buildings

MDPI

Article

Development of a Site Information Classification Model and a Similar-Site Accident Retrieval Model for Construction Using the KLUE-BERT Model

Seung-Hyeon Shin [1], Jeong-Hun Won [2,*], Hyeon-Ji Jeong [3] and Min-Guk Kang [2]

[1] Department of Big Data, Chungbuk National University, Cheongju 28644, Republic of Korea; shshin0317@chungbuk.ac.kr
[2] Department of Safety Engineering, Chungbuk National University, Cheongju 28644, Republic of Korea; kmk2415@chungbuk.ac.kr
[3] Department of Disaster Prevention Engineering, Chungbuk National University, Cheongju 28644, Republic of Korea; jeong1895@chungbuk.ac.kr
* Correspondence: jhwon@chungbuk.ac.kr

Abstract: Before starting any construction work, providing workers with awareness about past similar accident cases is effective in preventing mishaps. Based on construction accident reports, this study developed two models to identify past accidents at sites with similar site information. The site information includes 16 parameters, such as type of work, type of accident, the work in which the accident occurred, weather conditions, contract conditions, type of work, etc. The first model, the site information classification model, uses named entity recognition tasks to classify site information, which is extracted from accident reports. The second model, the similar-site accident retrieval model, which finds the most similar accidents that occurred in the past from input site information, uses a semantic textual similarity task to match the classified information with it. A total of 17,707 accident reports from South Korean construction sites were found; these models were trained to use Korean Language Understanding Evaluation–Bidirectional Encoder Representations from Transformers (KLUE-BERT) for processing. The first model achieved an average accuracy of 0.928, and the second model was precisely matched, with a mean cosine similarity score exceeding 0.90. These models could identify and provide workers with similar past accidents, enabling proactive safety measures, such as site-specific hazard identification and worker education, thereby allowing recognition of construction safety risks before starting work. By integrating site information with historical data, the models offer an effective approach to improving construction safety.

Keywords: past similar accident; accident report; site information; classification model; accident retrieval model

Citation: Shin, S.-H.; Won, J.-H.; Jeong, H.-J.; Kang, M.-G. Development of a Site Information Classification Model and a Similar-Site Accident Retrieval Model for Construction Using the KLUE-BERT Model. *Buildings* **2024**, *14*, 1797. https://doi.org/10.3390/buildings14061797

Academic Editor: Heap-Yih Chong

Received: 10 March 2024
Revised: 3 June 2024
Accepted: 11 June 2024
Published: 13 June 2024

1. Introduction

Although the construction industry is a significant contributor to the development of national economies worldwide [1,2], it is also considered one of the most dangerous industries [3–5]. Although construction workers account for ~7% of all workers in all industries globally, fatal accidents in construction constitute the majority of all fatal industrial accidents [6,7]. The rate of fatal occupational injuries per 100,000 full-time equivalent employees in the United States in 2021 was 3.6 for all industries; however, it was 9.4 in the construction industry, implying that the rate of fatal occupational injuries was 2.6 times higher than that in all industries (United States Bureau of Labor Statistics, 2022). In Korea, the fatal occupational injury rate per 100,000 workers for all industries was 4.3 in 2022, and 16.1 in the construction industry, indicating that the fatality rate in the construction industry was 3.7 times higher than that in all industries [8].

The high accident and fatality rates suggest that there is an urgent need to design and implement accident prevention measures in the construction industry. As an accident

prevention strategy, analyzing accident case studies is a crucial approach to understanding the patterns of accident occurrence and identifying potential risk factors [9–11]. Thus far, several effective accident prevention strategies have been developed by analyzing case studies, and the Korean government has taken measures based on the results, such as establishing public health and safety management guidelines and prohibiting hazardous subcontracting work [12–14]. However, accident case analysis entails not only understanding the causes and consequences but also identifying risk factors in specific situations and strategies to avoid them [15–17]. For construction site managers, workers, and safety managers, identifying the accidents that can occur on their own sites affects risk identification [11,18,19]. Therefore, identifying and proposing specific accidents for the construction site is central to the development and implementation of accident prevention strategies in the construction industry.

However, in practice, proposing suitable accidents is extremely complicated. Construction sites are continually affected by variable factors, such as weather changes, working conditions, and adherence to safety regulations [20–23]. These variables directly affect the likelihood of accidents and increase the complexity of the construction site. Owing to this complexity, previous studies focused on specific accident types or clearly identifiable causes when analyzing accidents because of the related technical or logistical limitations. For example, Marchelli et al. [24] proposed risk assessment and safety management methods for rockfall accidents on construction sites, whereas Luo et al. [25] constructed a framework for predicting the severity of construction site collapse accidents. Dogan et al. [26] developed safety measures for scaffolding accidents based on case studies of scaffolding accidents on construction sites, and Hwang et al. [8] analyzed accident trends based on case studies of accidents and falls from scaffolding or working platforms. These approaches to specific accidents and construction sites help prevent some types of accidents; however, they are restricted in their ability to reflect the complex reality of construction sites [27]. Analyzing suitable accident cases is essential for developing practical strategies for accident prevention; furthermore, combinations of past accidents occurring at different construction sites must also be considered.

Previous accident case analysis methods used fixed categories or limited variables to analyze the condition of sites, and as a result, they failed to comprehensively capture the multifaceted nature and diversity of sites. However, when people explain the site condition using natural language, the information is considered richer and more detailed [25,28]. These data can reveal subtle differences between sites and specific circumstances that cannot be identified easily when using fixed categories [28,29]. Therefore, the classification method of analyzing site information expressed in natural language can more accurately reflect the complex nature of construction sites and is effective in proposing suitable accident cases.

Identifying specific construction sites by considering different variables is essential for performing effective accident analysis and developing prevention strategies; however, the existing research has been limited by its focus on a narrow set of variables, thereby failing to fully capture the complexity of diverse construction site conditions. To overcome these limitations, this study considers the following objectives:

(1) To develop a model for the effective classification of construction site information using natural language processing (henceforth, "the site information classification model"). The main objective is to develop a model that classifies 18 specific types of construction site information from user-entered natural language descriptions, including weather, rate of work progress, work process, and construction type. This model uses a named entity recognition (NER) task to categorize these details.

(2) To develop a model that can find past accidents from sites with similar information by integrating the classification model (henceforth, "the similar-site accident retrieval model"). The objective is to develop a model that uses semantic textual similarity (STS) tasks for identifying construction sites with similar attributes, based on 18 specific types of classified site information from the "site information classification model". Upon determining the similarity, the model can find past accidents corre-

sponding to these sites, effectively linking the classified information with relevant past accident cases.

2. Literature Review

Recently, the importance of data-based approaches in the construction industry has been emphasized. For example, the application of natural language processing (NLP) to construction safety can aid in improving the efficiency of accident prevention and safety measures through the extraction, analysis, and interpretation of information. NLP-related studies in the field of construction safety can be categorized into various topics, including the development of text classification and information extraction methods and the establishment of safety control measures and strategies using NLP. Among these studies, text classification and information extraction studies have focused on classifying and extracting important information from diverse text-based data, including construction accident reports, project documents, and contracts. These studies focus on identifying and classifying specific patterns, types, and risk factors within text-based data. These studies use NLP techniques to extract the required information from the data, and methods are provided to analyze them.

The studies using NLP in the field of construction safety have changed significantly since 2010. Studies between 2016 and 2019 initially focused on using conventional machine learning (ML) techniques. For example, Tixier et al. [30,31] used conventional ML techniques, such as random forest and stochastic gradient tree boosting, to derive combinations of features affecting accidents from construction accident report forms. Goh and Ubeynarayana [32] used six ML algorithms, including support vector machines (SVMs), linear regression (LR), and naive Bayes (NB), to classify construction accident reports obtained from the United States Occupational Safety and Health Administration (OSHA) website. Goh and Ubeynarayana [32] claimed that the SVM was the most effective classification technique for analyzing construction accident reports and classified accident causes and types. In a similar manner to their study, Zhang et al. [33] proposed a model using five NLP techniques, including SVMs, LR, and NB, to classify accident causes from accident reports and used a sequential quadratic programming algorithm to optimize the weights of each classifier in the ensemble model. Furthermore, they proposed a new approach to extracting common causes of accidents using an optimized ensemble model and claimed that it was the best-performing model. The abovementioned studies could not sufficiently capture the diversity and nuances of complex linguistic models. Furthermore, most of these studies focused only on a specific accident type or considered only limited information about the accident location. Although it is useful to classify the causes of accidents, as performed by Zhang et al. [33], there are limitations when considering the combinations of multiple variables present at actual construction sites.

Since 2020, NLP research in the field of construction safety has shifted from using conventional ML algorithms to using deep learning models. For example, transformer-based models have been proposed in the field of NLP, such as long short-term memory (LSTM) or Bidirectional Encoder Representations from Transformers (BERT), to solve the vanishing gradient problem. Furthermore, these models are now being used in the field of construction safety. Baker et al. [34] investigated methods to automatically learn important information for injury prevention from construction accident reports. They investigated new possibilities for accident prediction using deep learning structures, such as convolutional neural networks (CNNs) and hierarchical attention networks, to analyze text patterns in accident reports. Zhong et al. [35] proposed another new framework that combines a latent Dirichlet allocation model with a CNN algorithm, a word co-occurrence network, and word cloud techniques to automatically analyze risk data. The authors claimed that the proposed framework could play an important role in visualizing construction accident reports and developing risk management strategies. Luo et al. [36] performed text mining, including SVM and content mining, to evaluate relationships between accident-causing factors and raw text for 557 falls on construction sites in China that occurred between

2013 and 2019. Using this approach, they demonstrated that certain causal factors affected the occurrence of falls. Moon et al. [37,38] proposed a semantic text-pairing method to automatically identify essential clauses in a review of construction regulations. They used BERT and Doc2Vec models to learn text features and evaluate the relevance of clauses based on cosine similarity. Tian et al. [39] presented an integrated, intelligent approach to text classification and on-site knowledge mining for large-scale construction projects. They aimed to efficiently manage construction site information by combining a CNN-based text classification model with knowledge mining using TF-IDF. Qiao et al. [40] evaluated the performances of models used for classifying specific information from 4770 construction accident reports submitted to OSHA and tested ten shallow learning methods, including SVMs, LR, and NB, and five deep learning methods, including CNNs and LSTM. Zhang [41] used the Word2Vec skip-gram model to learn word embeddings in a domain-specific corpus and proposed a hybrid deep neural network to classify accident reports. Luo and Hirogane [42] used Doc2Vec and BERT to analyze the similarity between accident cases from a total of 941 construction accidents reported to the Ministry of Health, Labor, and Welfare of Japan. Li and Wu [43] and Luo et al. [44] proposed NLP-based classification methods to extract important information by analyzing text from accident reports. Both studies developed models to classify text based on CNN models.

The abovementioned NLP studies in the field of construction safety have gradually shifted from conventional machine learning algorithms to more sophisticated deep learning models, especially those based on transformer architectures like BERT. These models address the vanishing gradient problem, which occurs when gradients used to update the model parameters during training become considerably small, causing the model to stop learning or learn remarkably slowly. By employing advanced architectures such as transformers, the models have been effectively applied in construction safety contexts. Studies such as those by Baker et al. [34] and Zhong et al. [35] have explored novel methods for extracting and analyzing text patterns from construction accident reports using deep learning structures like CNNs and hierarchical attention networks. However, despite these advancements, a significant gap remains in how these technologies are applied to real-world construction site accidents. Most of the existing studies focus narrowly on classifying text from accident reports or on specific types of construction accidents, with limited attention paid to the broader context in which these accidents occur. This narrow focus overlooks the complex interplay of factors at construction sites that can lead to accidents. Furthermore, while some studies, like those by Luo et al. [36] and Tian et al. [39], have begun to integrate NLP with site management practices, they still do not fully leverage NLP to understand the wide range of variables that affect site safety. This study aims to bridge this gap by developing comprehensive models that not only classify site information more effectively, but also utilize this information to identify and retrieve relevant accident cases. This approach allows for a more nuanced understanding of construction site dynamics and factors contributing to accidents, thus supporting the development of more effective prevention strategies. By integrating advanced NLP techniques with a deep understanding of construction site conditions, this study seeks to provide a robust tool for improving safety management practices in the construction industry.

3. Data Collection and Preparation of CSI Accident Reports

3.1. CSI Accident Report Data

The accident data from the construction industry in Korea are mainly reported by the Ministry of Employment and Labor (MOEL) and the Ministry of Land, Infrastructure, and Transport (MOLIT). The accident data from MOEL provide only statistical information, and detailed information is not released, while accident data from MOLIT provide detailed information through a construction safety management integrated information network (CSI). Thus, this study used accident data from the CSI, which has been operated by the Korea Authority of Land and Infrastructure Safety (KALIS) since 2019. The data are provided in text report form, and there is no application programming interface (API) for

data extraction or separately extracted data provided as a database. The accident reports provided by the CSI have been broadly categorized into accident cases, construction site characteristics, and accident surveys. For instance, the accident cases category includes subcategories such as the date of the accident, type of accident, and measures to prevent recurrence. Construction site characteristics may include data such as the construction type, costs, duration, progress ratio, bid price ratio, number of workers, and adherence to safety regulations.

3.2. Data Preprocessing—Creation of the Integrated and Shuffled Data

This study used data from 19,838 accidents that were reported to the CSI from 1 January 2019 to 13 June 2023. The data were systematically crawled using the BeautifulSoup Python library, a tool used for parsing documents to extract data from web pages. The collected data were then stored in CSV format, and the entries lacking complete information were treated as null. During the data cleaning process, the entries lacking necessary details, such as accident type, classification, location, or specific accident details, were excluded from further analysis. This resulted in a dataset of 17,707 accident reports. From the dataset, 18 types of site information, as detailed in Table 1, were utilized to develop and validate the models presented in this study. Since the main purpose of classifying the data is to input site information and find past accident sites with similar information, the site information that can classify the site should be selected according to the CSI database.

Table 1. Site information used in the model.

Type	Site Information	Number of Classes/Range
Categorical parameter	Month of accident occurrence	12 (month)
	Client	2 (public/private)
	Weather	6 (condition)
	Construction type (1st level)	4
	Construction type (2nd level)	15
	Construction type (3rd level)	56
	Accident classification—work type (1st level)	7
	Accident classification—work type (2nd level)	39
	Accident classification—detailed work process	59
	Obligation to prepare safety management plan	2 (yes/no)
	Obligation to prepare DFS (design for safety)	2 (yes/no)
Continuous parameter	Temperature	−50–50(°C)
	Humidity	0–100(%)
	Construction costs	≥0 (won)
	Bid price ratio	0–100(%)
	Work progress ratio	0–100(%)
	Approximate number of workers	≥0 (persons)
	Construction period	≥0 (days)

In Table 1, the "client" category plays a significant role in distinguishing between "public" and "private" construction projects. Here, "public" refers to government-funded or state-sponsored projects, whereas "private" indicates projects undertaken by individual or private entities. This distinction is crucial for understanding the scope and nature of various construction projects because it often influences the regulatory requirements, funding sources, and scale of the project in Korea.

The construction type, represented in Table 1 as "Construction type (1st level/2nd level, /3rd level)", refers to the structure being constructed. The work type in the accident classification, represented in Table 1 as "Accident classification: work type (1st level/2nd level)", indicates the subcategory of the construction type. For example, for the construction type "architecture/building/factory" and the construction type in the accident classification "machinery and equipment/machinery and equipment construction", the end result or intended structure from the construction is a factory, and the type of work being performed at the time of the accident is machinery and equipment construction. In the previous example, the 1st-, 2nd-, and 3rd-level classifications for the construction type are "architecture", "factory", and

"building", respectively. For the construction type in the accident classification, the 1st- and 2nd-level classifications are "machinery and equipment" and "machinery and equipment construction", respectively.

This study also investigated whether the construction projects must comply with the safety management plan and design for safety (DFS) outlined in the South Korean Construction Technology Promotion Act. The safety management plan, developed by contractors, aims at identifying and managing risk factors in the construction process before initiation. This involves developing safety management strategies, such as safety inspection and management organization, to prevent accidents during construction. In contrast, the DFS focuses on the proactive management of construction safety by assessing safety considerations and integrating them into the design stage, mitigating risks during construction. This two-part approach, which covers the entire construction process from planning to design to construction and maintenance, is crucial for preemptive safety management. This study investigated whether construction projects were obligated to include the safety management plan and DFS, thereby categorizing the status of each project as either "obligated" or "not obligated".

The site information used in the models was categorized as either categorical or continuous parameters, depending on the type of data. For example, if the user of the model enters the type of construction or the weather, the input data can be classified into a specific category. In contrast, if the user enters data such as the temperature or construction period, no category is defined. The distinction between categorical and continuous site information changes when the user enters site information and when the site information is output. Among the continuous data types, temperature, humidity, and construction duration are entered and output as continuous data. Construction costs, bid price rate, and number of workers are input by the user as continuous data; however, they are represented as categories in the accident reports, and therefore, they are exported as categorical data. The distinction between categorical and continuous site information is suggested based on the method used by the user to enter the site information into the model.

The subsequent data preprocessing involved two specific steps, each tailored to enhance its usability for modeling. These steps, which are detailed in Figure 1, include the creation of integrated site information, followed by shuffling to simulate natural language structures. In the first step, integration is performed by combining 18 types of site information from the cleaned dataset into a single-site information sentence. This process is referred to as "concatenation" and involves linking multiple pieces of information together into a cohesive one. The integrated site information data are created to provide a basis for processing and understanding the information that is similar to that of natural language. Furthermore, the method to create integrated site information is presented to provide the standard data required to derive the site with information that is most similar to the site information inputted by the user in the similar-site accident retrieval model.

The data were combined by retrieving the site information columns listed in Table 1, linking them to the corresponding investigations, and entering the actual data. Semicolons were used to separate the site information. If data were missing for a particular piece of site information, only the blank site information was excluded. In addition, accident classification data and construction duration data were preprocessed before being used to create the integrated site information. The accident classification refers to the type of construction or work processes performed on the day of the accident. Therefore, when the user enters site information for the accident classification categories, it is assumed that they have used the term "today", instead of entering the information as an accident. Therefore, when the accident classification data are preprocessed, the part corresponding to [Column name] in Figure 1 is not stored as "accident classification", but as "today's construction", "today's construction type", "today's work", or "today's working process". The construction duration data are in the form "year-month-date–year-month-date". However, as those involved in construction record the construction duration as the number of days, the construction duration data are converted and stored as the number of days.

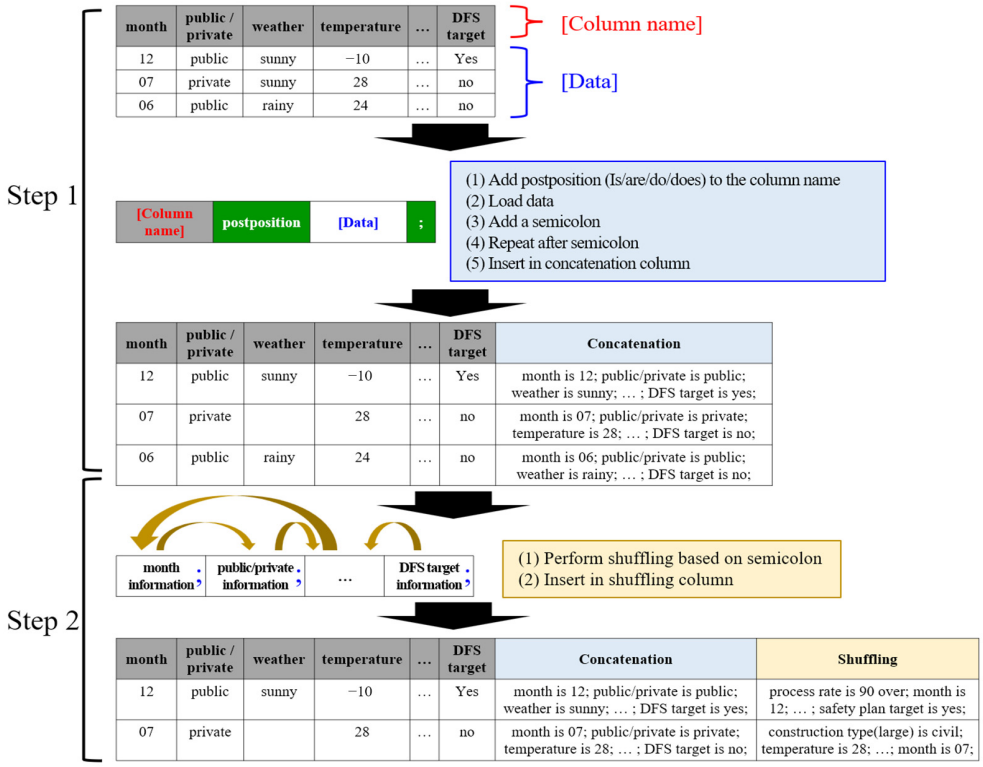

Figure 1. Integrated and shuffled data preparation steps.

In the second step, each piece of site information in the integrated site information is shuffled to create shuffled data. The integrated site information contains formats; hence, inaccuracies can occur in site classification if the user does not use the correct formats for the integrated site information. Therefore, integrated site information in the form of natural language is required to accurately classify raw site information. However, there are limitations in terms of costs when composing natural language site information sentences for all the data. Therefore, in this study, the integrated site information data were shuffled to ensure that the site information was as close as possible to a natural language environment. The site information was shuffled based on semicolons from the integrated site information data using the Python random library (Python 3.10.12).

4. Methodology—Modeling

4.1. Overview of the Proposed Approach

The Korean Language Understanding Evaluation (KLUE)-BERT-based model, which allows for more accurate analysis and processing of the unique linguistic features and context of the Korean language, was used for fine-tuning the two models proposed in this study because the accident cases were input in Korean.

The site information classification model converts and classifies natural language site information entered by the user into a standardized format. To this end, continuous data types are classified using regular expressions, and categorical data types are classified by fine-tuning the site information data with KLUE-BERT, which transforms text data into a high-dimensional vector space, thereby allowing the model to better understand and classify complex linguistic patterns and accidents.

The similar-site accident retrieval model identifies and searches for similar past accidents based on the classified site information. This model also uses the KLUE-BERT-based model to find other accident cases comparable to the classified site information. Processing natural language and understanding the context is particularly important for this model, and therefore, the BERT's powerful language modeling capabilities are used to determine the exact differences and similarities between accident cases.

The performances of both models were evaluated using standard indices, such as F1-score, precision, recall, and cosine similarity. Python 3.10.12, Tensor-Flow 2.15.0, and Pytorch 2.1.0 were used to develop and evaluate the KLUE-BERT-based models.

4.2. Bidirectional Encoder Representations from Transformers (BERT)—KLUE-BERT

BERT uses a transformer encoder architecture to effectively represent language models. The transformer encoder receives an input sequence and understands and expresses the language based on the context of each token. The BERT model, including the KLUE-BERT variant used in this study, operates based on the following key principles and equations, which are essential for its performance in natural language processing tasks. The BERT is pre-trained using two major techniques, i.e., the masked language model (MLM) and next-sentence prediction (NSP) [45]. These techniques enable BERT to understand the context and relationships between words and sentences; this is crucial for tasks such as the classification and similarity measurement in this study.

The MLM randomly masks some words in the input sequence and attempts to predict the masked words.

$$MLM(x) = BERT(x_{masked}) \tag{1}$$

where x and x_{masked} represent the original and masked input sequences, respectively. The masking process is performed by replacing a certain proportion of the words in the input sequence with "[MASK]" [45].

BERT also uses NSP to understand the relationships between sentences. NSP predicts whether the second sentence follows the first sentence when consecutive sentences are provided. This technique is used in question answering and natural language inference [45].

The input processing mechanism of BERT uses WordPiece tokenization. Each input is expressed as the sum of WordPiece embedding, positional embedding, and segment embedding [45], i.e.,

$$e_i = t_i + p_i + s_i \tag{2}$$

where e_i, t_i, s_i, and p_i represent the final embedding vector, WordPiece embedding, segment embedding, and positional embedding, respectively.

BERT uses WordPiece tokenization to convert words into tokens and generate WordPiece embedding vectors for each token.

$$t_i = emb_{wp}(x_i) \tag{3}$$

The transformer model does not contain sequential information, and therefore, positional embedding is added to provide positional information for each token.

$$p_i = emb_{pos}(i) \tag{4}$$

Segment embedding is used to distinguish between two different sentences [45].

$$s_i = emb_{seg}\left(seg_{x_i}\right) \tag{5}$$

These fundamental principles and equations are the basis for the more intensive fine-tuning of models for the downstream tasks of classifying construction site information or measuring the similarity of the information between sites.

KLUE-BERT is used as the baseline model in the KLUE benchmark data, and it was pre-trained with 3 GB of data extracted from texts such as ModuCorpus, CC-100-Kor, Namuwiki, news articles, and petitions [46]. A morpheme-based sub-word tokenizer was used (vocabulary size = 32,000 words), and the model size was 111 M params [46]. The performance of the KLUE-BERT model was evaluated using the KLUE Benchmark, which includes eight tasks (NER; STS; topic classification; and natural language inference) [46]. For NER, KLUE-BERT showed an entity-level macro F1-score of 83.97 and a character-level F1-score of 91.39; for STS performance, Pearson's r was 90.85 and the F1-score was 82.84 [46].

The NER and STS tasks of the KLUE-BERT model were used for the modeling in this study. NER classifies object names into previously defined categories, whereas STS measures the semantic similarity between two sentences. Thus, the NER task is used to classify site information from the natural language site information entered by the user, whereas the STS task is used to retrieve similar site information based on the input site information and output accident cases from the corresponding sites.

4.3. Site Information Classification Model

The site information classification model focuses on classifying construction site information, as depicted in Figure 2. This model's purpose is to extract specific site information from the natural language site descriptions provided by users and to arrange this information into a standardized template. The output of this model is not directly related to accidents but serves as a preparatory step for the retrieval of similar-site accident data. Therefore, the primary function of this model is to organize the natural language site information into a structured format before it is processed by the similar-site accident retrieval model.

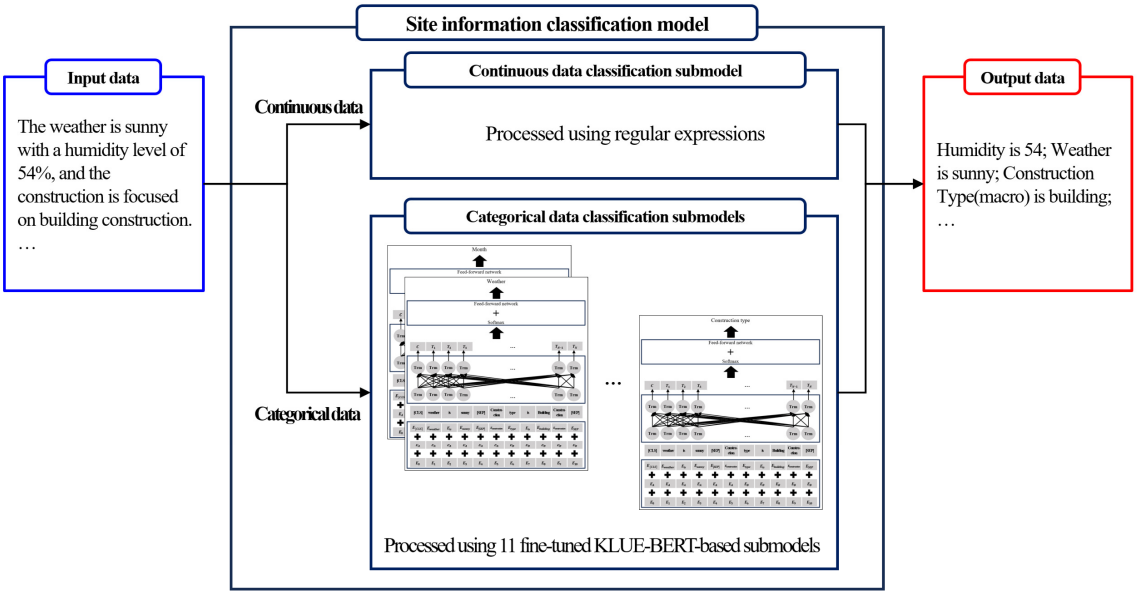

Figure 2. The architecture of the site information classification model.

The model analyzes the natural language construction site information input by a user and identifies specific circumstances and related accident types for the construction sites. The developed model uses integrated construction site data generated in the preprocessing stage and individually classifies 18 types of construction site information. The output site information is provided in the same format as the integrated site information generated in the preprocessing stage.

The site information classification model includes two submodels. The continuous data classification submodel is designed for classifying continuous type site information, and the categorical data classification submodels include multiple submodels for classifying categorical type site information. For classifying continuous site information, the site information entered by the user is processed using regular expressions, which may differ from those in English because the continuous data classification submodel was created to process Korean. The data on construction expenses, progress ratio, number of workers, and construction duration are converted from the data derived with regular expressions to categorical data consistent with the accident data. For example, data such as the user-entered progress ratio, construction expenses, bid price ratio, number of workers, and construction duration are input as continuous data; however, the accident data consists of categorical data, and therefore, these variables are converted to match the categorical

format of the accident data. Temperature and humidity data are processed as continuous data within the accident data, and they are therefore output as continuous data without further transformation.

All the categorical data submodels of the site information classification model have a common architecture, as exemplified by the weather classification submodel shown in Figure 3. After breaking down (tokenizing) the site information sentences into individual components (tokens), we add a special [CLS] token at the beginning of each sentence to indicate the start of the sequence and an [SEP] token at the end to signify the end of the sequence. For example, the sentence "weather is sunny. construction type is building construction." is tokenized and transformed into ["[CLS]", "weather", "is", "sunny", "[SEP]", "construction", "type", "is", "building", "construction", "[SEP]"]. These tokenized sentences, with the added special tokens, are fed into the pretrained KLUE-BERT model. The model generates vector representations (embeddings) for each token, capturing their contextual meanings. These token embeddings are then passed through a classifier that consists of a feed-forward network followed by a softmax function. The classifier assigns each token to its respective category, which corresponds to specific site information. In this example, the classifier identifies the token "sunny" as the category "weather". This process is shown in Figure 3, where the tokens are processed to produce the output "sunny" for the weather category. The submodels determine the categories corresponding to the object names (specific site information).

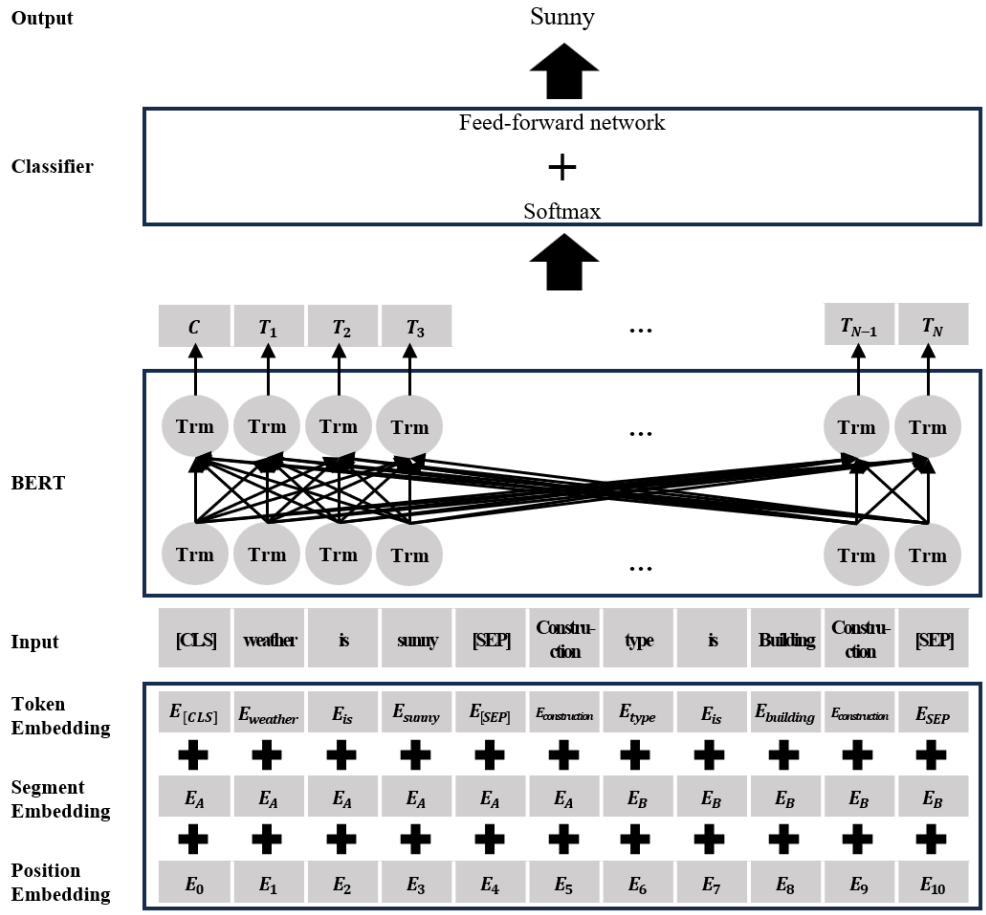

Figure 3. The architecture of a categorical data classification submodel (weather classification example).

Throughout the fine-tuning process, the categorical data classification submodels underwent training to comprehend the contextual significance and site information category of each token by utilizing the natural language format (shuffled data from Figure 1).

Through this process, the submodels were able to identify the keywords and contexts indicating information about the site and to classify them into categories.

In the actual execution stage, the categorical data classification submodels are provided with natural language information about the site entered by the users. In this case, the submodels identify important site information from the sentences entered based on the previously fine-tuned knowledge and determine the category corresponding to each piece of information. For example, if the user enters, "The weather is clear, The construction type is building construction". The submodels identify two types of site information—"weather" and "construction type"—and assign them to the categories "clear" and "building construction", respectively.

The data were split into training, validation, and test sets to train the categorical data classification submodels. The training data were created from 80% of the 17,707 total accident data points, while the validation and test datasets were each created from 10% of the total dataset. The missing data for each piece of site information were omitted from the training, validation, and test sets. Although the amount of data for each submodel varies, a ratio of 80%–10%–10% was consistently maintained between training, validation, and testing across all submodels. The Adam optimizer, known for its efficiency and low memory requirements, was utilized with a learning rate of 5×10^{-5} for the categorical site information submodel. This optimizer adjusts the learning rate for each parameter, making it well suited for the complex nature of our NLP tasks. The batch size for training was 64. The model was trained for 20 epochs, where an epoch refers to one complete pass through the entire training dataset. The optimal epoch was defined as the one with the lowest validation loss or the highest training accuracy. When the natural language site information is entered into the classification model, the classifier outputs the site information in the same format as the data in the integrated site information column.

4.4. Similar-Site Accident Retrieval Model

The similar-site accident retrieval model evaluates the similarity between user-entered site information and previously generated integrated site information and finds the most similar construction site. The model then finds past accidents specific to that site by identifying the similar sites. Cosine similarity is used to assess the similarity between sentences in various fields in addition to the STS task of the BERT model [46,47].

The maximum sequence length was set to 128 based on the analysis of the length of the accident reports in the dataset, which typically did not exceed 60 sub-words, with the majority being 20 sub-words in length. Consequently, the maximum sequence length for the embeddings was limited to 128 to efficiently manage the computational resources of the model and maintain accuracy, despite BERT's capacity to handle up to 512 sub-words.

The method for calculating the cosine similarity between two pieces of site information is shown in Figure 4. In this figure, u and v_n represent the vectors of the input site information and integrated site information within the accident data, respectively. For each piece of site information, the two sentences are input into two pre-trained BERT models, and the embedding vectors are determined by pooling. Here, v_n is previously stored in a separate column (hereafter, the embedding column) after the integrated site information is embedded in the accident data. Next, when site information is input into the model, u is calculated, and the cosine similarity of u is calculated for all values of v_n in the embedding column. The cosine similarity is calculated as

$$cosine - similarity(u, \mathrm{v}_n) = \frac{\mathrm{u} \cdot \mathrm{v}_n}{\|u\| \|v_n\|} \quad (6)$$

where $u \cdot v_n$ represents the dot product of the two vectors, and $\|u\|$ and $\|v_n\|$ represent the respective magnitudes of the vectors.

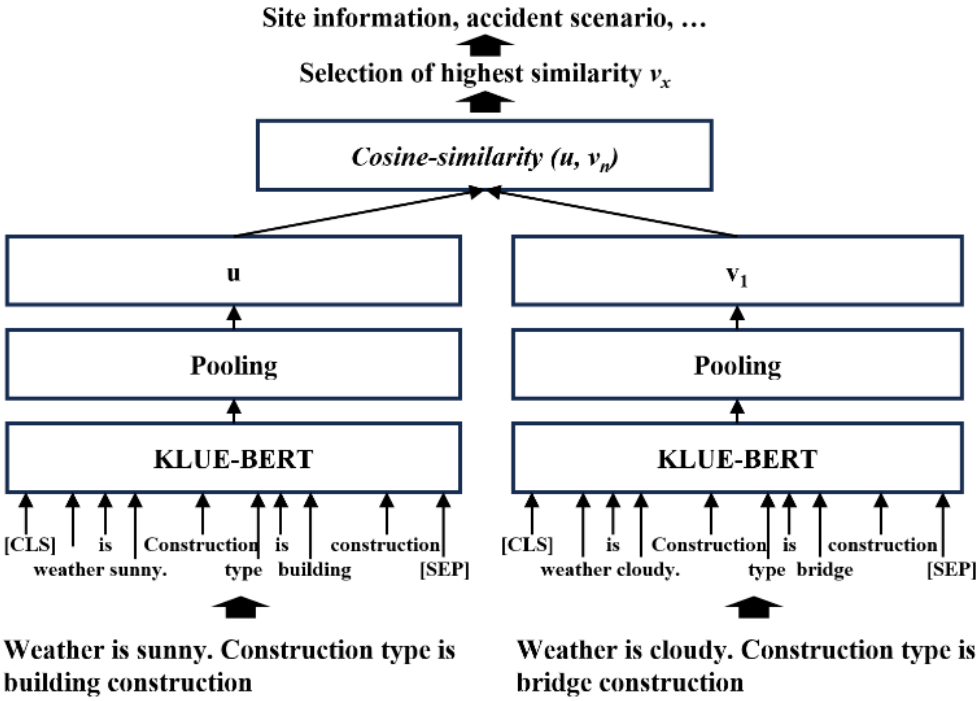

Figure 4. The architecture of the similar-site accident retrieval model.

The column of the integrated site information is ordered by similarity based on the calculated cosine similarity values, and v_n with the highest similarity is selected. The integrated site information data and accident circumstances for this v_x are output. Thus, the output data provide the user with an accident that can occur in a similar situation and the characteristics of that accident.

For example, Figure 4 is not to indicate that these two specific sites ("building construction" and "bridge construction") are similar. Figure 4 illustrates the architecture of the similar-site accident retrieval model, showing how the model calculates the cosine similarity between user-entered site information (u) and each piece of previously stored site information (v_n) from the accident data. The sentence "Weather is sunny. Construction type is building construction" represents user-entered site information (u), while "Weather is cloudy. Construction type is bridge construction" is an example of one of the many site information entries (v_n) from the dataset. The model calculates the cosine similarity between u and each v_n to find the most similar construction site, denoted as v_x, which has the highest cosine similarity score. This process allows the model to identify the site information and accident scenario from the most similar past accidents.

4.5. Model Validation

The proposed models were validated. However, for the site information classification model, the continuous data classification submodel was not validated. When evaluating the performance of the categorical data classification submodels to validate the proposed model, each submodel was evaluated independently. Instead of evaluating each class within these submodels (such as "sunny", "cloudy", or "rain" for the weather classification submodel), the overall performance of each submodel was evaluated. The performance of the model was evaluated using accuracy, loss, precision, recall, and F1-score, which are widely used in machine learning to assess the performance of classification models [48]. In addition, a confusion matrix was output for each model as a measure of model performance. In the confusion matrix, true positive (TP) refers to cases where the predicted and actual value are positive, whereas true negative (TN) refers to cases where the predicted and actual values are negative. False positive (FP) refers to cases where the actual value is

negative but is mistakenly predicted to be positive. False negative (FN) refers to cases in which the actual value is positive but is incorrectly predicted as positive [14,49,50].

Accuracy is calculated as the proportion of the total sample that was correctly classified.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

Precision is the proportion of predicted positive values that were actually positive.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

Recall is the proportion of actual positive values that the model correctly predicted as positive.

$$Recall = \frac{TP}{TP + FN}. \tag{9}$$

F1-score is the harmonic mean of the precision and recall and is a balanced measure of both indices.

$$F1\text{-}score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{10}$$

The performance was evaluated indirectly in terms of the similarity of the site information retrieved by the model using cosine similarity to evaluate the similar-site accident retrieval model. The performance of the model was assessed utilizing cosine similarity because of the inherent challenge in quantifying accuracy, which hinders the use of other evaluation metrics. The site information used in this study consists of a total of 18 categories of site information. An estimated 30 billion potential combinations of site information exist, solely considering combinations of the eleven categories of the basic site information. Given the extreme difficulty in producing site information that precisely matches the input, the performance was assessed in terms of similarity.

The similar-site accident retrieval model was used to measure the cosine similarity between the retrieved site information vector and the actual site information, and the ratio was measured to express the degree of similarity of the output site information. The validity of the developed model was evaluated for all 17,707 accident data points using the following approaches:

1. Similarity between the classified site and actual site information, when only the site information classification model is used on natural language site information (shuffled data from Figure 1)
2. Similarity between the retrieved similar site and actual site information occurs when only the similar-site accident retrieval model is used on the shuffled data.
3. Similarity between the retrieved similar site and actual site information when the site information classification model for the shuffled data is used to classify integrated site information, and the similar-site accident retrieval model is used on the classified site information.

5. Results

5.1. Performance of the Site Information Classification Model

In the evaluation process of the site information classification model, the performance of the categorical data classification submodels was analyzed for 11 categorical data types, except for the continuous data classification submodel. Table 2 shows the size of the training, validation, and test datasets for each categorical data classification submodel. Although the amount of data for each submodel varies, a ratio of 80%–10%–10% was consistently maintained between training, validation, and testing across all the submodels. It can be shown that there is transparency regarding the proposed model's training scope and the consistency of the data split ratios across all the submodels.

Table 2. Number of training, validation, and test datasets and optimal epoch number for each categorical data classification submodel.

Model	Training Data	Validation Data	Test Data	Optimal Epoch Number	Training Accuracy	Validation Loss
Date of accident (month)	14,146	1770	1791	6	0.8629	0.3829
Public/private	14,137	1768	1790	8	0.9914	0.0474
Weather (condition)	14,146	1770	1791	9	0.9585	0.1262
Construction type (1st level)	14,146	1770	1791	5	0.9840	0.0505
Construction type (2nd level)	14,069	1760	1781	4	0.9583	0.1284
Construction type (3rd level)	13,657	1708	1717	6	0.9222	0.3796
Accident classification—work type (1st level)	14,146	1770	1791	5	0.9449	0.2332
Accident classification—work type (2nd level)	14,146	1770	1791	5	0.8955	0.4417
Accident classification—detailed work process	14,146	1770	1791	5	0.8615	0.5604
Obligation to prepare safety management plan	14,145	1770	1791	4	0.9735	0.0822
Obligation to prepare DFS (design for safety)	14,145	1770	1791	5	0.9791	0.0581

We focused on the process of determining the optimal epoch number to the prevent overfitting of each categorical data classification submodel and to maintain stable performance. Furthermore, each of the 11 submodels was trained for 20 epochs to determine the optimal epoch number. The change in the accuracy and loss of each submodel over 20 epochs is shown in Appendix A. The optimal epoch number was determined as the epoch that minimized the validation loss. The optimal epoch numbers for each submodel are shown in Table 2.

The categorical data classification submodels developed in this study were evaluated based on the accuracy, precision, recall, and F1-score of each submodel, as shown in Table 3. In addition, the results of the confusion matrix analysis for each submodel are presented in Appendix B. These confusion matrices provide a detailed breakdown of the model's performance by showing the number of correct and incorrect predictions for each class.

Table 3. Results of the evaluation for each categorical data classification submodel.

Number of Classes	Model	Accuracy	Precision	Recall	F1-Score
2	Public/private	0.9821	0.9833	0.9797	0.9814
2	Obligation to prepare safety management plan	0.9631	0.9590	0.9400	0.9490
2	Obligation to prepare DFS (design for safety)	0.9788	0.9782	0.9743	0.9762
4	Construction type (1st level)	0.9711	0.8987	0.8838	0.8905
6	Weather (condition)	0.9542	0.9614	0.9158	0.9355
7	Accident classification—work type (1st level)	0.9458	0.9623	0.8913	0.9240
12	Date of accident (month)	0.8598	0.8862	0.8581	0.8663
15	Construction type (2nd level)	0.9373	0.9073	0.8460	0.8723
39	Accident classification—work type (2nd level)	0.8813	0.9522	0.8748	0.9082
56	Construction type (3rd level)	0.8892	0.9494	0.9203	0.9303
59	Accident classification—detailed work process	0.8513	0.9487	0.9081	0.9231

The performances showed clear differences when the submodels were categorized into three groups based on the number of classes. In the group with relatively few classes (two to six classes), there were five classification submodels: public/private, weather (condition), construction type (1st level), obligation to prepare a safety management plan, and obligation to prepare a DFS. In this group, the public/private classification submodel showed the best performance, while the weather (condition) and construction type (1st level) classification submodels showed better results than those of the other group's submodels. The obligation to prepare a safety management plan and the obligation to prepare a DFS classification submodels also showed relatively high performance. When the confusion matrices for each submodel were analyzed, the accuracy of all the classifiers was between 0.70 and 1.00.

5.2. Performance of Similar-Site Accident Retrieval Model

The performance evaluation of the similar-site accident retrieval model was roughly divided into three approaches. Specifically, the cosine similarity was measured when using only the site information classifier, when using only the similar-site accident retrieval model, and when using the site information classifier and the similar-site accident retrieval model in parallel. Table 4 shows the results of the measurement of cosine similarity for each approach.

Table 4. Results of the evaluation for each similar-site accident retrieval model.

Evaluation Results	Site Information Classification Model	Similar-Site Accident Retrieval Model	Site Information Classification Model + Similar-Site Accident Retrieval Model
Mean cosine similarity	0.9101	0.9641	0.9803
Ratio with cosine similarity $\geq$ 0.99	0.00%	30.46%	56.77%
Ratio with cosine similarity $\geq$ 0.95	18.58%	65.52%	84.22%
Ratio with cosine similarity $\geq$ 0.90	59.88%	97.31%	99.81%

When the site information classification model was used as a single model, the mean cosine similarity was the lowest of all three approaches. The mean cosine similarity between the classified site information and the actual site information was 0.9101, with 0.00% of the results having a cosine similarity of $\geq$0.99 with the actual site information, 18.58% of the results having a cosine similarity of $\geq$0.95, and 59.88% of the results having a cosine similarity of $\geq$0.90. Thus, when using only the site information classification model, sites were retrieved that were less similar than in the other approaches.

When the similar-site accident retrieval model was used as a single model, the mean cosine similarity between the similar sites retrieved and the actual site information was 0.9641, which was the middle level of performance of the three approaches. The ratio of the results that had a cosine similarity with the actual site information of $\geq$0.90 was 97.31%, but the ratio of the results with a similarity of $\geq$0.99 was 30.46%. This demonstrated a certain level of performance, but also an inability to retrieve similar site information at a detailed level.

When the performance was evaluated using the site information classification model and the similar-site accident retrieval model together, the mean cosine similarity between the similar site information and the actual site information was 0.9803, showing better performance than when using either model individually. When the two models were used in parallel, more than half of the results of the similar site information (56.77%) had a cosine similarity of $\geq$0.99 with the actual site information. In particular, most of the results of the derived site information (99.81%) showed a cosine similarity of $\geq$0.90 with the actual site information.

5.3. Example of Retrieving Accidents from Similar Sites

Figure 5 shows an example of a result that found a past accident from a similar site that fit the entered site information using the developed model. When detailed site information was entered in the input window (black box) and the model was executed, the results indicated that the accident occurred at sites with the most similar site information that occurred in the work of the entered information. The model presented the final results of running both the site information classification model and the similar-site accident retrieval model. The results included a description of the accident (red box), the similarity of the accident site (blue box), and the details of the site where the accident occurred (green box). In this example, the similarity of the site where the accident occurred was 89% of the fieldwork information entered. Before starting on-site work, construction site managers can educate workers about the accidents derived from the model to alert them.

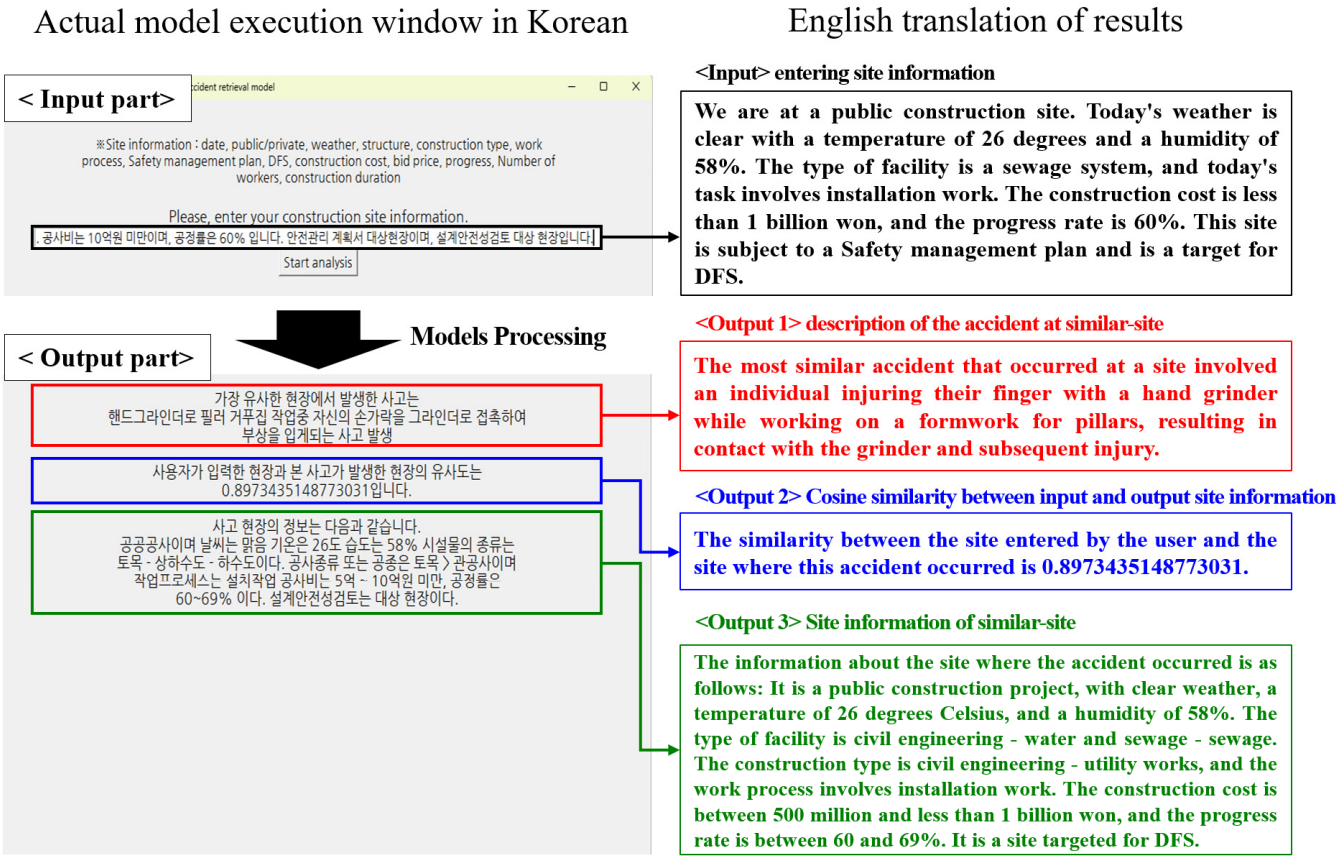

Figure 5. An example of finding a similar accident using development models.

6. Conclusions

This study developed two models for identifying and providing similar past accidents to workers using input site information prior to starting work. The first model is the site information classification model, which uses a named entity recognition task to classify input site information. The second model is a similar-site accident retrieval model that finds the most similar accidents that occurred in the past from input site information and uses semantic textual similarity. These models employ advanced NLP techniques to accurately categorize diverse construction site data into structured formats and identify accidents corresponding to classified site information.

(1) The site information classification model effectively categorized both continuous and categorical types of input site information. This categorization facilitates a deeper understanding of site-specific conditions, which are critical for assessing potential hazards and implementing appropriate safety measures.

(2) The similar-site accident retrieval model could identify similar past accidents with high accuracy, as indicated by a mean cosine similarity score of $\geq$0.90, when paired with the site information classification model. This result indicated that the combination of these two models could give birth to correct results that were more comparable to the actual site information.

(3) The developed models can be effectively used by construction site managers for safety management. For example, site managers can follow these steps to utilize the models. Before starting daily work, the site manager inputs the current site information, including environmental conditions, contract details, and specific work processes, into the models. The models process the input data to retrieve records of similar past accidents from the database. By examining the retrieved accident records, the site manager can identify potential risk factors associated with the current site conditions. The manager can use the details of past accidents to educate workers about specific

hazards and safety measures relevant to the day's tasks. Informed safety briefings may be conducted with workers by using the retrieved accident data to highlight specific risks and preventive actions. This proactive approach allows site managers to implement targeted safety measures, thereby enhancing overall safety awareness and preparedness among workers.

Despite the models' effectiveness, this study had some limitations.

(1) The efficiency of the models strongly depends on the scope and quality of the datasets. The datasets used in this study were based on temporally and geographically limited accident cases; therefore, they could not accurately reflect the characteristics of accidents in other environments or conditions. Furthermore, the structure of the model needs to be adjusted when applied to different datasets because the results of this study are based on a specific dataset. The change in the structure can help ensure that the model is effective in various environments and conditions.
(2) This study focused on Korean language data, and therefore, further studies on transferability to other language environments are required. Despite focusing on Korean data, the methodologies developed in this study can aid in creating global safety management models. The framework proposed in this study to classify site information and retrieve relevant past accidents is adaptable and offers a blueprint for researchers to tailor safety measures according to their specific regional needs.
(3) This study did not conduct an empirical validation of the models with actual construction site users owing to constraints.

Future studies should address the limitations of this study and focus on developing models that are applicable to diverse construction sites and linguistic environments. In addition, recognizing the significance of real-world applicability, future research should also include empirical validation to further ascertain the effectiveness of the proposed models in practical settings.

Author Contributions: S.-H.S., conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft, visualization, writing—review and editing; J.-H.W., conceptualization, validation, resources, and writing—review and editing, investigation, supervision, and project administration; H.-J.J., investigation, writing—review and editing; M.-G.K., investigation, writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are available in the Construction Safety Management Integrated Information Network (CSI) at https://www.csi.go.kr/index.do (accessed on 10 June 2024). These data were derived from the resources available in the public domain and managed by the Korea Authority of Land and Infrastructure Safety (KALIS).

Acknowledgments: This research was supported by Chungbuk National University Korea National University Development Project (2022).

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Accuracy and Loss for Site Information Classification Submodels over 20 Epochs

This appendix details the changes in accuracy and loss across 20 epochs for 11 categorical data classification submodels, as shown in Figures A1–A11.

Appendix A.1. Date of Accident (Month) Classification Submodel

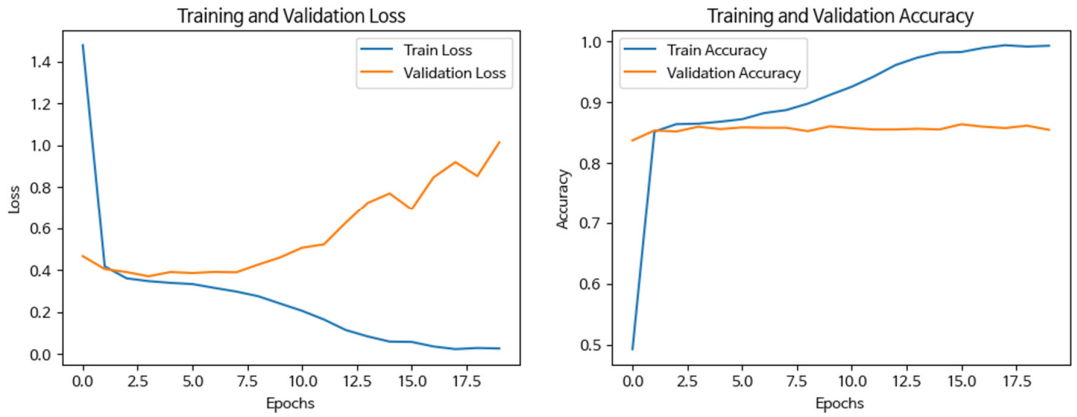

Figure A1. Visualization of accuracy and loss for the training and validation stages of date of accident (month) classification submodel.

Appendix A.2. Public/Private Classification Submodel

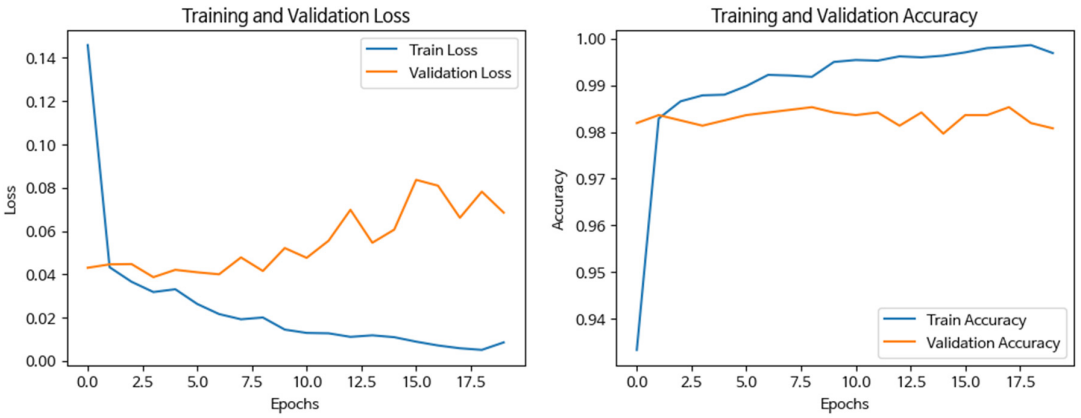

Figure A2. Visualization of accuracy and loss for the training and validation stages of public/private classification submodel.

Appendix A.3. Weather (State) Classification Submodel

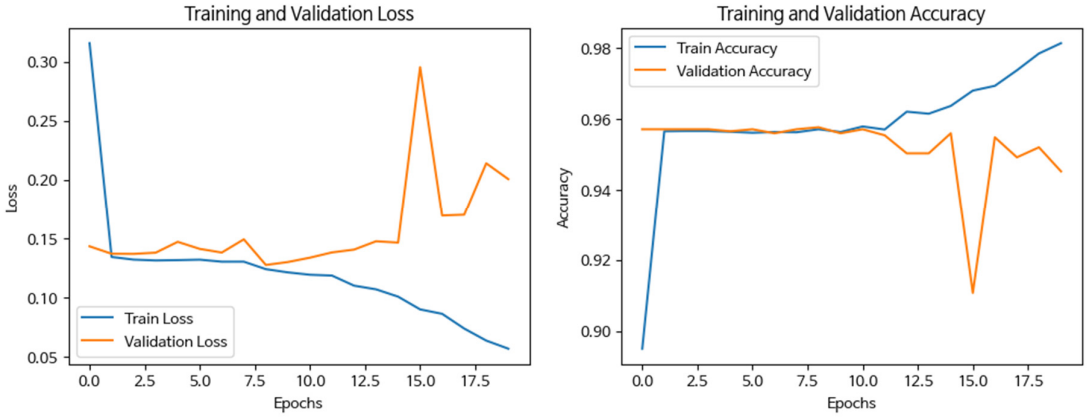

Figure A3. Visualization of accuracy and loss for the training and validation stages of weather classification submodel.

Appendix A.4. Construction Type (Macro) Classification Submodel

Figure A4. Visualization of accuracy and loss for the training and validation stages of construction type (macro) classification submodel.

Appendix A.5. Construction Type (Meso) Classification Submodel

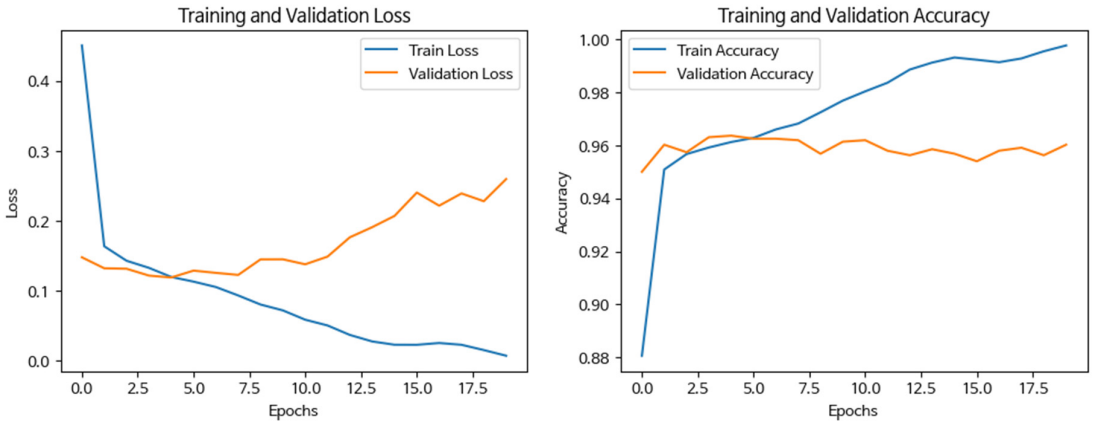

Figure A5. Visualization of accuracy and loss for the training and validation stages of construction type (meso) classification submodel.

Appendix A.6. Construction Type (Micro) Classification Submodel

Figure A6. Visualization of accuracy and loss for the training and validation stages of construction type (micro) classification submodel.

Appendix A.7. Accident Classification—Construction Type (Macro) Classification Submodel

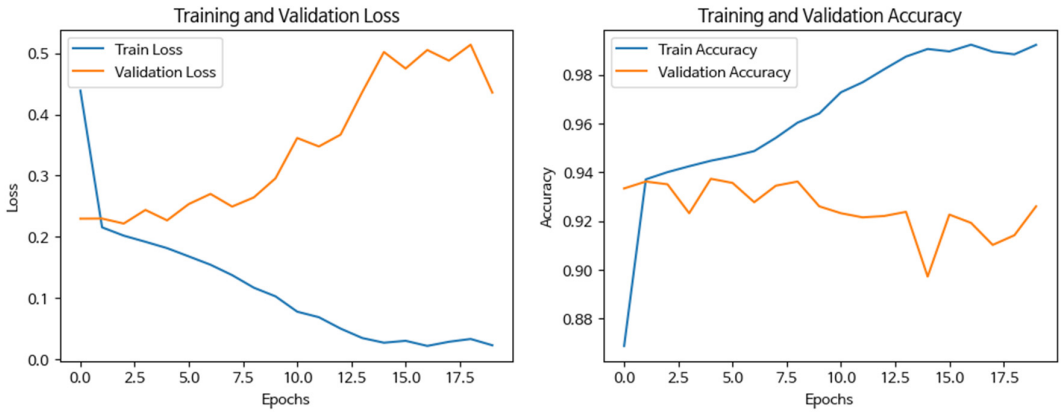

Figure A7. Visualization of accuracy and loss for the training and validation stages of accident classification—construction type (macro) classification submodel.

Appendix A.8. Accident Classification—Construction Type (Micro) Classification Submodel

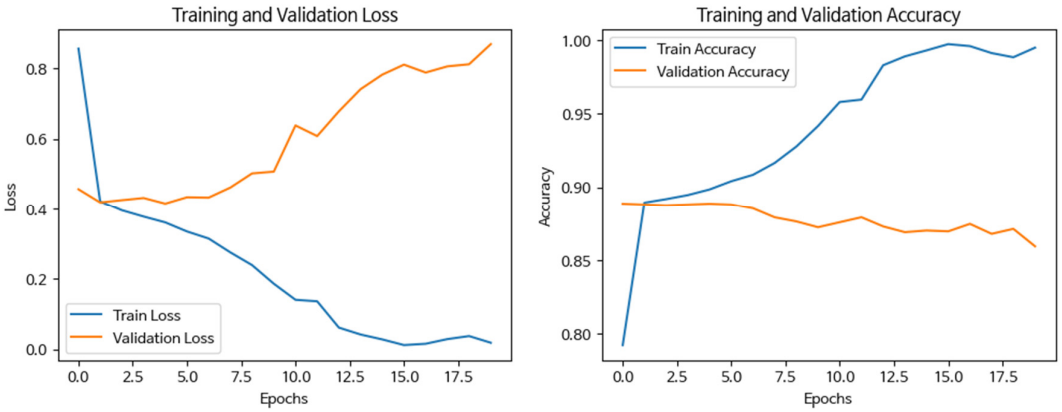

Figure A8. Visualization of accuracy and loss for the training and validation stages of accident classification—construction type (micro) classification submodel.

Appendix A.9. Accident Classification—Work Process Classification Submodel

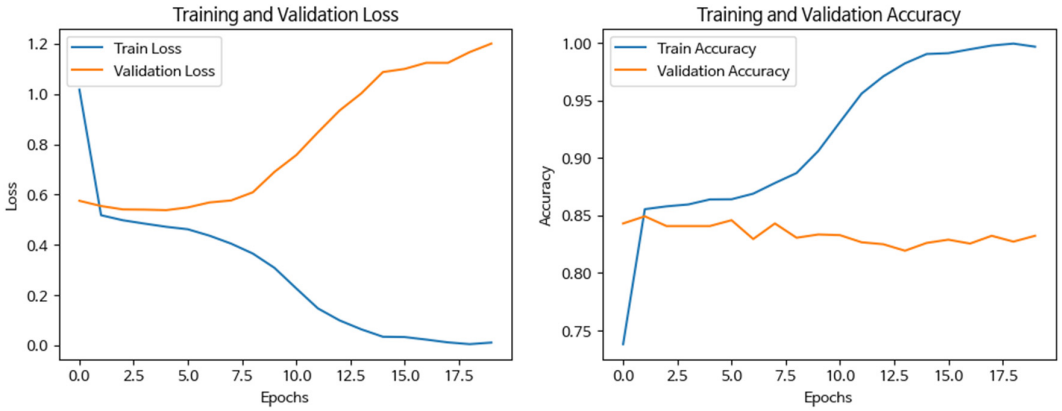

Figure A9. Visualization of accuracy and loss for the training and validation stages of accident classification—work process classification submodel.

Appendix A.10. Safety Management Plan Classification Submodel

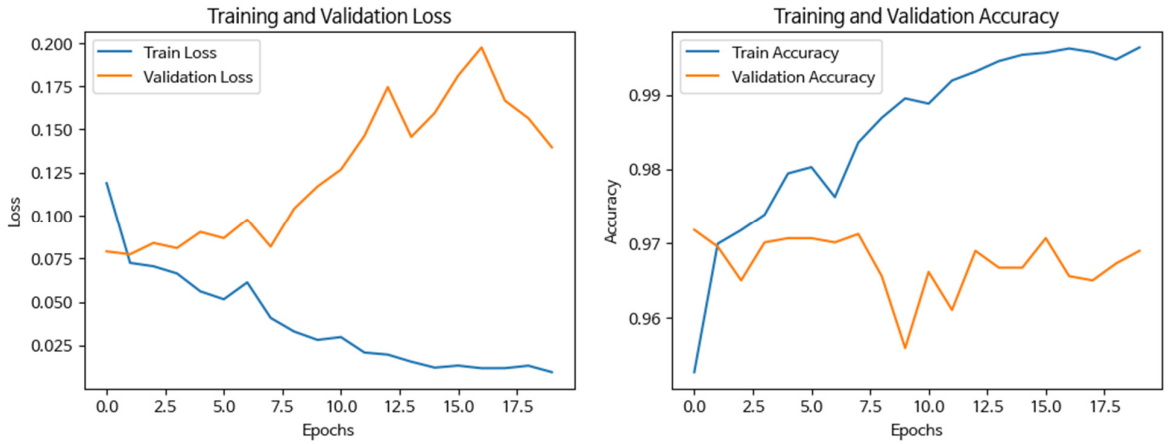

Figure A10. Visualization of accuracy and loss for the training and validation stages of safety management plan classification submodel.

Appendix A.11. DFS Classification Submodel

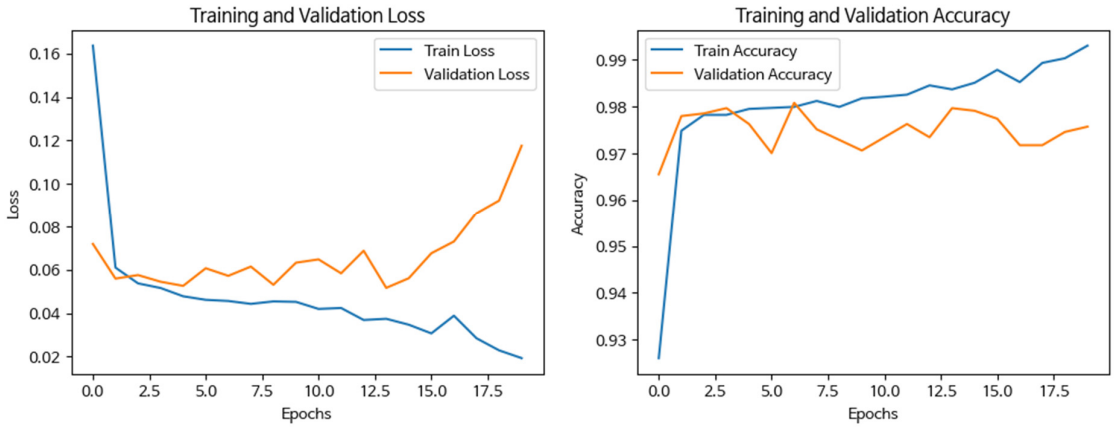

Figure A11. Visualization of accuracy and loss for the training and validation stages of DFS classification submodel.

Appendix B. Confusion Matrix Results for Site Information Classification Submodels

This appendix presents the confusion matrix results for each of the site information classification submodels, as shown in Figures A12–A22.

Appendix B.1. Date of Accident (Month) Classification Submodel

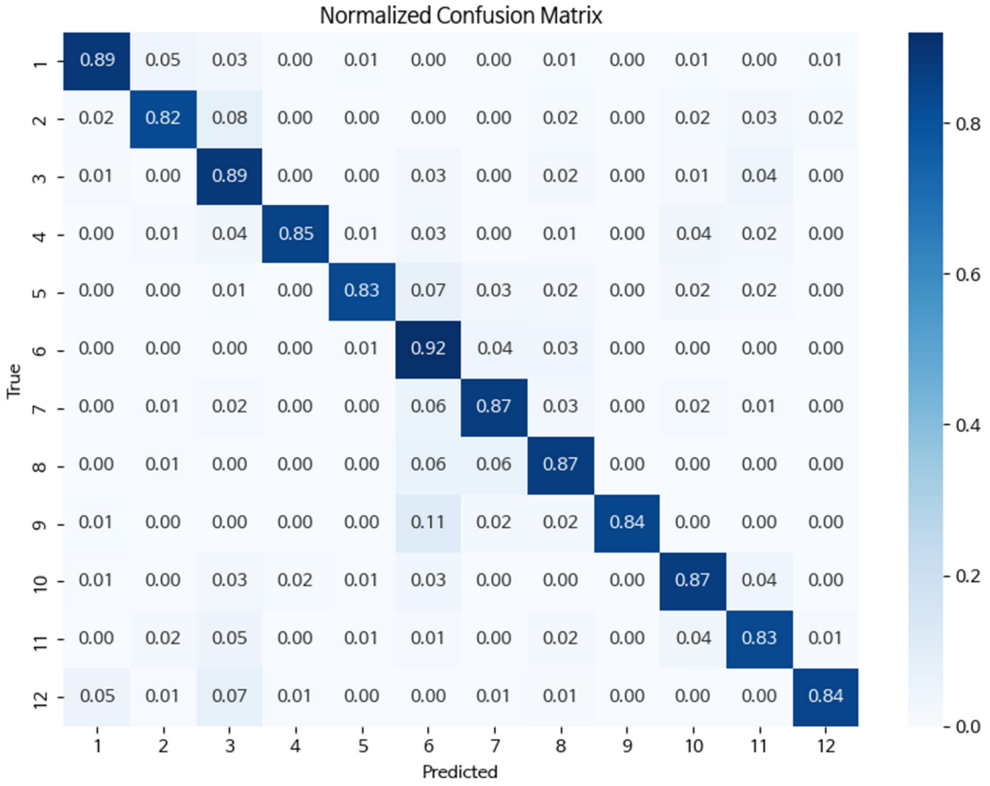

Figure A12. Confusion matrix for date of accident (month) classification submodel.

Appendix B.2. Public/Private Classification Submodel

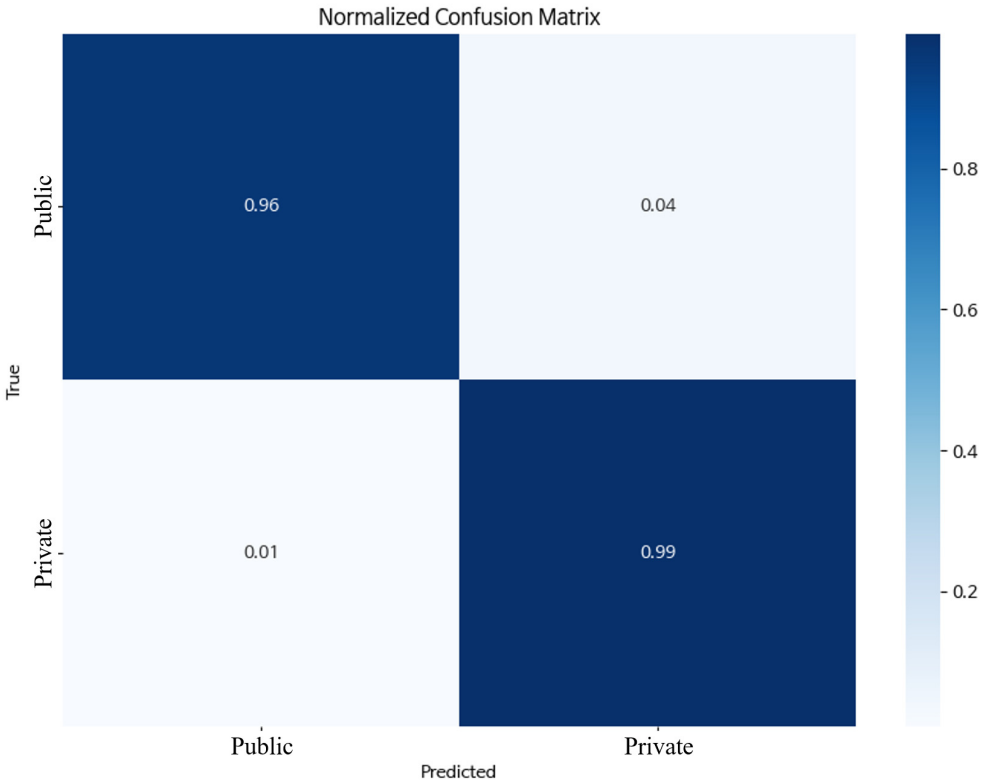

Figure A13. Confusion matrix for public/private classification submodel.

Appendix B.3. Weather (State) Classification Submodel

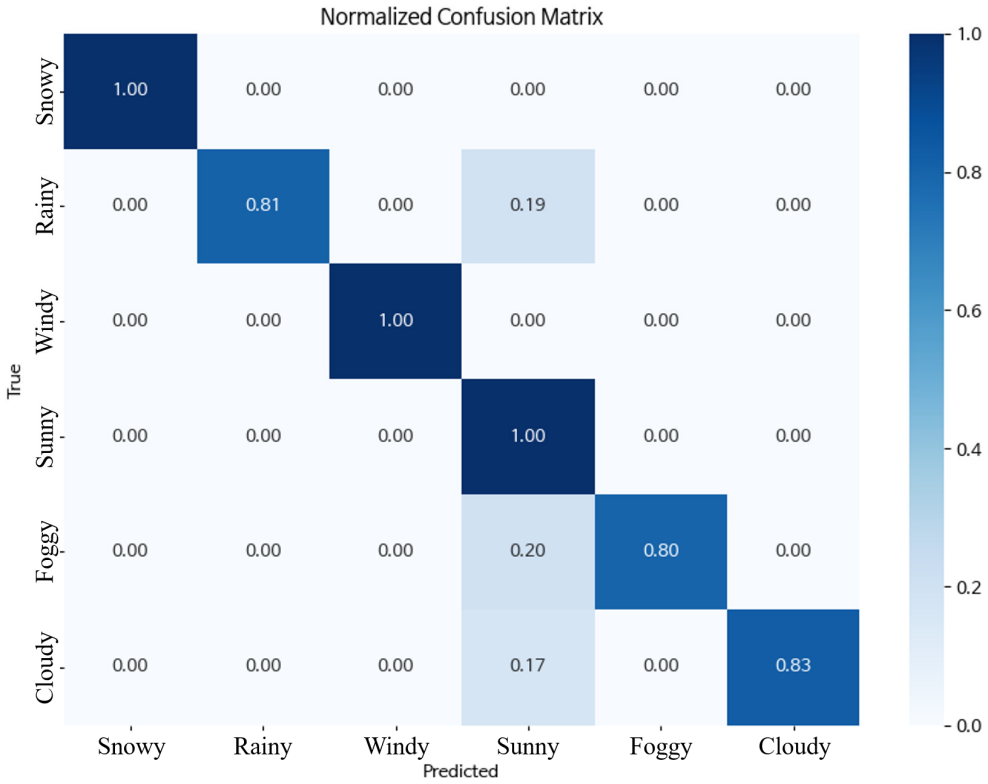

Figure A14. Confusion matrix for weather (state) classification submodel.

Appendix B.4. Construction Type (Macro) Classification Submodel

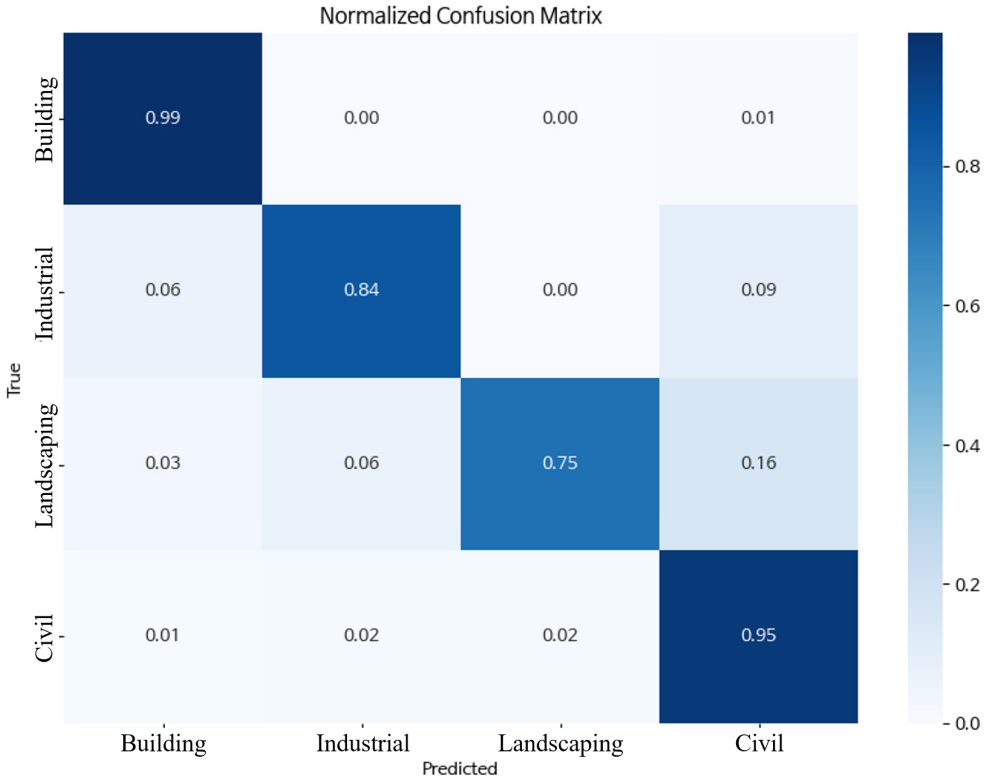

Figure A15. Confusion matrix for construction type (macro) classification submodel.

Appendix B.5. Construction Type (Meso) Classification Submodel

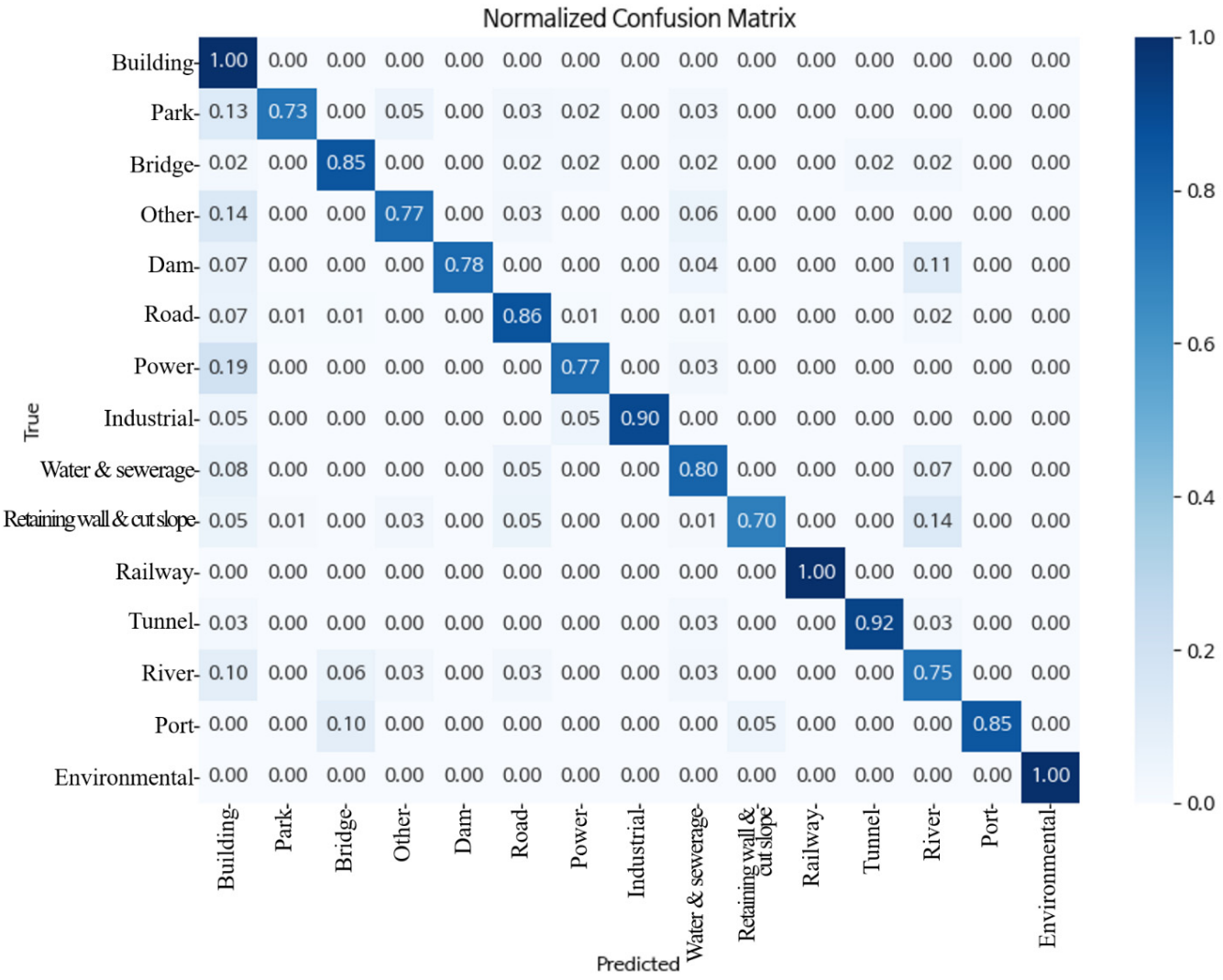

Figure A16. Confusion matrix for construction type (meso) classification submodel.

Appendix B.6. Construction Type (Micro) Classification Submodel

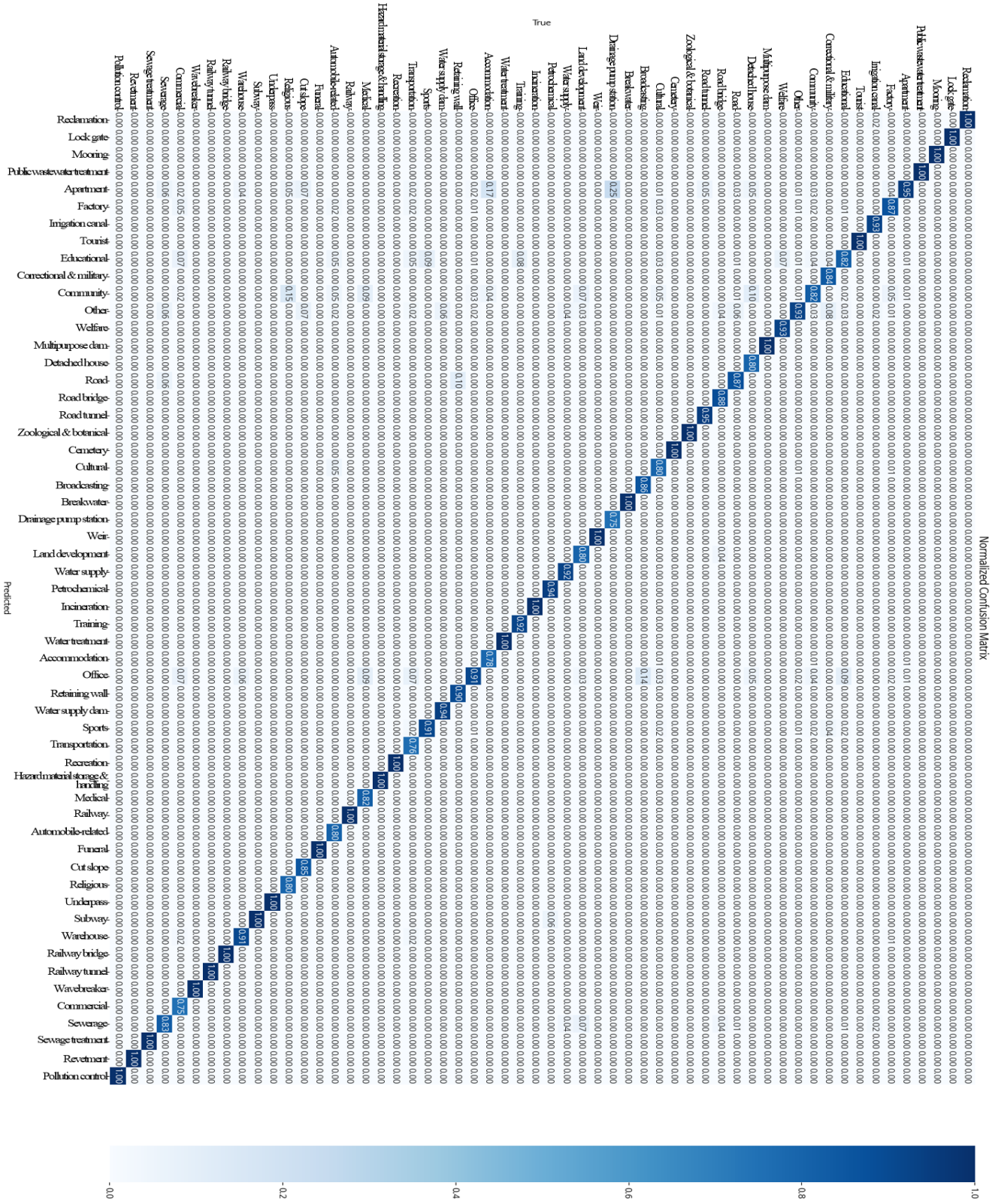

Figure A17. Confusion matrix for construction type (micro) classification submodel.

Appendix B.7. Accident Classification—Construction Type (Macro) Classification Submodel

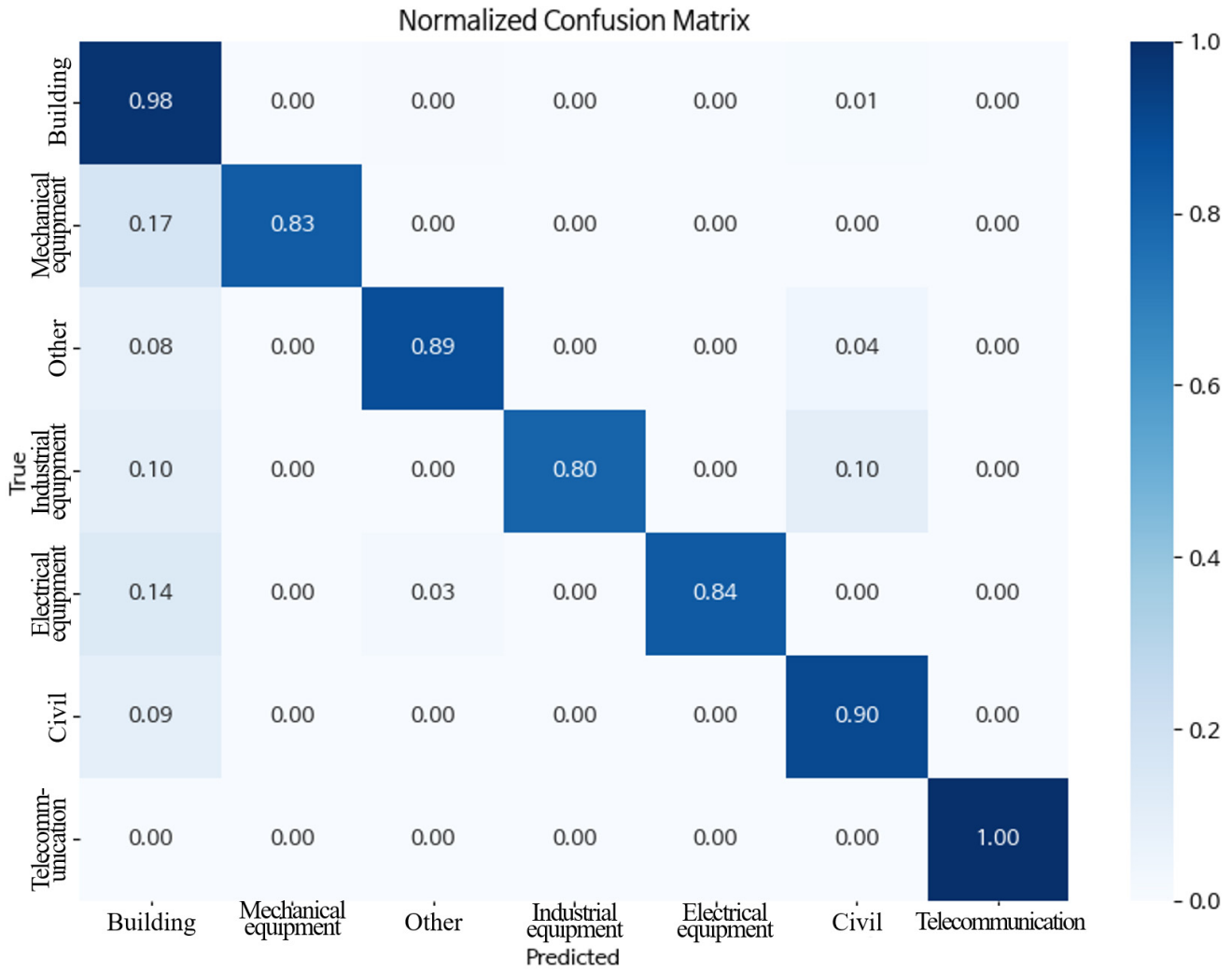

Figure A18. Confusion matrix for accident classification—construction type (macro) classification submodel.

Appendix B.8. Accident Classification—Construction Type (Micro) Classification Submodel

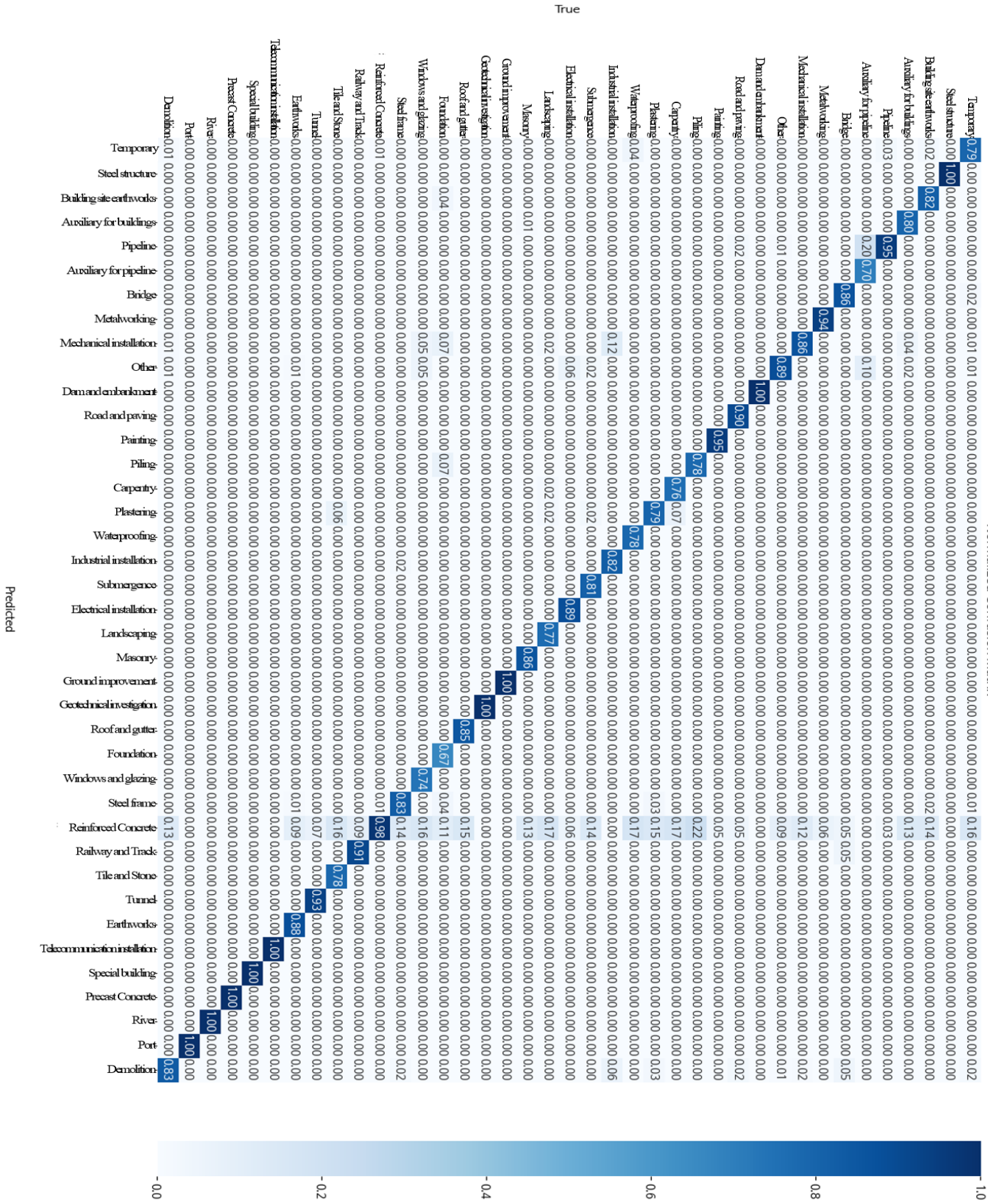

Figure A19. Confusion matrix for accident classification—construction type (micro) classification submodel.

Appendix B.9. Accident Classification—Work Process Classification Submodel

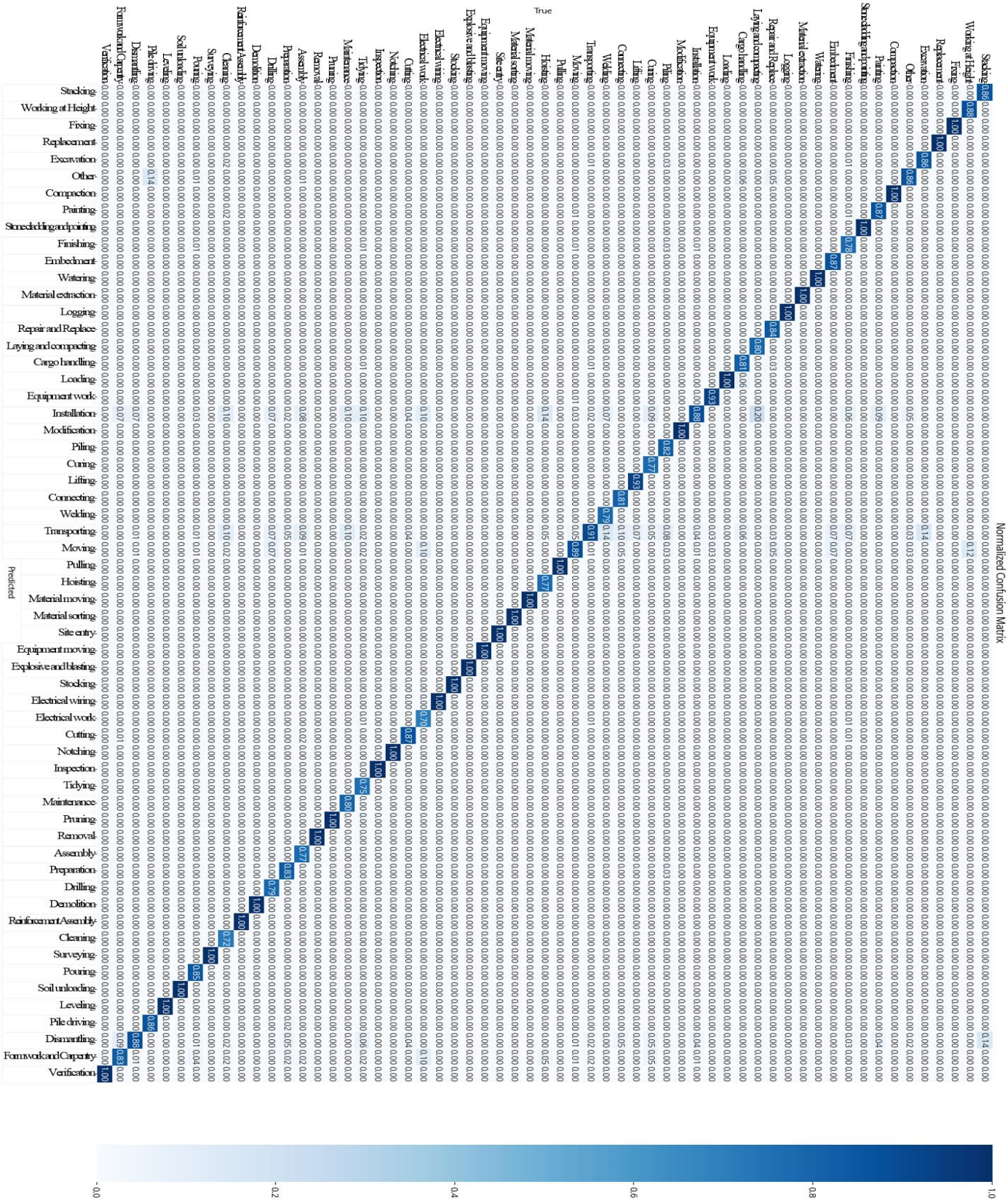

Figure A20. Confusion matrix for accident classification—work process classification submodel.

Appendix B.10. Safety Management Plan Classification Submodel

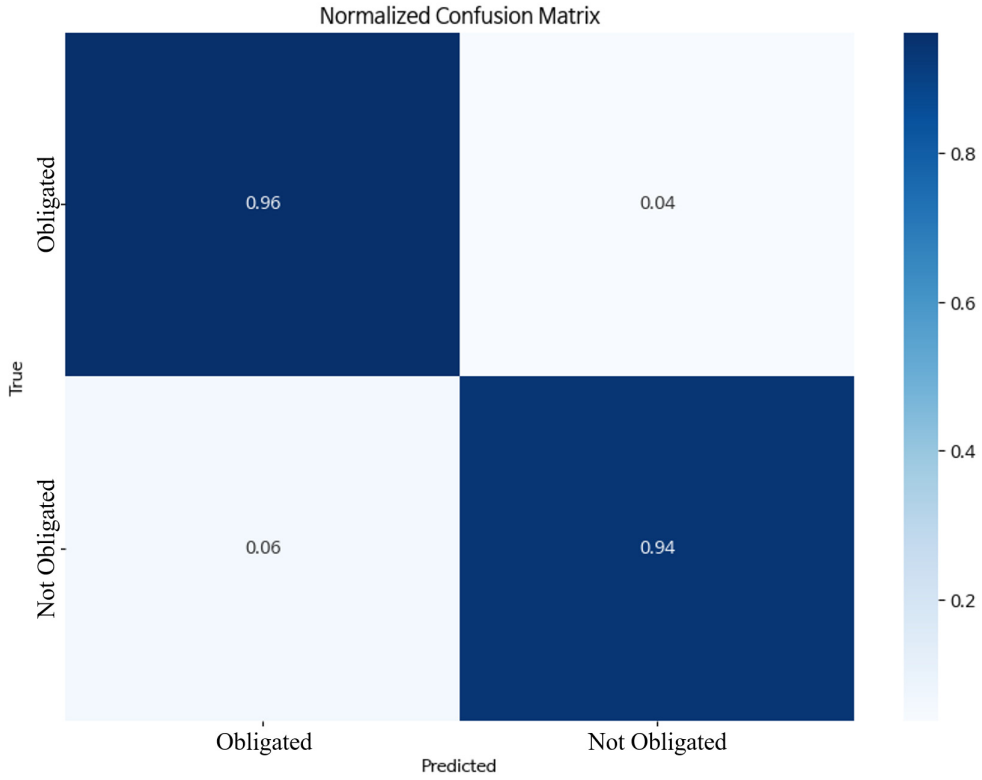

Figure A21. Confusion matrix for safety management plan classification submodel.

Appendix B.11. DFS Classification Submodel

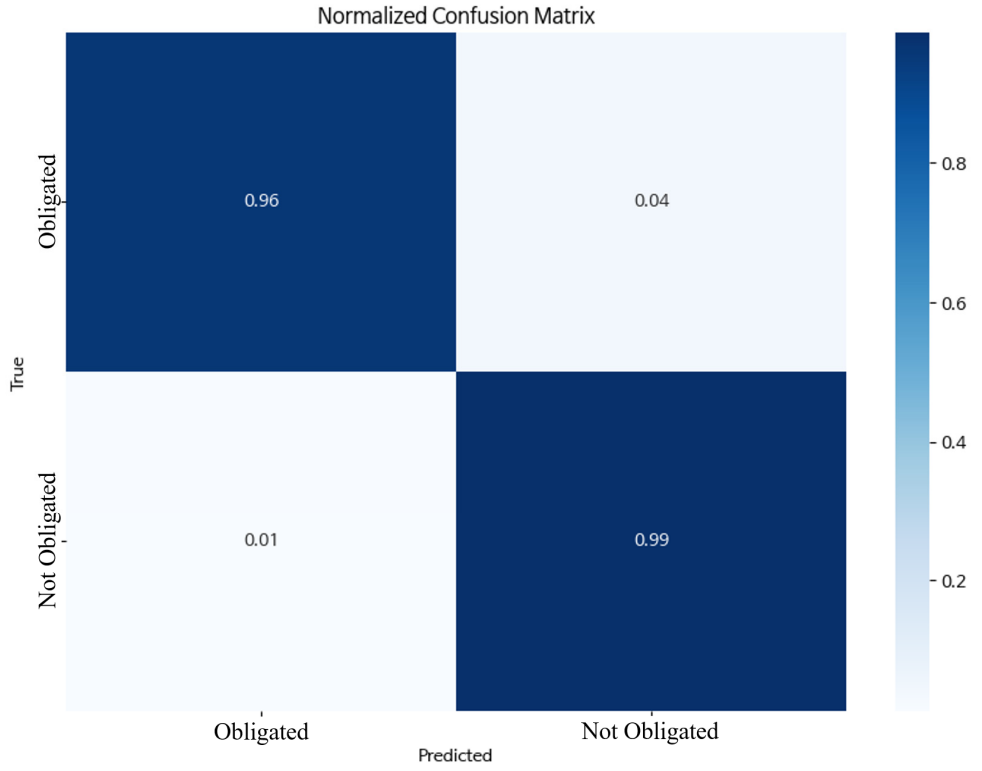

Figure A22. Confusion matrix for DFS classification submodel.

References

1. Alaloul, W.S.; Musarat, M.A.; Liew, M.S.; Qureshi, A.H.; Maqsoom, A. Investigating the impact of inflation on labour wages in Construction Industry of Malaysia. *Ain Shams Eng. J.* **2021**, *12*, 1575–1582. [CrossRef]
2. Hillebrandt, P.M. *Economic Theory and the Construction Industry*; Springer: Berlin/Heidelberg, Germany, 1985.
3. Wang, D.; Qin, Y.; He, Y. The effect of leader–follower psychological capital congruence on safety behavior. *Buildings* **2024**, *14*, 1. [CrossRef]
4. Omer, M.M.; Mohd-Ezazee, N.M.A.; Lee, Y.S.; Rajabi, M.S.; Rahman, R.A. Constructive and destructive leadership behaviors, skills, styles and traits in BIM-based construction projects. *Buildings* **2022**, *12*, 2068. [CrossRef]
5. Tam, C.M.; Zeng, S.X.; Deng, Z.M. Identifying elements of poor construction safety management in China. *Saf. Sci.* **2004**, *42*, 569–586. [CrossRef]
6. Shuang, Q.; Zhang, Z. Determining critical cause combination of fatality accidents on construction sites with machine learning techniques. *Buildings* **2023**, *13*, 345. [CrossRef]
7. Mock, C.N.; Nugent, R.; Kobusingye, O.; Smith, K.R. *Disease Control Priorities, Third Edition (Volume 7): Injury Prevention and Environmental Health*; World Bank Publications: Washington, DC, USA, 2017.
8. Hwang, J.M.; Won, J.H.; Jeong, H.J.; Shin, S.H. Identifying critical factors and trends leading to fatal accidents in small-scale construction sites in Korea. *Buildings* **2023**, *13*, 2472. [CrossRef]
9. Harms-Ringdahl, L. *Guide to Safety Analysis for Accident Prevention*; IRS Riskhantering AB: Stockholm, Sweden, 2013.
10. Hollnagel, E. *Barriers and Accident Prevention*; Routledge: London, UK, 2016.
11. Zhang, W.; Zhu, S.; Zhang, X.; Zhao, T. Identification of critical causes of construction accidents in China using a model based on system thinking and case analysis. *Saf. Sci.* **2020**, *121*, 606–618. [CrossRef]
12. Feng, Z.; Lovreglio, R.; Yiu, T.W.; Acosta, D.M.; Sun, B.; Li, N. Immersive virtual reality training for excavation safety and hazard identification. *Smart Sustain. Built Environ.* **2023**. [CrossRef]
13. Halabi, Y.; Xu, H.; Long, D.; Chen, Y.; Yu, Z.; Alhaek, F.; Alhaddad, W. Causal factors and risk assessment of fall accidents in the U.S. construction industry: A comprehensive data analysis (2000–2020). *Saf. Sci.* **2022**, *146*, 105537. [CrossRef]
14. Kang, K.; Ryu, H. Predicting types of occupational accidents at construction sites in Korea using random forest model. *Saf. Sci.* **2019**, *120*, 226–236. [CrossRef]
15. Abdelhamid, T.S.; Everett, J.G. Identifying root causes of construction accidents. *J. Constr. Eng. Manag.* **2000**, *126*, 52–60. [CrossRef]
16. Leveson, N. A new accident model for engineering safer systems. *Saf. Sci.* **2004**, *42*, 237–270. [CrossRef]
17. Yousri, E.; Sayed, A.E.B.; Farag, M.A.M.; Abdelalim, A.M. Risk identification of building construction projects in Egypt. *Buildings* **2023**, *13*, 1084. [CrossRef]
18. Albert, A.; Hallowell, M.R.; Skaggs, M.; Kleiner, B. Empirical measurement and improvement of hazard recognition skill. *Saf. Sci.* **2017**, *93*, 1–8. [CrossRef]
19. Carter, G.; Smith, S.D. Safety hazard identification on construction projects. *J. Constr. Eng. Manag.* **2006**, *132*, 197–205. [CrossRef]
20. Buchholz, B.; Paquet, V.; Punnett, L.; Lee, D.; Moir, S. PATH: A work sampling-based approach to ergonomic job analysis for construction and other non-repetitive work. *Appl. Ergon.* **1996**, *27*, 177–187. [CrossRef] [PubMed]
21. Jannadi, O.A.; Bu-Khamsin, M.S. Safety factors considered by industrial contractors in Saudi Arabia. *Build. Environ.* **2002**, *37*, 539–547. [CrossRef]
22. Yi, K.J.; Langford, D. Scheduling-based risk estimation and safety planning for construction projects. *J. Constr. Eng. Manag.* **2006**, *132*, 626–635. [CrossRef]
23. Rani, H.A.; Radzi, A.R.; Alias, A.R.; Almutairi, S.; Rahman, R.A. Factors affecting workplace well-being: Building construction projects. *Buildings* **2022**, *12*, 910. [CrossRef]
24. Marchelli, M.; Coltrinari, G.; Alfaro Degan, G.A.; Peila, D. Towards a procedure to manage safety on construction sites of rockfall protective measures. *Saf. Sci.* **2023**, *168*, 106307. [CrossRef]
25. Luo, X.; Li, X.; Goh, Y.M.; Song, X.; Liu, Q. Application of machine learning technology for occupational accident severity prediction in the case of construction collapse accidents. *Saf. Sci.* **2023**, *163*, 106138. [CrossRef]
26. Dogan, E.; Yurdusev, M.A.; Yildizel, S.A.; Calis, G. Investigation of scaffolding accident in a construction site: A case study analysis. *Eng. Fail. Anal.* **2021**, *120*, 105108. [CrossRef]
27. Gürcanli, G.E.; Müngen, U. An occupational safety risk analysis method at construction sites using fuzzy sets. *Int. J. Ind. Ergon.* **2009**, *39*, 371–387. [CrossRef]
28. Wu, C.; Li, X.; Guo, Y.; Wang, J.; Ren, Z.; Wang, M.; Yang, Z. Natural language processing for smart construction: Current status and future directions. *Autom. Constr.* **2022**, *134*, 104059. [CrossRef]
29. Fang, W.; Luo, H.; Xu, S.; Love, P.E.D.; Lu, Z.; Ye, C. Automated text classification of near-misses from safety reports: An improved deep learning approach. *Adv. Eng. Inform.* **2020**, *44*, 101060. [CrossRef]
30. Tixier, A.J.-P.; Hallowell, M.R.; Rajagopalan, B.; Bowman, D. Application of machine learning to construction injury prediction. *Autom. Constr.* **2016**, *69*, 102–114. [CrossRef]
31. Tixier, A.J.-P.; Hallowell, M.R.; Rajagopalan, B.; Bowman, D. Construction safety clash detection: Identifying safety incompatibilities among fundamental attributes using data mining. *Autom. Constr.* **2017**, *74*, 39–54. [CrossRef]
32. Goh, Y.M.; Ubeynarayana, C.U. Construction accident narrative classification: An evaluation of text mining techniques. *Accid. Anal. Prev.* **2017**, *108*, 122–130. [CrossRef] [PubMed]

33. Zhang, F.; Fleyeh, H.; Wang, X.; Lu, M. Construction site accident analysis using text mining and natural language processing techniques. *Autom. Constr.* **2019**, *99*, 238–248. [CrossRef]
34. Baker, H.; Hallowell, M.R.; Tixier, A.J.-P. Automatically learning construction injury precursors from text. *Autom. Constr.* **2020**, *118*, 103145. [CrossRef]
35. Zhong, B.; Pan, X.; Love, P.E.D.; Sun, J.; Tao, C. Hazard analysis: A deep learning and text mining framework for accident prevention. *Adv. Eng. Inform.* **2020**, *46*, 101152. [CrossRef]
36. Luo, X.; Liu, Q.; Qiu, Z. A correlation analysis of construction site fall accidents based on text mining. *Front. Built Environ.* **2021**, *7*, 690071. [CrossRef]
37. Moon, S.; Lee, G.; Chi, S. Semantic text-pairing for relevant provision identification in construction specification reviews. *Autom. Constr.* **2021**, *128*, 103780. [CrossRef]
38. Moon, S.; Chi, S.; Im, S.B. Automated detection of contractual risk clauses from construction specifications using bidirectional encoder representations from transformers (BERT). *Autom. Constr.* **2022**, *142*, 104465. [CrossRef]
39. Tian, D.; Li, M.; Shi, J.; Shen, Y.; Han, S. On-site text classification and knowledge mining for large-scale projects construction by integrated intelligent approach. *Adv. Eng. Inform.* **2021**, *49*, 101355. [CrossRef]
40. Qiao, J.; Wang, C.; Guan, S.; Shuran, L. Construction-accident narrative classification using shallow and deep learning. *J. Constr. Eng. Manag.* **2022**, *148*, 04022088. [CrossRef]
41. Zhang, F. A hybrid structured deep neural network with Word2Vec for construction accident causes classification. *Int. J. Constr. Manag.* **2022**, *22*, 1120–1140. [CrossRef]
42. Luo, Z.; Hirogane, M. Utilization of similar accident cases for safety education. In Proceedings of the 2022 Joint 12th International Conference on Soft Computing and Intelligent Systems and 23rd International Symposium on Advanced Intelligent Systems (SCIS&ISIS), Ise, Japan, 29 November–2 December 2022; pp. 1–4. [CrossRef]
43. Li, J.; Wu, C. Deep learning and text mining: Classifying and extracting key information from construction accident narratives. *Appl. Sci.* **2023**, *13*, 10599. [CrossRef]
44. Luo, X.; Li, X.; Song, X.; Liu, Q. Convolutional neural network algorithm–based novel automatic text classification framework for construction accident reports. *J. Constr. Eng. Manag.* **2023**, *149*, 04023128. [CrossRef]
45. Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pretraining of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
46. Park, S.; Moon, J.; Kim, S.; Cho, W.I.; Han, J.; Park, J.; Song, C.; Kim, J.; Song, Y.; Oh, T.; et al. Klue: Korean language understanding evaluation. *arXiv* **2021**, arXiv:2105.09680. Available online: https://arxiv.org/abs/2105.09680 (accessed on 10 June 2024).
47. Robinson, S.D.; Irwin, W.J.; Kelly, T.K.; Wu, X.O. Application of machine learning to mapping primary causal factors in self reported safety narratives. *Saf. Sci.* **2015**, *75*, 118–129. [CrossRef]
48. Kaya, G.K.; Ustebay, S.; Nixon, J.; Pilbeam, C.; Sujan, M. Exploring the impact of safety culture on incident reporting: Lessons learned from machine learning analysis of NHS England staff survey and incident data. *Saf. Sci.* **2023**, *166*, 106260. [CrossRef]
49. Alkaissy, M.; Arashpour, M.; Golafshani, E.M.; Hosseini, M.R.; Khanmohammadi, S.; Bai, Y.; Feng, H. Enhancing construction safety: Machine learning-based classification of injury types. *Saf. Sci.* **2023**, *162*, 106102. [CrossRef]
50. Rawson, A.; Brito, M.; Sabeur, Z.; Tran-Thanh, L. A machine learning approach for monitoring ship safety in extreme weather events. *Saf. Sci.* **2021**, *141*, 105336. [CrossRef]