems of these concrete structures can deteriorate rapidly due

to harsh environmental conditions. A postponement with identifying and analyzing concrete structures can bring about sudden structural and functional failures that could leak harmful substances into the environment and demand high rehabilitation costs. As a result, they must always be maintained in the best manner in order to mitigate the loss of life during natural disasters, such as earthquakes, floods, or criminal acts [4, 5].

Manual inspection is the primary way of periodically evaluating the structural and functional requirements in order to guarantee that it meets the basic service specifications. However, it is labor-intensive, time-consuming, and strenuous, because the structural inspection companies have to hire professional inspectors to manually perform the structural inspections using various equipment [2]. Robots and scanning devices [6] have been increasingly used in recent years to inspect and maintain concrete structures in order to reduce maintenance costs and improve the effectiveness of an automated inspection. Thus, there is an enormous contest among the top industrial robot manufacturers to create a better line of robots, such as the latest utility hole inspection vehicle from the Electronics and Telecommunication Research Institute (ETRI), South Korea, which supports a 3D point cloud and a 360-degree field of view for accurate utility hole analysis that can go up to 50 feet in sanitary sewers. Moreover, the high-resolution vision sensor that is attached to the robot's head enables it to precisely record the inner conditions in medium and large sewer pipes. Closed-circuit television (CCTV) recorded by the robots is a cost-effective and suitable means to monitor the pipe's condition in an unsanitary environment or complex surveillance circumstances where humans cannot reach.

A training dataset is then generated, which is based on the collected CCTV videos. For the sewer fault classification problem, the datasets contain different types of defects, and each type contains numerous images that are extracted from the videos and validated manually. However, the collected datasets usually suffer from the imbalanced data problem (IDP), which refers to a dataset where the samples between the classes are not represented evenly [7]. Three groups of regularly used solutions to deal with the problem are the data-based approaches, the ensemble-based methods, and the cost-sensitive learning-based methods [8]. However, only the ensemble-based approach and the cost-sensitive learning-based approach are implemented in this study. The data-based approach, which includes the over-sampling and the under-sampling techniques, is performed by removing or adding the duplicate samples from the original dataset, which can significantly influence the system’s performance [9, 10].

Due to the massive amount of collected CCTV videos, it is crucial to implement an AI-powered crack detection framework in order to identify the cracks and extract the contextual

information about them automatically [7]. The traditional computer vision (CV) approaches achieved poor performance and could only be applied to the small datasets, because they required the manual selection and extraction of the distinctive defect features [6]. The deep learning-based defect detection systems have been proved to improve the overall performance and save a considerable amount of time and effort compared to the conventional machine learning (ML) techniques [1, 2]. Moreover, they have also motivated a consistent improvement of the structural inspection technologies and proactive asset management strategies. The existing literature surveys revealed that the computationally intensive algorithms usually achieved a high classification performance [5], whereas the customized classification methods compromised the false-positive rates and accuracy [1]. Furthermore, the defect classification for concrete structures is a challenging subject that depends on many factors, such as the input quality, lighting environment, and background noise. Although various structural assessment studies have been introduced, they did not achieve convincing results and contained a limited number of the defect types [2, 4].

The printed subtitles on each frame of a CCTV video provide the contextual information about the fault inside a sewer pipe, such as the position, date, time, pipe diameter, and pipe type. This information is automatically recognized based on the text recognition frameworks, so it can be used later to provide in-depth details about a defect. A standard text recognition system involves three stages, which include text localization, enhancement, and text recognition. Localization is implemented in order to identify the appearance of the text. Common text features, such as intensity, color, and geometry, are usually applied to perform localization. Text localization is followed by text enhancement methods in order to increase the image quality, which improves the text and the background contrast. Lastly, optical character recognition (OCR) models are applied to perform text recognition [11].

Based on reviewing the different aspects of a sewer defect detection system, a pressing need is needed in order to promote a robust defect detection framework in sewer pipelines that can efficiently identify the different types of defects, extract textual information, and are resilient to the IDP. This manuscript proposes a deep convolutional neural network (CNN)-based sewer defect detection framework to detect the 7 types of defects, including crack, debris silty, faulty joint, open joint, protruding lateral, surface damage, broken pipe, and a normal class without the defects as illustrated in Fig. 1 in order to automatically extract the defect's abstract features and deliver a state-of-the-art defect classification accuracy for CCTV videos. The introduced framework is motivated by the recent successes of the CNN-based models on numerous applications that showed its ability to extract multiple abstract features from the

training datasets [1, 2]. After that, three hybrid models that apply different algorithms into the existing framework are also introduced in order to solve the IDP. Lastly, various tests are performed in order to confirm the effectiveness of the defect detection framework and the hybrid models with different dataset settings. In the initial test, the performance of the framework is examined on a balanced dataset and compared with the existing sewer faulty classification models. After that, the second experiment is conducted in order to verify the decisions that are made by the proposed model, which is based on the explainable artificial intelligence (XAI) methods. Three hybrid models are then tested on different imbalanced data settings that range from 1:1 to 1:100. Finally, the system is applied to generate the defect reports for the real CCTV videos, and the generated reports are then compared with the ground truth reports that are created by the inspectors.

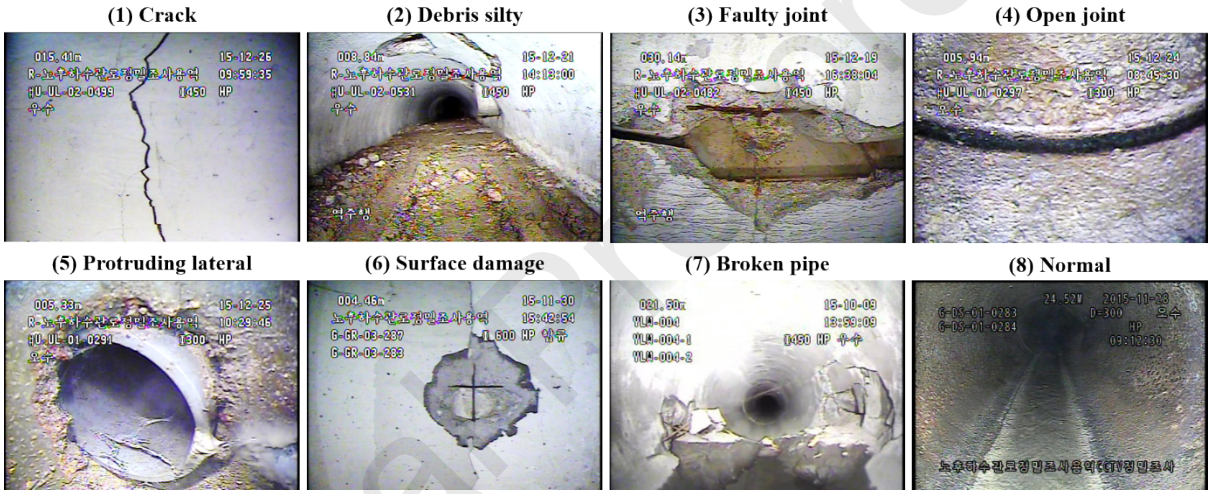


Figure 1. Sample images of eight classes of the sewer dataset.

The proposed framework contributes to the advancement of sewer defect analysis in many ways. Firstly, it offers a deep learning-based defect detection framework that can be applied to detect various types of sewer defects. Secondly, this paper evaluates a range of methods that help the proposed model achieve a robust performance in extremely imbalanced data settings, which is the problem that usually occurs with structure defect detection projects. Thirdly, the contextual information of a defect, which was ignored in the previous research, is also recognized using a text recognition module. This data is useful in order to provide in-depth details about the defect. Fourthly, a novel frame reduction algorithm, which is based on the recognized contextual data, is presented in order to reduce the number of CCTV frames that need to be processed. Finally, some XAI techniques have been implemented in order to gain insight into how the proposed model was trained. These contributions collectively generate a

technical framework that is required to achieve a precise and efficient sewer defect detection framework using CCTV videos.

The manuscript is divided as follows. Section 2 thoroughly summarizes the previous approaches about sewer defect detection, the IDP, and the XAI. The sewer defect detection framework is described in Section 3. Next, the collected sewer defect image dataset is introduced in Section 4. In Section 5, multiple experiments are conducted in order to examine the robustness of the defect detection framework on both the imbalanced dataset and the balanced dataset. Finally, we conclude the research in Section 6 by analyzing the strengths and the weaknesses of the proposed model and introduce the future work.

2. Related work

2.1. Sewer defect classification

Due to technological advancements, especially in CV, there is an ever-increasing amount of vision-based sewer defect classification research [12]. For example, Myrans et al. proposed an automatic method that recognized various types of defects in CCTV videos [13]. Firstly, a feature descriptor for each frame from a video was computed. After that, two ML algorithms were then implemented in order to analyze the content from the individual frames. The Hidden Markov Model and the filtering method were applied to extract information from a sequence of frames in order to smooth the prediction. The experimental results on a dataset that was collected by Wessex Water showed that the model obtained a detection accuracy of over 80%. Moreover, the smoothing technique on the sequence of frames decreased the false-negative rate and considerably increased the performance. However, the dataset used in this research is small, which contained only 1000 images, and over half of the dataset, which totaled 623 images, belonged to the *no faults* class. Ye et al. introduced a sewer faults recognition approach based on feature extraction and an ML algorithm [14]. Various features, such as texture features, Hu invariant moment, Daubechies (DBn) wavelet transform, and lateral Fourier transform, were extracted from the defect regions. After that, these features were used to train a support vector machine (SVM) model in order to categorize the seven types of sewer pipe defects. The model performance on 28,760 m of sewer pipes reached 84.1%. However, the proposed model performed poorly with the two classes, which included collapse and joint damage, because it suffered from the IDP. Fang et al. introduced a defect identification framework, which used an unsupervised ML-based fault detection algorithm on CCTV footage [15]. Moreover, the authors obtained the related features from a collection of images and

combined the extracted features in order to increase the accuracy. The evaluations were conducted on small and big image sequences, which produced the highest recorded accuracy above 90%. Even though traditional CV methods have been extensively used for automated defect detection in CCTV footage, the traditional CV algorithms rely heavily on the pre-processing methods and the appropriate feature extraction for certain cases, which is both error-prone and labor-intensive. In addition, these models are usually examined in artificial testing setups that lead to biased results and poor performances on real-world data.

In the recent decades, deep learning, which is a subset of ML that applies multi-layered artificial neural networks, has proved to deliver state-of-the-art performances in different fields, which include object detection, object classification, natural language process, and many others. Deep models effectively learn abstract features from training data without human invention, and conventional image processing techniques are not compulsory. Therefore, deep learning-based models have been applied extensively to the defect classification for public infrastructures during the last few years. For example, Hassan et al. introduced a CNN-based sewer crack identification framework on images that were extracted from CCTV videos as the input [2], which was based on performing transform learning of the AlexNet model. The proposed model can recognize the 6 main types of cracks with an average accuracy of 96.33% on the testing dataset. However, the model's performance was influenced by the imbalanced distribution of the images between the classes. In another study, Cheng et al. performed automatic sewer crack detection using a faster region-based convolutional neural network (Faster R-CNN) approach [1]. Many hyperparameters were tailored in order to examine the most influential factors with the proposed model's performance. Several experiments were implemented in order to evaluate the model's performance, which include the accuracy and the computational costs. The change of parameters, such as the stride values and the filter dimensions, contributed to the high detection accuracy, which led to a mean average precision (mAP) of 83%. The proposed framework can only be applied to static images, so interpreting the videos should be studied. Xie et al. applied a two-level hierarchical deep CNN model in order to automatically extract the representative features for sewer defect identification [5]. The framework proved to generalize the new data well and solved the difficulties of the IDP through several experiments, which obtained a classification accuracy that was over 94% on multiple benchmark datasets. However, the framework required high computational costs and failed to point out the correlations, such as crack defects and deformations, which usually occur together. Meijer et al. introduced a self-collected sewer defect dataset from CCTV footage with over 21 thousand images [4]. The authors then fed the collected dataset, which was based on a

deep learning approach, in order to perform the defect detection. In addition, a multiclass classification was applied to detect multiple defect types in a single image. They also proposed a *leave-two-inspections-out* cross-validation approach that effectively eliminated a data leakage bias. However, the suggested classifier did not achieve the standard performance for real crack classification systems.

Even though many studies applied deep learning in order to perform the sewer defect detection, the subtitles printed on the sewer CCTV videos were not considered, and the datasets used for training were small. This manuscript addresses the mentioned problems by introducing a huge sewer defect dataset, analyzing the subtitles in order to perform an in-depth analysis of the detected defects, and proposing a frame reduction algorithm to reduce the number of frames that need to be processed.

2.2. Imbalanced data problem (IDP)

Sewer images that contain defect region(s) are minor compared to non-defect images during the data collection process, which was previously described. As a result, the previous sewer defect detection frameworks usually suffer from the IDP [16]. There are three main approaches that are usually used to solve the IDP, which include the resampling approach, cost-sensitive learning approach, and ensemble learning approach. The resampling approach is usually applied to the tabular data by removing or adding the samples in order to balance the samples between the classes, which is inappropriate for the sewer defect dataset used in this study [9]. On the other hand, the cost-sensitive learning approach is introduced to solely solve the IDP, whereas the ensemble learning method combines the results of multiple base classifiers to enhance the model generalization, which can also be used to deal with the IDP [8, 9]. Therefore, this work integrates the cost-sensitive learning technique and the ensemble learning technique into the deep learning model in order to create different hybrid extensions of the model to cope with the IDP.

Ensemble learning approaches use several techniques and learning algorithms in order to acquire a higher system's performance than the performance of any of the learning algorithms by themselves [17]. The main idea is to incorporate a group of weak learners in order to form a better classifier, which consequently improves the system's performance and robustness. Bagging and boosting are two main techniques that are used in ensemble learning [18, 19]. Bagging approaches generate multiple new training subsets by randomly taking with replacement from the original dataset. In the boosting methods, the learners are trained consecutively with the initial learners, which apply simple models to the data, and then probe

the data for mistakes. Following trees are fit, and at every step, the intention is to improve the performance of the previous learner. In the case of the defect detection framework, only the boosting technique is examined, because the final learner has lower errors as it improves the performance and reduces the pitfalls of using only one model. In addition, bagging has demonstrated that it rarely achieved a better bias when a single model has a low performance.

Extreme gradient boosting (XGBoost) [20] and a light gradient boosting machine (LightGBM) [21] are popular extensions of the gradient boosting technique, which is popular for its speed and robustness. They have shown excellent results with several classification algorithms [22]. They both implement the left-wise growth strategy when growing the tree in order to obtain the best tree and prevent the possible loss of information, which is a problem that remains in the gradient boosted tree approaches. However, the main difference is that while LightGBM applies a method call Gradient-based One-Side Sampling (GOSS) in order to analyze and lower the data instances, which helps to figure out a suitable split value, XGBoost relies mainly on pre-sort-based and histogram-based algorithms in order to estimate the most appropriate split. In contrast, the cost-sensitive learning approach attempts to solve the problem by assigning different costs for each class [8]. Conventional ML algorithms consider that each class has equivalent misclassification costs, which leads to a significant drop in their performance when facing the IDP [2, 4].

2.3. Explainable artificial intelligence (XAI)

Explainable AI or XAI refers to a collection of methods and approaches that help to interpret decisions that are made by the AI models and make them more comprehensible to the users and the researchers. These methods aim at solving the current *black box* nature of AI algorithms, because AI algorithms cannot explain what features lead to a decision. Moreover, XAI presents the reasons for decision-making in a way that people can understand, so it can be applied to areas that require transparency and user trust, such as medical [23], banking [24], and law [25]. Previous work on the sewer defect classification ignored the importance of XAI that can explain the results, which were predicted by the AI models. As a result, this study attempts to implement the Class Activation Map (CAM) and the layer activation visualization methods in order to interpret and explain the classification model's predictions. Both methods are based on analyzing and visualizing the feature maps of a specific convolutional layer in order to enable the interpretation and explanation of how a CNN model learns the notable characteristics from each type of the defect.

CAM results from performing Global Average Pooling (GAP) in a feature map that is extracted from the final convolution layer and putting the value in Softmax in order to get the probability value. A paper proposed by Dang et al. proposes a manipulated face image detection framework that is based on deep learning [22]. CAM is used in the paper in order to explain the decisions that are made by the proposed deep learning model by transferring the weights of the Softmax layer back to the feature maps of the fifth convolutional layer, which thereby emphasizes the significant parts in the image that affect the prediction. The manipulated regions were marked correctly in the CAM output, which showed that the framework categorized the tampered images by detecting the tampered regions. On the other hand, layer activation visualization is an XAI technique that is implemented in order to visualize how the convolutional layer extracts the significant features of an image through the filter, which is the feature maps. For example, Xie et al. introduced the automated identification and classification of sewer cracks using a hierarchical deep learning approach [5]. The former layers output relatively fine detail from an image through the feature map visualization method. For example, some feature maps from the Conv0 layer highlight features, such as color, shape, background, and foreground from the sewer pipe images. On the other hand, the feature maps show less detail as the network gets deeper.

3. Methodology

Fig. 2 shows a comprehensive structure of the suggested sewer defect detection framework with four key modules: data collection, model training, subtitle recognition, and report generation. First of all, the proposed sewer defect dataset is generated based on investigating the frames that were extracted from the CCTV videos and filtering out the images that contain defects. After that, they are classified into one of the seven types of defects and the normal class. Prior to the training process, the data augmentation and preprocessing techniques are implemented in order to refine the training dataset. Next, the dataset is used to train two different models. One model is used for the defect classification, and the other model is used for the subtitle recognition. For the defect classification model, the abstract features are extracted by a series of convolutional layers. The final output layer eventually uses the extracted features in order to decide the output label, which is one of the eight classes, for an input image. In addition, the IDP is also addressed by applying two different IDP techniques to the proposed model. These models will be examined using different imbalanced data settings in Section 5.4.

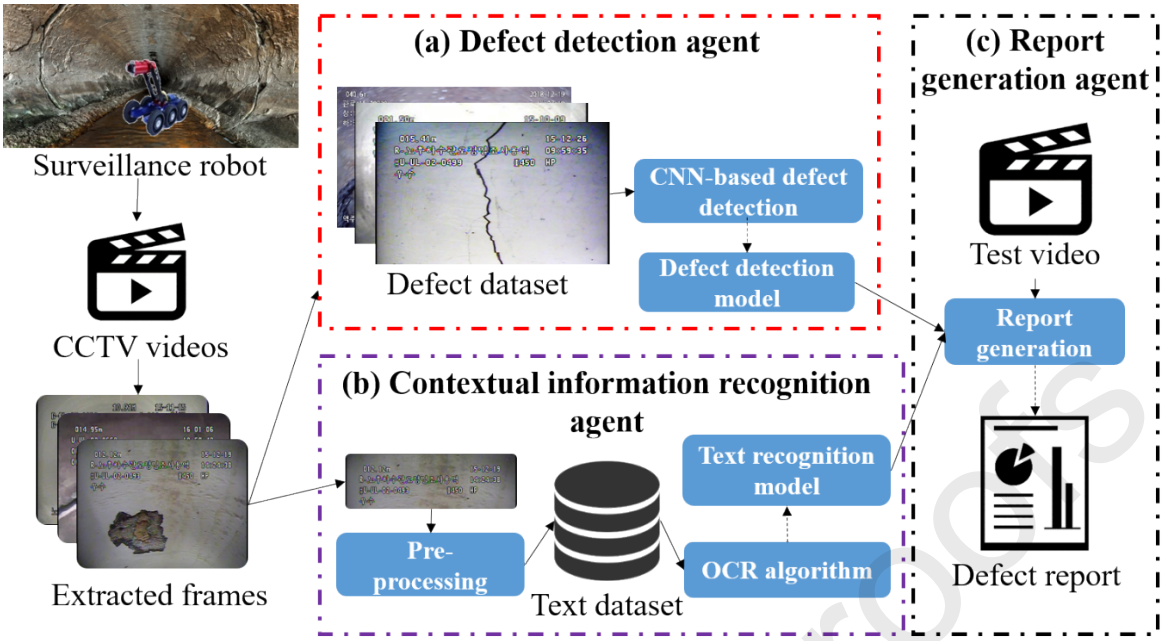


Figure 2. Detailed visualization from input to output of the sewer defect detection framework.

3.1. Defect detection agent

3.1.1. Fine-tuned sewer defect classification model

The successful use of CNNs for various tasks motivated us to apply them for the crack classification application [27, 28]. VGG is one of the most commonly used deep CNNs for the image classification task, which obtained the highest performance for the ILSVRC-2014 [28]. As a result, a model that is based on the VGG19's architecture with some essential adjustments regarding the previously mentioned defect detection topic was implemented.

All the layers from the pre-trained VGG-19 were initially initialized. However, the defect dataset is unassociated with the benchmark ImageNet dataset, so the top 17 convolutional layers are frozen to act as an abstract features extractor. The blue box in Fig. 3 indicates the customization for the sewer defect detection model during the fine-tuning phase. The model accepts color images of size $224 * 224 * 3$ as the input. All the layers are kept similar to the original VGG-19 model except for two convolution layers with a kernel size of $3 * 3$, which were placed after the convolution layer that belongs to the fourth convolution block and before the pooling layer. Thus, the number of convolutional layers increases from 4 to 6 in the convolutional block 5. A series of five convolutional blocks is followed by three fully-connected layers. The top two fully-connected layers contain 4096 neurons each, whereas the number of output neurons in the final fully-connected layer is decreased from 1000 to 8 in order to fit the total number of 8 classes in this manuscript. Batch normalization is a standard approach that is used to optimize the deep learning models by unifying the scattered data,

avoiding the vanishing gradient, and increasing the network's robustness. Several research papers have demonstrated that batch normalization remarkably stabilized the training process and reduced the number of training epochs [11]. As a result, a batch normalization layer is put after the first two fully-connected layers.

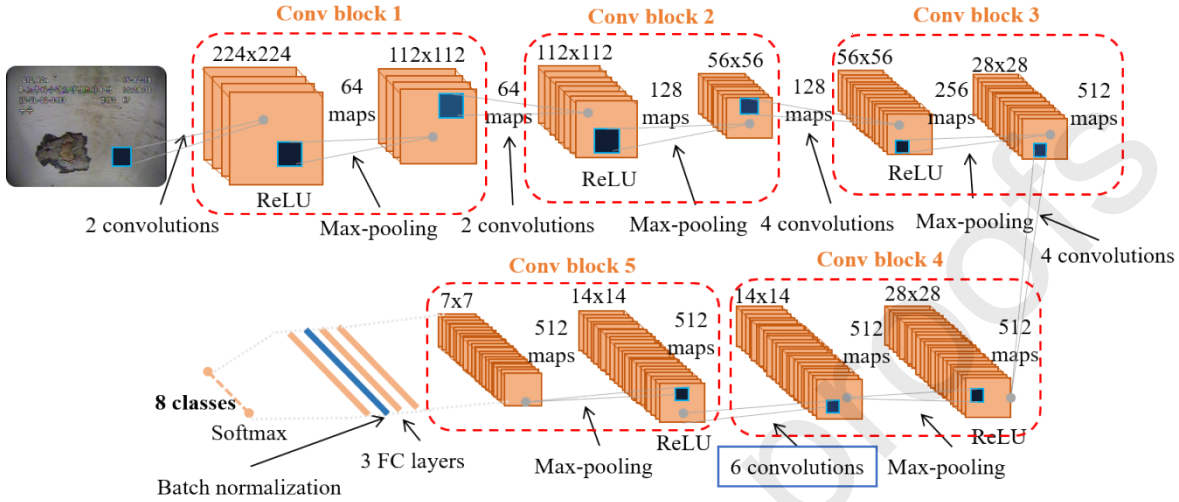


Figure 3. A fine-tuned VGG19 structure with five convolutional blocks for the sewer defect detection.
Note. Two more convolutional layers (blue box) are added to the fourth convolutional block. Additionally, a batch normalization layer (blue line) is put between the second and the third fully-connected layers.

3.1.2. Hybrid sewer defect detection model

After the initial experiments, the classification accuracy dropped significantly when the number of images was unevenly distributed between the classes. As a result, the previous sewer defect identification models experienced poor performances due to the IDP [17]. Therefore, this section applies two standard methods to cope with the IDP that include the cost-sensitive learning method and the ensemble learning method, which are shown in Fig. 4. Two algorithms, which include XGBoost and LightGBM, are implemented in the ensemble learning-based approach, whereas the misclassification cost customization is implemented in the cost-sensitive learning approach.

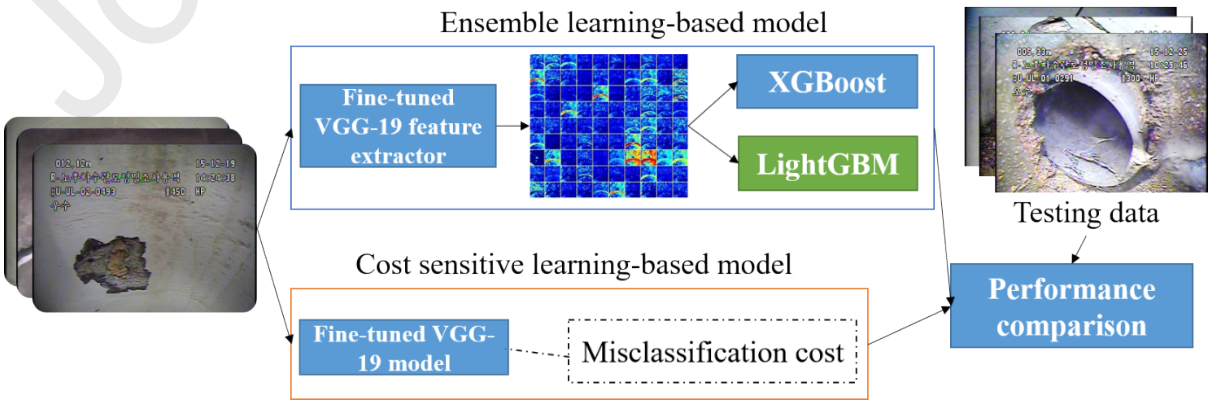


Figure 4. Two primary hybrid approaches that are applied to solve the IDP.

3.1.2.1. Extreme gradient boosting (XGBoost)

XGBoost is an efficient approach in order to force the algorithm to concentrate on the misclassification from the minority class when the imbalance data problem occurs. It includes the construction of a series of two-stage learners from the initial data and then combines their predictions [20]. The objective function of XGBoost at iteration t, which was described in [20], was defined by the equation below.

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (1)$$

where l is the loss function that estimates the difference between the ground truth y_i and the predicted $\hat{y}_i$. For the regression problem, l is usually the root-mean-square deviation (RMSE), and it is logloss or mlogloss for the classification problem. On the other hand, the Ω regularization parameter assists XGBoost in order to smooth the learned weights to control the overfitting. The regularization value is assigned by analyzing the number of samples and the prediction threshold of the samples.

3.1.2.2. Light gradient boosting machine (LightGBM)

LightGBM concentrates on reducing the model training time based on combining several learning methods [21]. The primary objective of this method is to implement the Exclusive Feature Bundling (EFB) and the Gradient-based One-Side Sampling (GOSS), which are two methods that are used in order to cope with the issue remaining in XGBoost, which scans the data repeatedly [15]. GOSS decreases the computing complexity by excluding data with small gradients that have a minor impact on the calculation of the information gain. The EFB method enables the dependent features to be combined in order to minimize the number of features. Moreover, LightGBM has a fast convergence rate using the weighted quantile sketch algorithm, a histogram-based splitting algorithm, and the leaf-wise tree growth strategy in order to manage big datasets.

3.1.2.3. Cost-sensitive learning

Cost-sensitive learning is another common approach that is used to solve the imbalanced data problem, which focuses on specifying the separate costs to the kinds of misclassification errors. After that, different methods are implemented in order to consider those costs. For instance, the cost matrix $cm_{c_1c_2}$ indicates the average cost of classifying an observation from class c_1 to class c_2. In the matrix, all the diagonal elements are equal to 0, which is an accurate

classification. The risk R for deciding output o_i for the input x is defined in [8], which is specified using the equation below.

$$R(o_i|x) = \sum_i cm_{c_1c_2} P(v_j|x) \quad (2)$$

The probability of selecting class i depends on the predetermined misclassification cost through the equation, and the posterior probabilities determine the uncertainty about the ground truth of x. The main objective of cost-sensitive learning is to minimize the misclassification cost by producing class v_j with the least risk R.

3.2. Text recognition and report generation

Subtitle information printed on the CCTV video contains critical information for an in-depth analysis of the detected sewer defects, which include the travel distance, the pipe ID, the sewer type, the diameter, the inspection date, and the inspection time. It provides useful information about the sewer pipe under investigation. The information can be recognized automatically using text detection and recognition algorithms in order to provide in-depth details of a defect. This section is dedicated to a fully automated process that supports operators in detecting, analyzing, and reporting all defects that occur inside a CCTV video in order to make a final report file.

3.2.1. Template matching

Template matching is a method in order to discover regions of an image that are similar to a template image. The similarity can be computed using a popular method that is called Normalized Correlation Coefficient matching proposed in [26], which is described in Equation 3.

$$\eta(x',y') = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}\{x'(i,j).y'(i.j)\}}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}\{x'(i,j)\}^2}\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}\{y'(i,j)\}^2}}\epsilon\,[0,1] \quad (3)$$

where x' is the original image of size $P * Q$, y' that refers to the template of size $M * N$, and $\eta(x',y')$ is the dividend of the fraction corresponds to the cross-correlation between the reference image and the original input. The sum-up is done for each image spot: $x' = 0..w-1; y' = 0..h-1$. The template is considered a match if the degree of match between the original image and the template is greater than the predefined threshold.

The template matching method is the translation invariant, which is sensitive to the image size and template. Thus, if the text size of the template is slightly different from the one in the original image, it is likely that the algorithm fails to match the text. A three-step multi-scale

approach is introduced to make the template matching method more robust to the image scaling changes.

- It loops over the source image at various scales by reducing the size gradually.
- It implements template matching on each of the scaled source images and keeps track of the match with the highest correlation coefficient.
- It selects a match with the most significant correlation coefficient and considers it as the *matched* region.

3.2.2. Text recognition module

The text recognition module that is utilized to perform the subtitle recognition in this study was introduced by Dang et al. [10], and it includes four main processes, which include the multi-frame integration, the preprocessing, the text detection, and the text recognition, as described in Fig. 5.

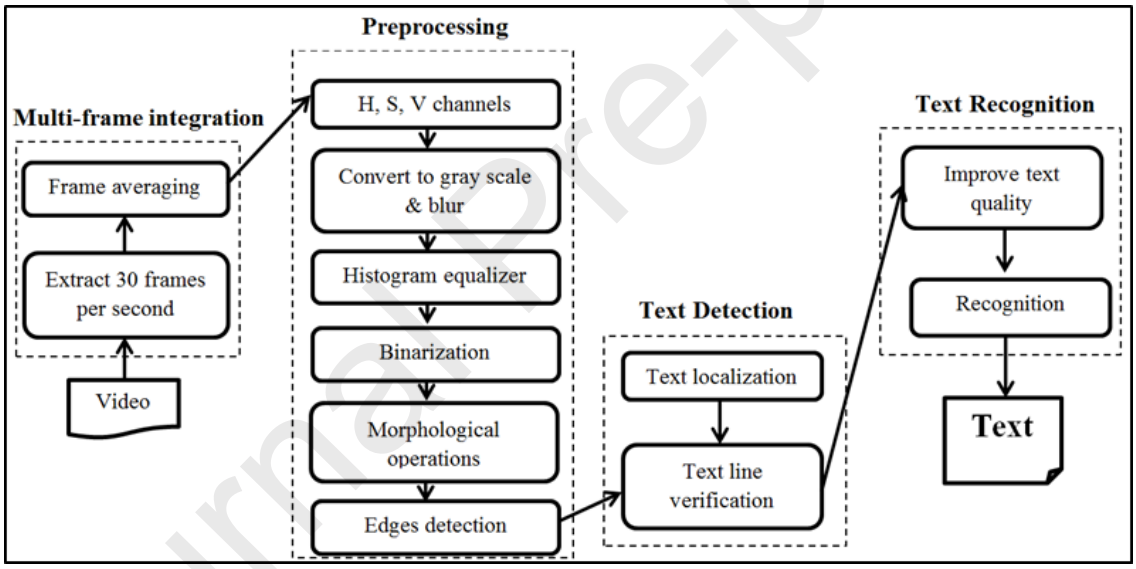


Figure 5. Overall architecture of the text detection and recognition module that was introduced by Dang et al. [10].

Firstly, the multi-frame integration process is implemented in order to enhance the contrast and simplify the background by computing an average image based on 30 consecutive frames (frame rate: 30 fps). For a series of frames C_i ($i \in [0,29]$), the output average image is created by using the equation that is given next.

$$AverageImage_i(x,y) = \underset{j \in C_i}{avg}\,(p_j(x,y)) \tag{4}$$

where $p_j(x,y)$ refers to the pixel value at location (x,y) of frame j.

The multi-frame integration process is followed by a series of image preprocessing methods, such as blurring, histogram equalizer, and morphological operations, in order to

reduce the possible noise. The third process involves two main sub-process that include text localization, which detects the subtitle information from the output image of the previous step, and text verification, which reduces noise and eliminates false alarms. Finally, the text recognition is trained based on the Tesseract OCR engine in order to recognize the detected text.

3.3. Report generation agent

3.3.1. Defect detection and extraction based on the nature of the CCTV subtitle

The operator's reaction can be identified and simulated when a defect is detected by analyzing distinct characteristics of the subtitles that were printed on the sewer CCTV video using the proposed text detection and recognition agent. The robot usually starts recording video before it is placed inside a sewer pipe, so parts of the CCTV recorded outside the sewer pipe can be discarded in order to reduce the processing time. The proposed template matching approach (Section 3.2.1) is useful to remove unnecessary frames by detecting the *Investigation starts* template, which is depicted in Fig. 8. If the *Investigation starts* template is seen in one frame, all the frames following that frame are extracted. On the other hand, all the frames are processed if the template is not found.

Even after conducting the frame reduction using the template matching approach, the sequence of the images that is fed into the frameworks is still huge. Therefore, one more method is proposed in order to reduce further the number of frames that are fed into the deep learning model. While the robot is being controlled to move inside the sewer pipeline in order to perform the investigation, if a sewer fault is identified by the streaming CCTV camera, the operator stops the robot at that position for about 3-8 seconds and utilizes the mounted camera to inspect the defect thoroughly, which is displayed in Fig. 6. As a result, Algorithm 1 is proposed in order to reduce the number of frames processed by the defect detection model, because the *travel distance* subtitle remains unchanged during the 3-8 seconds window.

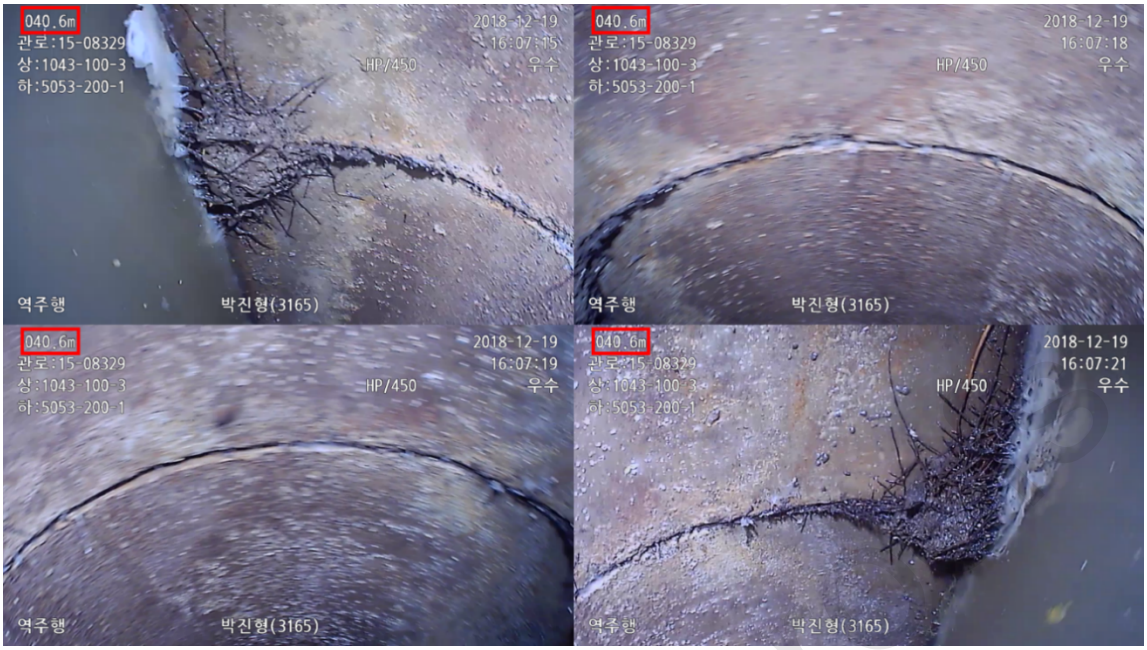


Figure 6. Four sample images of an open joint defect when the robot stops in order to investigate the defect for 6 seconds from 16:07:15 to 16:07:21.
Note. The red boxes indicate that the *Travel distance* subtitle of 040.6m remains unchanged.

Algorithm 1. *Defects detection using the travel distance subtitle*

For frame ***x*** **in** *extracted_frames* ***X*** **do**
    current_distance = recognize_distance(***x***)
    time = get_time(***x***)
    If current_distance **is** NULL **then**
        start_time = time
        end_time = start_time
        add current_distance, start_time, end_time **to** the temp_array
    Else
        end_time = time
        update end_time **in** the temp_array
    End If
    If (current_distance is different than (select current_distance from temp_array)) and ((end_time – start_time) > 3000) **then**
        extract all the frames that have the current_distance information
        current_distance = -1, temp_array = NULL
    End If
End For

Initially, the proposed text recognition module is applied to recognize the distance information for all frames that were extracted from a CCTV video. Suppose that the difference between the end time and start time of a specific distance is equal to or higher than 3000ms, which is 3 seconds, and a defect appears at this particular distance. Only the first frame is extracted to perform the defect classification in order to reduce computational complexity. Finally, the subtitle information that describes the crucial contextual information regarding the defect is recognized using the text recognition module.

3.3.2. Report generation module

This module's primary purpose is to support the operators in order to automatically analyze the sewer videos and generate a detailed report about the defects that appear in the video. This module fills the template, which is shown in Fig. 7, using two primary data sources that include the detected defect images using the proposed model, and the recognized contextual information for each detected defect using the text recognition module.

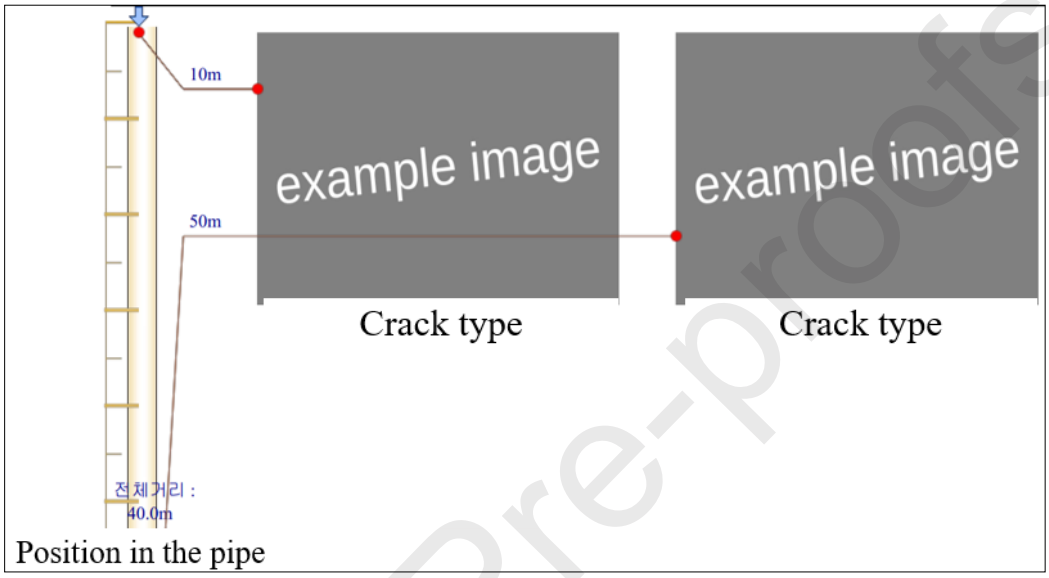


Figure 7. Report template that is created to report detailed information regarding the detected defects in a sewer CCTV video.

Note. The template contains all defects that appear in the video, their corresponding label, and the robot's travel distance information when a defect occurs.

4. Sewer defect dataset

There are 7733 sewer CCTV videos used in this study, which include lengths that range from 30 seconds to 15 minutes, and a video size of 1280x720. They were given to us from the Korea Institute of Civil Engineering and Building Technology. All the videos were captured by a commercial Robo Cam 6 (Tap Electronics Ind. Co., Ltd). The robot has a 1.3-megapixel Exmor CMOS sensor with the ability to perform full 360 rotation, and 240-degree side views up/down tilt. Moreover, six high-power led lamps, which were 35W, allow the robot to capture the videos in various environments. The inspectors controlled the robots in order to perform the investigation of several concrete sewer pipes across South Korea.

From the original collection of CCTV videos, all the frames were extracted and manually investigated in order to create a new sewer defect dataset, which included seven kinds of sewer defects and one normal class with a total of 38,386 images. The normal class indicates images that do not contain any defects. The number of images was expanded to 115,170 after the

implementation of three different data augmentation techniques, which included horizontal flip, shear range, and zoom range. Table 1 describes in detail the number of images for each class before and after applying the data augmentation process. Overall, an extreme class imbalance scenario can be observed between the normal class that contains 41,804 images and the crack class that has 7,142 images after implementing the data augmentation process.

Table 1. Number of images for each class before and after applying the image augmentation process.

Class	Before augmentation	After augmentation
Crack	2380	7142
Debris silty	4036	12,108
Faulty joint	5397	16,193
Open joint	3174	9524
Protruding lateral	3754	11,264
Surface damage	2542	7628
Pipe broken	3169	9507
Normal	13,934	41,804
Total	38,386	115,170

The subtitle information printed on any single frame of a CCTV video contains the crucial information for in-depth analysis, such as the travel distance, the pipe identification number (ID), the inspection date, the pipe diameter, and the sewer type, which is illustrated in Fig. 8.

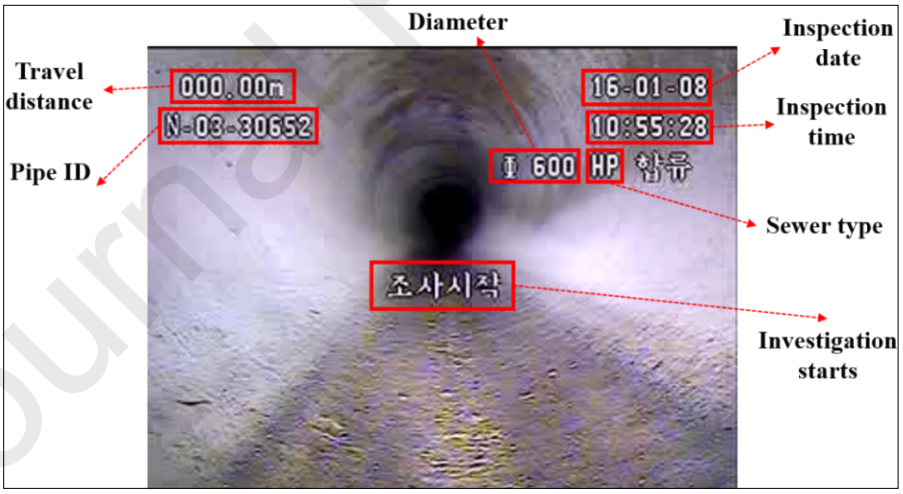


Figure 8. Essential subtitles printed on a sample frame that was extracted from a CCTV inspection video.

5. Experimental results

Extensive experiments are conducted in this section using the proposed dataset in order to show the model's effectiveness in identifying and analyzing the different sewer fault types. The first experiment, which is conducted in Section 5.1, was performed to examine the framework performance in the normal data scenario. Moreover, the proposed model was also

compared with two previous sewer faulty detection frameworks. Next, the model robustness against different attacks is examined in Section 5.2. After that, two XAI methods, which include layer activation visualization and class activation map, were implemented in order to interpret the proposed model, which is shown in Section 5.3. The fourth experiment, which is included in Section 5.4, then verifies the performances of the different hybrid models in the imbalanced dataset scenario. Finally, we also deploy the proposed model in a real-world application in order to detect sewer defects, which is described in Section 5.5.

5.1. Sewer fault classification on a balanced dataset

The model performance was evaluated in a balanced dataset setting and compared with the previous sewer defect detection models. From the augmented dataset, 56,000 images were selected randomly, which include 7000 images per class. The dataset was then separated into two parts that included the training sets, which contained 90% of the dataset or 50,400 images, and the remaining 10% of the dataset, which was used as the testing sets and contained 5600 images. The training dataset is further divided into two subsets that include the training subset, which contains 75% of the training dataset, and the validation subset, which contains 25% of the training dataset. The deep learning model was built and trained on Keras, which is a Python-based high-level open-source deep learning API. Adam optimization is used as the main optimization function in the model. At first, we set the learning rate equal to 0.001, and it is slowly minimized to 0.0001 depending on the validation error. The system was trained with 50 epochs and a batch size of 64. The total training time lasted for 1 hour 10 minutes. Fig. 9 presents the accuracy and loss of training and validation processes. The training accuracy and validation accuracy rise remarkably to above 90%, but the training loss and validation loss declines notably to below 0.25 after epoch 20. The accuracy and loss continue to improve gradually and become stable before stopping at 97.6% for the validation accuracy and 0.024 for validation loss at epoch 49th.

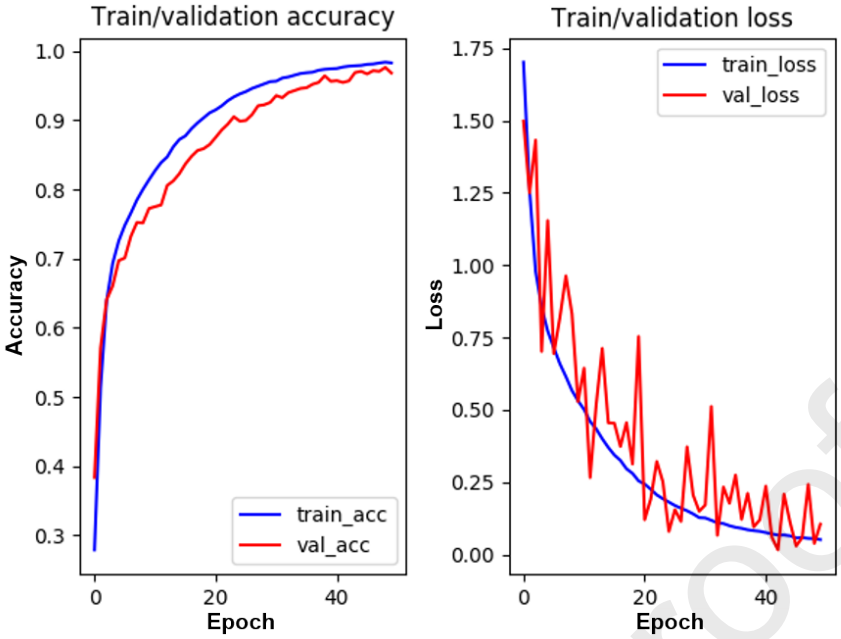


Figure 9. Training and validation results of the sewer defect classification framework.

Next, the model performance on the testing dataset that contains a total of 6300 images for eight classes, which include 700 images per class, was described in the confusion matrix in Table 2. The obtained accuracy demonstrates the robustness of the model, which could detect 7 classes of defects and the normal class with an average accuracy of 97.6%.

Table 2. Confusion matrix for the defect classification results on the testing set.
Note. Abbreviations for the table include crack (CC), debris silty (DS), faulty joint (FJ), open joint (OJ), protruding lateral (PL), surface damage (SD), broken pipe (BP), normal (NO), and accuracy (ACC).

Class	CC	DS	FJ	OJ	PL	SD	BP	NO	ACC (%)
CC	**666**	0	0	7	0	15	6	6	95.1
DS	0	**684**	3	5	2	0	1	5	97.7
FJ	11	1	**658**	23	0	0	5	2	94
OJ	5	3	8	**674**	0	2	0	8	96.2
PL	0	11	8	2	**673**	3	2	1	96.1
SD	17	0	25	0	1	**652**	0	5	93.1
BP	4	13	0	3	0	1	**677**	2	96.7
NO	5	3	6	3	0	2	1	**680**	97.1

The proposed framework was also compared with two previous deep learning-based defect classification models. The customized model proposed by Kumar et al. classifies whether an input sewer image is normal or contains defects [29]. It contains two convolutional layers with ELU activation and two fully connected layers. On the other hand, Perez et al. applied a pre-trained VGG-16 model in order to classify four types of sewer defects, which included mold, stains, deterioration, and normal [16]. These two models were implemented using the detailed description from the papers. The hyper-parameters were set similar to the settings of [16, 29]

except for the final Softmax layer was reconfigured to 8 for the performance comparison on the proposed sewer defect dataset. A performance comparison between the three approaches to the testing set is revealed in Table 3.

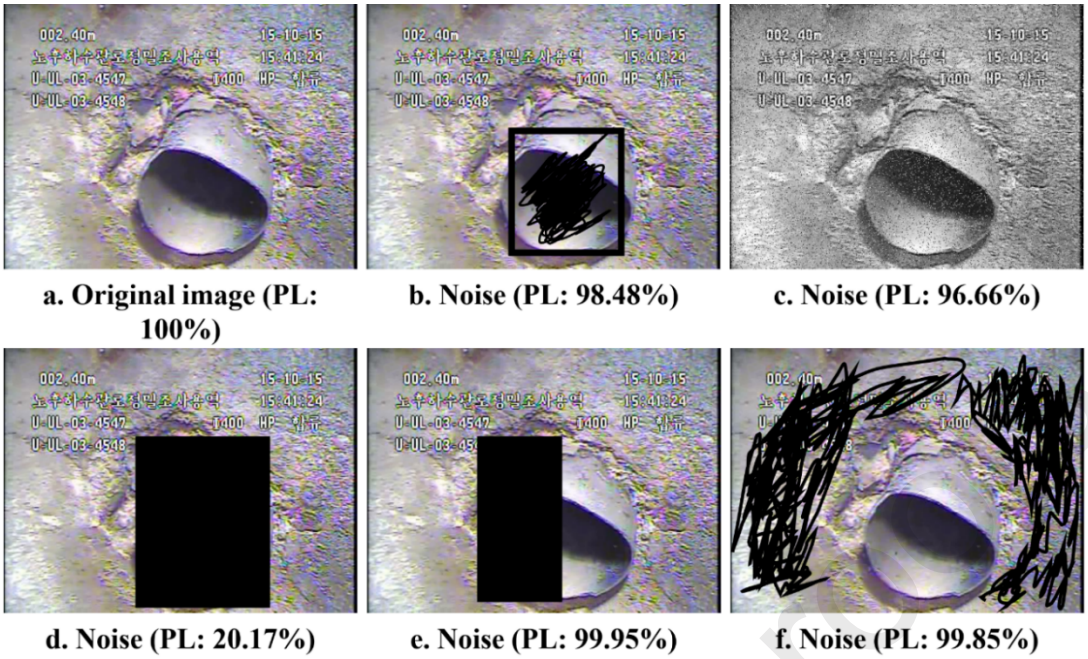
Table 3. The experimental results of the three different models on the collected sewer defect dataset.

Model	Input size	Number of convolutional layers	Testing time/image	Accuracy (%)
Fine-tuned model	(224, 224, 3)	19	0.15	**97.6**
Kumar et al. [1]	(256, 256, 3)	2	0.057	85.4
Perez et al. [2]	(224, 224, 3)	13	0.094	95.2

The fine-tune VGG-19 model yields the highest accuracy at 97.6%, whereas the model proposed by Perez achieves a 95.2% accuracy. Finally, the two convolutional layers' structure introduced by Kumar obtained the lowest accuracy of 85.4%. The results prove that the deeper the model, the better performance it gets.

5.2. Model robustness evaluation

In this section, we examine the proposed defect identification method robustness by applying several attacks, which include adding noise, cropping, and rotating. Fig. 10(a) displays an input protruding lateral (PL) defect, which was correctly predicted by the model as the PL with the confidence of 1. In Fig. 10(b) and Fig. 10(e), the major parts of the defect are obstructed, and the images are then fed into the proposed system. These two inputs were still predicted as the PL with an accuracy of 98.48% and 99.95%, respectively. Different types of noise were added to Fig. 10(c) and Fig. 10(f), but the model predicted them as the PL with a high accuracy of over 95%. In the last scenario, which is described in Fig. 10(d), the defect region was blocked and fed into the proposed framework. The model accurately predicted it as the normal class with the accuracy of the PL class being only 20.17%. Based on these results, the proposed method proved that it detected the defects under varying conditions despite the noisy images, which can occur in the real surveillance videos.

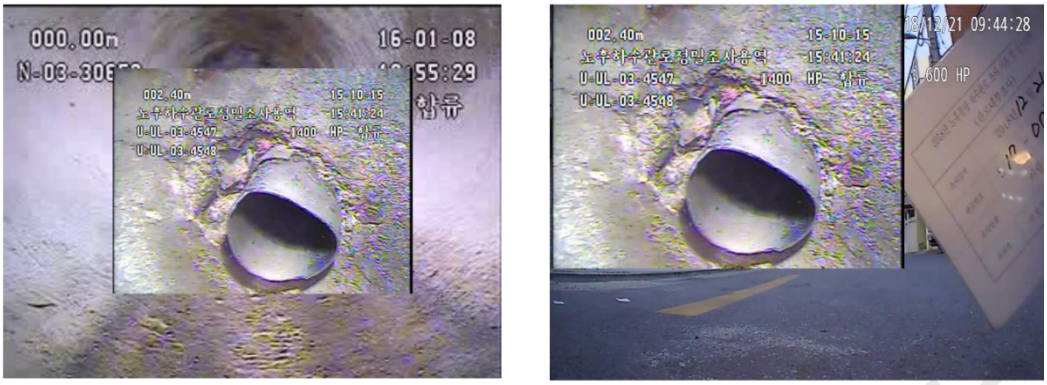


a. Original image (PL: 100%) b. Noise (PL: 98.48%) c. Noise (PL: 96.66%)

d. Noise (PL: 20.17%) e. Noise (PL: 99.95%) f. Noise (PL: 99.85%)

Figure 10. The impact of noise on the classification results of the proposed sewer defect detection model for the protruding lateral (PL) defect.
Note. Images a, b, c, e, and f are correctly predicted as the PL, while the image d is predicted as not PL.

In Fig. 11, we investigated the model performance against challenging cases where the PL image is injected into other images, which makes it more challenging for sewer defect detection. Fig. 11(a) and Fig. 11(c) show the image's prediction inside the pipe where the PL image is being injected into different locations. Both of them are predicted as the PL with an accuracy of 96.83% and 100%, respectively. Fig. 11(b) and Fig. 11 (d) show the image's prediction outside the pipe where the PL image is being injected into different locations. Both of them are predicted as the PL with an accuracy of 73.21% and 85.83%, respectively. The outputs from the proposed model were the PL despite the small defect sizes, which shows the robustness of the defect classification model.



a. Challenging case (PL: 96.83%)

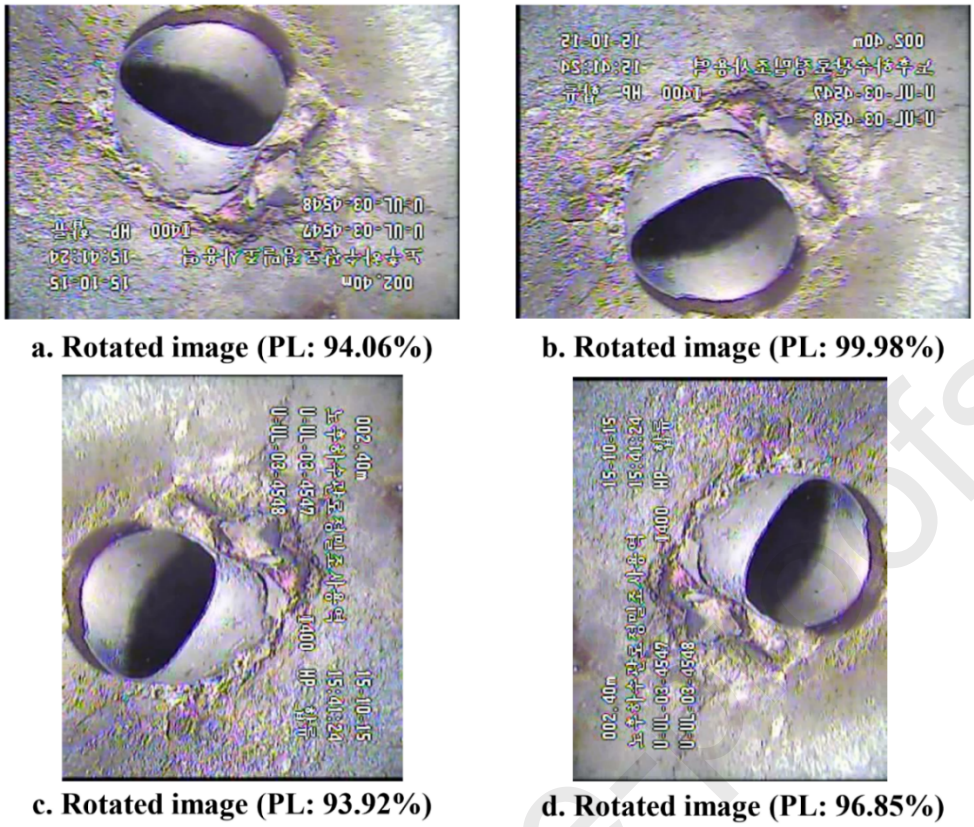
b. Challenging case (PL: 73.21%)

c. Challenging case (PL: 100%)

d. Challenging case (PL: 85.83%)

Figure 11. The experimental results of the proposed sewer defect detection model on challenging cases.

Fig. 12 shows the impact of different types of rotations on the model's performance. The model accurately predicted the defect class with an overall accuracy of over 90%, which is helpful to the CCTV-based sewer defect detection framework.



a. Rotated image (PL: 94.06%) b. Rotated image (PL: 99.98%)

c. Rotated image (PL: 93.92%) d. Rotated image (PL: 96.85%)

Figure 12. The experimental results of the proposed sewer defect detection model on the rotation cases.

5.3. Explainable AI for the proposed model

In this section, three different visualization methods, which include layer activation and CAM [8], show how the model identifies and recognizes a class. These methods were implemented using the Keras-vis library [11], which is a high-level toolkit that is used in order to explain the trained model.

5.3.1. Layer activation visualization

The intermediate activations are useful to gain an understanding of how successive layers transform their input. Thus, the intermediate activations of the CNN were visualized in this section in order to show how the model managed to learn specific abstract features, such as shape, edge, and color from an image through several convolutional layers. Intermediate activations can be displayed by visualizing the feature maps from a specific convolutional layer. Each feature map is dedicated to different types of features, so the most appropriate approach to represent them is by separately displaying the content of each feature map as a 2D image. Fig. 13 shows the visualization of the intermediate activations for the first and the last convolutional layer for the input *open joint* image. The intermediate activations from the first layer still retain the full shape of the input, which focuses on the input image's outer border,

and most of the information from the input image is maintained. However, more abstract features, such as single borders, corners, color, and angles, are learned when the model gets deeper. Therefore, the activation layer visualization becomes less visually interpretable.

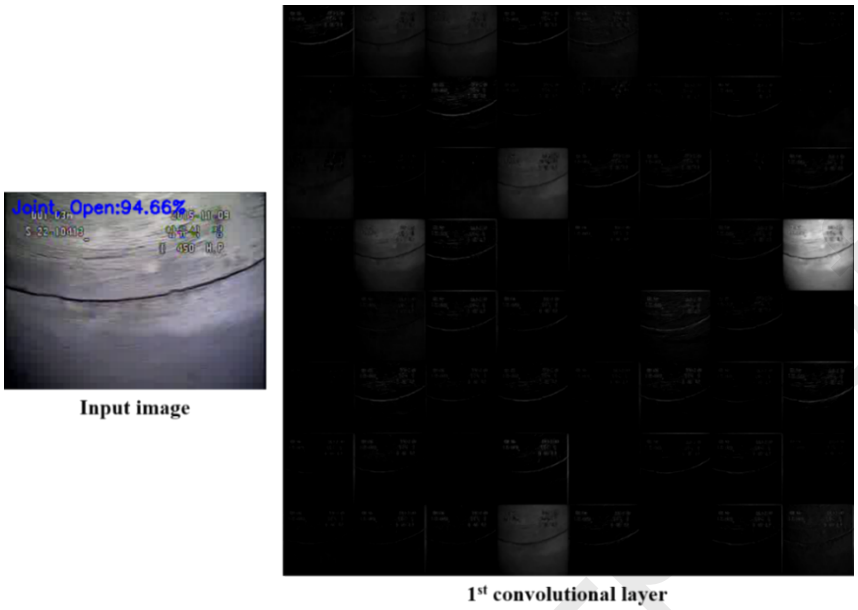


Figure 13. Visualization of 64 intermediate activations belongs to the first convolutional layer (convolution_1) of the proposed model.

5.3.2. Class activation maps (CAM)

CAM allows humans to monitor the essential regions in the image relevant by projecting the class-specific weights of the Softmax layer back to the feature maps from the last convolutional layer. Fig. 14 shows that defect regions for a specific type of defect are accurately highlighted in the corresponding CAM image. As a result, the proposed model shows that it correctly classifies the sewer defect images, which is based on detecting the defect regions.

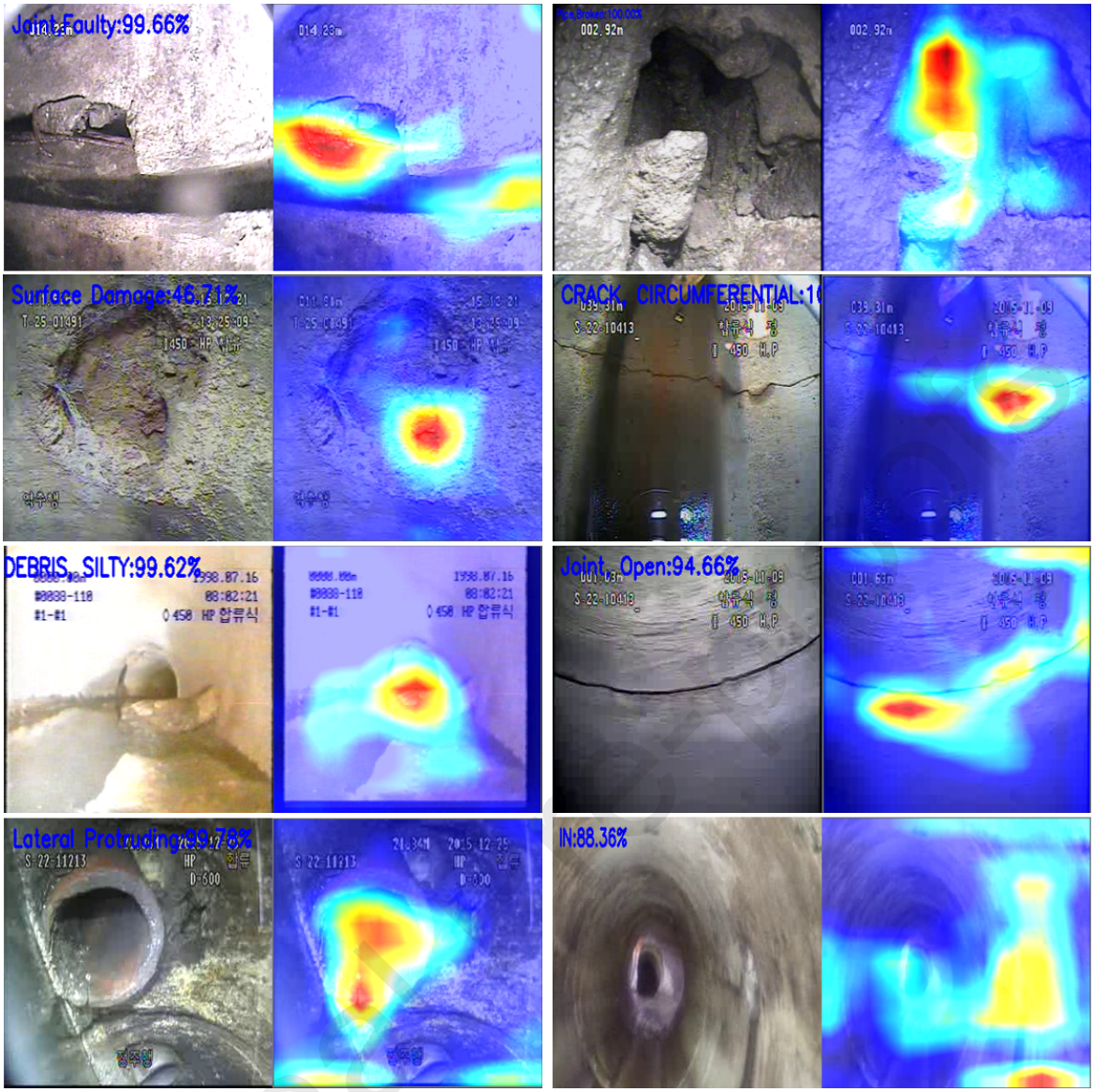


Figure 14. CAM visualization for eight different classes of the collected dataset.
Note. The left image shows the original image, whereas the corresponding CAM image depicts the CAM for the input.

5.4. Defect classification on various imbalanced dataset settings

In this section, two classes, which include crack and normal, were selected to perform the imbalanced experiments. Based on the original images from these two classes (crack: 2380 and normal: 13,934), various data ratios from balance (1:1) to severely imbalance (1:100) were set in Table 4.

Table 4. Number of images for the crack class (CC) and the normal class (NO) with different balancing ratios.

	1:1	**1:2**	**1:3**	**1:4**	**1:5**	**1:6**	**1:7**	**1:8**	**1:9**	**1:10**	**1:100**
CC	1000	1000	1000	1000	1000	1000	1000	1000	1000	1000	130
NO	1000	2000	3000	4000	5000	6000	7000	8000	9000	10,000	13,000

Two different approaches were implemented to cope with the IDP. The data ratio between the different classes is used to set the corresponding weight for each class before the model is trained using the cost-sensitive learning method. Every image that belongs to the minor class is regarded as *n*, which is customizable, images in the majority class. As a result, the model is compelled to handle the imbalanced classes and balanced classes evenly. On the other hand, the ensemble learning-based approach contains two main parts, which include feature extraction and ensemble learning. The fine-tuned VGG-19 model is turned into a feature extractor in the feature extraction part by removing the final Softmax classifier. 4096 feature vectors extracted from the first part are then used to train the XGBoost / LightGBM classifier of the second part. Fig. 15 presents the AUC values of the four different models, which include the proposed model, the cost-sensitive learning-based model (CS), the XGBoost model (XGB), and the LightGBM model (LGB) on various data ratios.

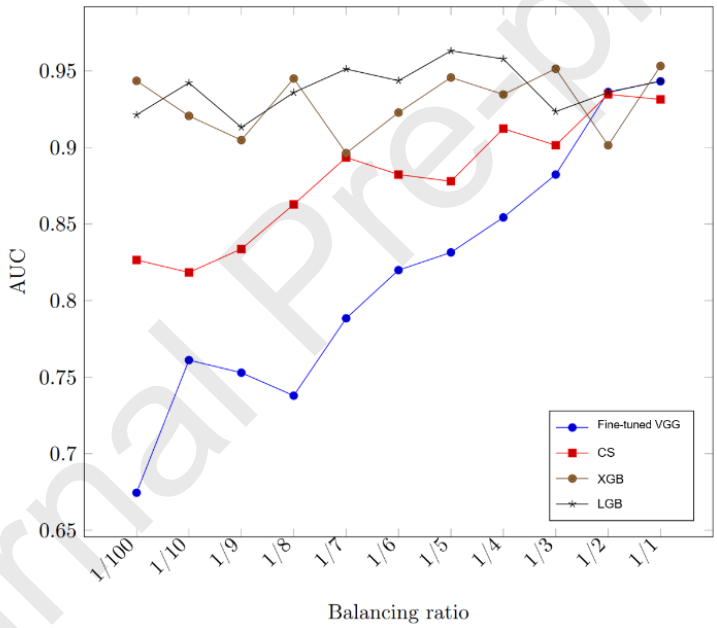


Figure 15. AUC values of four different methods using the introduced imbalanced dataset.
Note. The ratio of crack images (CC) is minor compared to the number of normal images (NO) (CC/NO=1/100, 1/10, 1/9, 1/8, 1/7, 1/6, 1/5, 1/4, 1/3, 1/2, and 1/1).

There is a small difference between the AUC values of the four models when the data ratio is equal to 1/1 and 1/2. However, the fine-tuned VGG model's AUC value decreases significantly when the imbalanced data becomes more severe and reaches 0.67 when the balancing factor is 1/100. Moreover, the three approaches, which are based on the proposed model effectively deal with the IDP even under severely imbalanced data. The XGB model obtains the highest AUC value of 0.94, LGB achieves a slightly lower AUC value of 0.92, and CS gets the lowest AUC value of 0.82 when the data ratio is 1/100. Overall, the LGB and XGB show more robust performance in terms of the AUC value under different data ratio settings

compare to the CS model and the fine-tuned VGG model. In addition, the LGB model's AUC value is better than the XGB model in most cases when the imbalanced ratio gets higher.

5.5. Evaluation of the report generation module

This section attempts to investigate the performance of the report generation module by comparing the model predictions with the ground truth defect reports, which are created manually by the ETRI. The manually generated report is generated by inspectors who assess a sewer video in order to investigate defects that appear, classify them into a specific type, and finally put them in the report. A test sewer CCTV video that has not been utilized for the training process was selected in the first part. The video lasted for 10 minutes and 30 seconds. Table 5 compares the results of the proposed model and the ground truth. Overall, the model correctly classified most defects that are reported in the ground truth report with only one misclassified case, which involved the lateral protruding being misclassified as the *broken pipe*. The ground truth report (xxx.xx m) is different from the distance recognized by the text recognition module (xxx m), because the first three digits indicate the meters. In contrast, the last two are the more extract locations in centimeters.

Table 5. The experimental results of the automated defect classification and location recognition compare with the ground truth data.
Note. The red color indicates the wrong prediction.

#	Defect type		Location	
	Ground truth	**Prediction**	**Ground truth**	**Prediction**
1	Faulty joint	Faulty joint	2.47m	2m
2	Faulty joint	Faulty joint	10.51m	10m
3	Faulty joint	Faulty joint	14.44m	14m
4	Protruding lateral	Protruding lateral	15.63m	15m
5	Protruding lateral	Protruding lateral	24.02m	24m
6	Protruding lateral	Protruding lateral	26.78m	26m
7	Open joint	Open joint	30.38m	30m
8	Protruding lateral	Broken pipe	41.97m	41m
10	Faulty joint	Faulty joint	42.50m	42m
11	Surface damage	Surface damage	43.61m	43m

Fig. 16 shows the generated sewer report that contains the defect image, the defect type, and the location. Moreover, next to the defect type is the current time of the video when the defect occurred. In the second part, twelve sewer CCTV videos and their corresponding reports were selected randomly in order to verify the proposed defect classification system.

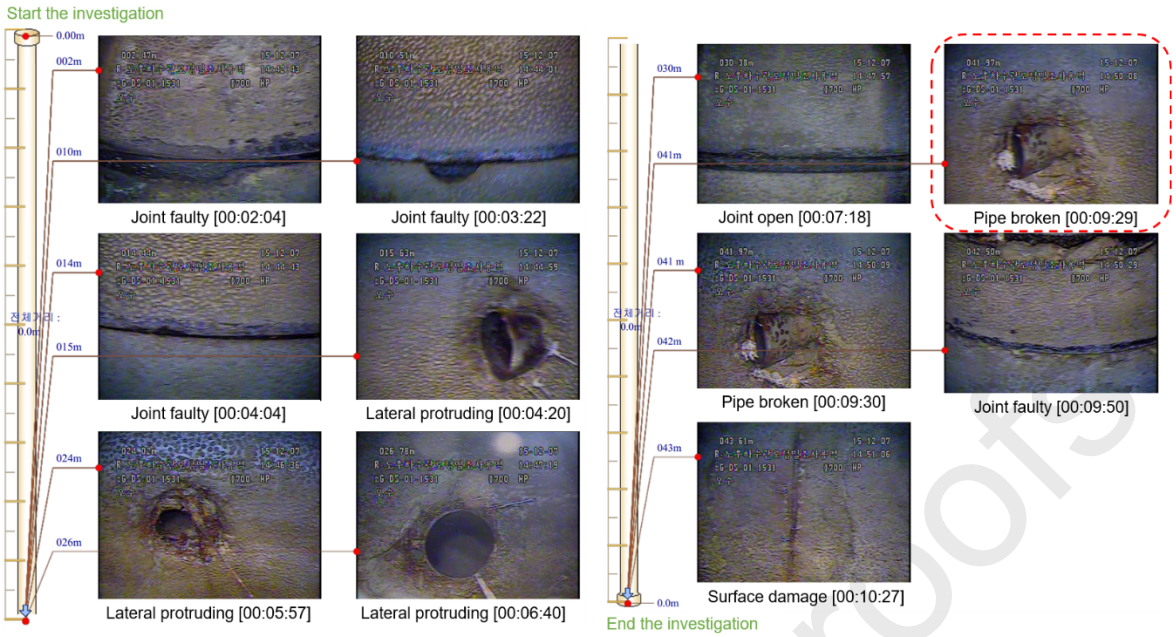


Figure 16. A sample of the sewer inspection report generated by the proposed model.
Note. The red dotted box indicates the wrong prediction, where the lateral protruding image was predicted as pipe broken class.

Fig. 17 supplies the classification performance of the system compare with the manually generated report. Overall, 85 defects were automatically extracted from 12 videos out of 83 defects in the manual report, and our proposed system correctly classified 81 defects. The overall classification accuracy recorded on the twelve videos is 95.3%, which indicates 81/85 defects are correctly classified.

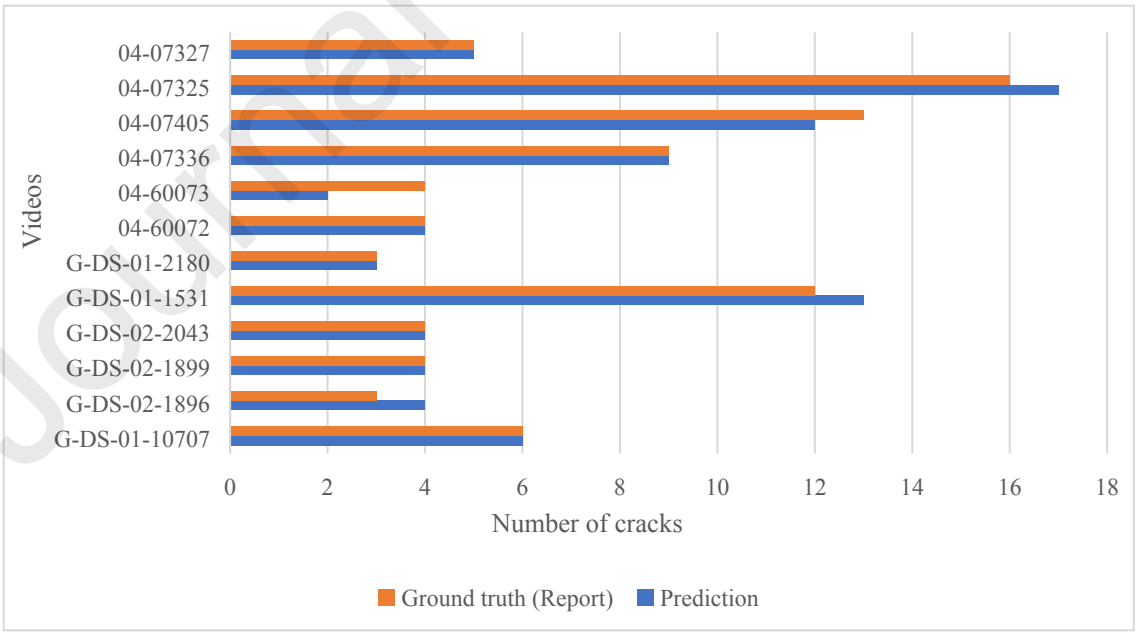


Figure 17. The number of defects detected by the proposed system (prediction) compared to the manually generated ground truth report for twelve random CCTV videos.

6. Conclusion

This paper proposes a deep learning-based automated sewer defect classification framework for the collected images by CCTV videos. In addition, a huge sewer defect dataset, which contains about 38,386 images for the seven defect classes and one normal class, is introduced along with the deep learning model. We also addressed the IDP by applying three methods, which include XGBoost, LightGBM, and misclassification cost customization, to the proposed CNN model.

The experimental results demonstrated that the proposed network could efficiently classify the sewer pipe defects and provide in-depth information for the detected defects by recognizing the subtitles that were printed on the video frames, which include the location, diameter, and type. The proposed method proved to achieve a better defect classification performance than the previous learning-based methods through various experiments with the highest recorded accuracy of 95.7% on the testing set. In addition, the LightGBM extension showed a robust performance in solving the IDP even with the most extreme imbalanced case (1/100). Finally, a novel frame reduction algorithm is implemented based on recognizing the robot travel distance subtitle in order to reduce the number of frames to be processed. The experimental results showed that the model detected various defects precisely with low false alarm rates and robust against several types of attacks, such as noise, rotation, injection, which is appropriate to be integrated into the sewer defect detection applications that use the CCTV videos. Moreover, several XAI approaches were implemented to interpret and explain the proposed model's predictions in order to improve the user's trust and demonstrate that it was possible to apply the proposed model to real-life applications.

However, the system failed to recognize an image with over two types of defects, because it can only recognize the defect with the highest probability. Therefore, a meta-heuristic learner, such as particle swarm optimization (PSO), can be used in the future to train a tailored loss function in a separate layer and independently return the output to enable it to perform defect classification on images that contain more than one type of defect. Another possible solution is to customize the cut-off probability for the multi-classes case. The system can also be extended to perform real-time defect detection and classification on live-streaming CCTV videos that are recorded by a robot, which can provide more precise investigations and support inspectors during the decision-making process.

Acknowledgment

This work was supported by a grant from the project entitled, “Underground Space DB Accuracy Improvement and Underground Utilities Safe Management Technology“, which was funded by Korea Institute of Civil Engineering and Building Technology (KICT) and by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (2020R1A6A1A03038540).

References

[1] J. C. Cheng and M. Wang, "Automated detection of sewer pipe defects in closed-circuit television images using deep learning techniques," Automation in Construction, vol. 95, pp. 155-171, 2018.
[2] S. I. Hassan, L. M. Dang, I. Mehmood, S. Im, C. Choi, J. Kang, et al., "Underground sewer pipe condition assessment based on convolutional neural networks," Automation in Construction, vol. 106, p. 102849, 2019.
[3] J. Gibson and F. Rioja, "Public infrastructure maintenance and the distribution of wealth," Economic Inquiry, vol. 55, pp. 175-186, 2017.
[4] D. Meijer, L. Scholten, F. Clemens, and A. Knobbe, "A defect classification methodology for sewer image sets with convolutional neural networks," Automation in Construction, vol. 104, pp. 281-298, 2019.
[5] Q. Xie, D. Li, J. Xu, Z. Yu, and J. Wang, "Automatic detection and classification of sewer defects via hierarchical deep learning," IEEE Transactions on Automation Science and Engineering, vol. 16, pp. 1836-1847, 2019.
[6] J. B. Haurum and T. B. Moeslund, "A Survey on Image-Based Automation of CCTV and SSET Sewer Inspections," Automation in Construction, vol. 111, p. 103061, 2020.
[7] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," Expert Systems with Applications, vol. 73, pp. 220-239, 2017.
[8] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," IEEE transactions on neural networks and learning systems, vol. 29, pp. 3573-3587, 2017.
[9] L. M. Dang, K. Min, S. Lee, D. Han, and H. Moon, "Tampered and Computer-Generated Face Images Identification Based on Deep Learning," Applied Sciences, vol. 10, p. 505, 2020.
[10] T. N. Nguyen, H. Nguyen-Xuan, and J. Lee, "A novel data-driven nonlinear solver for solid mechanics using time series forecasting," Finite Elements in Analysis and Design, vol. 171, p. 103377, 2020.
[11] L. M. Dang, S. I. Hassan, S. Im, I. Mehmood, and H. Moon, "Utilizing text recognition for the defects extraction in sewers CCTV inspection videos," Computers in Industry, vol. 99, pp. 96-109, 2018.
[12] W. Guo, L. Soibelman, and J. Garrett Jr, "Visual pattern recognition supporting defect reporting and condition assessment of wastewater collection systems," Journal of computing in civil engineering, vol. 23, pp. 160-169, 2009.

[13] J. Myrans, R. Everson, and Z. Kapelan, "Automated detection of faults in sewers using CCTV image sequences," Automation in Construction, vol. 95, pp. 64-71, 2018.
[14] X. Ye, R. Li, Y. Wang, L. Gan, Z. Yu, and X. Hu, "Diagnosis of sewer pipe defects on image recognition of multi-features and support vector machine in a southern Chinese city," Frontiers of Environmental Science & Engineering, vol. 13, p. 17, 2019.
[15] X. Fang, W. Guo, Q. Li, J. Zhu, Z. Chen, J. Yu, et al., "Sewer Pipeline Fault Identification Using Anomaly Detection Algorithms on Video Sequences," IEEE Access, vol. 8, pp. 39574-39586, 2020.
[16] H. Perez, J. H. Tah, and A. Mosavi, "Deep learning for detecting building defects using convolutional neural networks," Sensors, vol. 19, p. 3556, 2019.
[17] B. Krawczyk, "Learning from imbalanced data: open challenges and future directions," Progress in Artificial Intelligence, vol. 5, pp. 221-232, 2016.
[18] B. Wang and J. Pineau, "Online bagging and boosting for imbalanced data streams," IEEE Transactions on Knowledge and Data Engineering, vol. 28, pp. 3353-3366, 2016.
[19] H. Wang, Y. Li, L. M. Dang, J. Ko, D. Han, and H. Moon, "Smartphone-based bulky waste classification using convolutional neural networks," Multimedia Tools and Applications, vol. 79, pp. 29411-29431, 2020.
[20] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 2016, pp. 785-794.
[21] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, et al., "Lightgbm: A highly efficient gradient boosting decision tree," in Advances in neural information processing systems, 2017, pp. 3146-3154.
[22] L. M. Dang, S. I. Hassan, S. Im, and H. Moon, "Face image manipulation detection based on a convolutional neural network," Expert Systems with Applications, vol. 129, pp. 156-168, 2019.
[23] Z. Papanastasopoulos, R. K. Samala, H.-P. Chan, L. Hadjiiski, C. Paramagul, M. A. Helvie, et al., "Explainable AI for medical imaging: deep-learning CNN ensemble for classification of estrogen receptor status from breast MRI," in Medical Imaging 2020: Computer-Aided Diagnosis, 2020, p. 113140Z.
[24] H. Hagras, "Toward human-understandable, explainable AI," Computer, vol. 51, pp. 28-36, 2018.
[25] A. Deeks, "The Judicial Demand for Explainable Artificial Intelligence," Columbia Law Review, vol. 119, pp. 1829-1850, 2019.
[26] Tsai, Du-Ming, and Chien-Ta Lin. "Fast normalized cross correlation for defect detection." Pattern Recognition Letters 24.15 (2003): 2625-2631.
[27] A. Sengupta, Y. Ye, R. Wang, C. Liu, and K. Roy, "Going deeper in spiking neural networks: Vgg and residual architectures," Frontiers in neuroscience, vol. 13, 2019.
[28] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[29] S. S. Kumar, D. M. Abraham, M. R. Jahanshahi, T. Iseley, and J. Starr, "Automated defect classification in sewer closed circuit television inspections using deep convolutional neural networks," Automation in Construction, vol. 91, pp. 273-283, 2018.