Nuclear Engineering and Technology 54 (2022) 3336–3346

Contents lists available at ScienceDirect

Nuclear Engineering and Technology

journal homepage: www.elsevier.com/locate/net

Original Article

Machine learning-based categorization of source terms for risk assessment of nuclear power plants

Kyungho Jin, Jaehyun Cho*, Sung-yeop Kim

Korea Atomic Energy Research Institute, (34057) 111, Daedeok-daero 989, Daejeon, Republic of Korea

ARTICLE INFO

Article history:
Received 30 December 2021
Received in revised form
24 March 2022
Accepted 9 April 2022
Available online 13 April 2022

Keywords:
Level 2 probabilistic safety assessment
Source term category
Accident consequence
Autoencoder
Clustering

ABSTRACT

In general, a number of severe accident scenarios derived from Level 2 probabilistic safety assessment (PSA) are typically grouped into several categories to efficiently evaluate their potential impacts on the public with the assumption that scenarios within the same group have similar source term characteristics. To date, however, grouping by similar source terms has been completely reliant on qualitative methods such as logical trees or expert judgements. Recently, an exhaustive simulation approach has been developed to provide quantitative information on the source terms of a large number of severe accident scenarios. With this motivation, this paper proposes a machine learning-based categorization method based on exhaustive simulation for grouping scenarios with similar accident consequences. The proposed method employs clustering with an autoencoder for grouping unlabeled scenarios after dimensionality reductions and feature extractions from the source term data. To validate the suggested method, source term data for 658 severe accident scenarios were used. Results confirmed that the proposed method successfully characterized the severe accident scenarios with similar behavior more precisely than the conventional grouping method.

1. Introduction

Probabilistic safety assessment (PSA) is an effective tool for the risk assessment of nuclear power plants (NPPs). Since risk is defined as the product of frequency and consequence for a given scenario [1], the results of PSA for NPPs cover information on how often severe accidents occur and how the accidents impact the public. In order to evaluate such information systematically, PSA is divided into three levels [2]; Level 1 PSA focuses on whether a reactor core is damaged or not when an initiating event occurs. If damaged, the severe accident (SA) scenarios for the given core damage case are identified to quantify the probabilities of containment failures in Level 2 PSA. In addition, the types and amounts of radioactive materials released to the environment (i.e., source terms) are also evaluated in the same phase. In Level 3 PSA, the impacts on the public as the consequence of the accident (e.g., early or late fatality) are estimated using the source terms derived in Level 2 PSA. In sum, the risks of NPPs are composed of SA scenarios, accident frequencies, and accident consequences.

In particular, Level 2 PSA plays an important role in overall PSA as it addresses the source terms of accidents, which are closely related to the accident consequence, as well as the frequency of accidents [3]. For example, Level 2 PSA can provide surrogate metrics such as large early release frequency (LERF) [4] by identifying a number of SA scenarios based on a logical event tree approach. In addition, the amount of hazardous materials that will be released to the environment for a given SA scenario can also be estimated through source term analysis codes. As these elements cover all components of risk (i.e., scenario, frequency, and consequence), it is important to reduce the uncertainties in Level 2 PSA as much as possible to improve the overall risk quantification results for NPPs.

However, the results of Level 2 PSA in reality involve a lot of uncertainties due to the lack of data and to the reduction of model sizes to save analysis costs. For example, source term analysis for a limited number of SA scenarios has been performed to reduce the computational burden. In other words, before analyzing the source terms, a number of SA scenarios are grouped into several categories by qualitative logical trees or expert judgements to sort out similar source term characteristics without quantitative information [5,6] (Fig. 1, upper panel). After this categorization, source term analysis is implemented for only one representative scenario of each source

* Corresponding author.
E-mail address: chojh@kaeri.re.kr (J. Cho).

https://doi.org/10.1016/j.net.2022.04.006

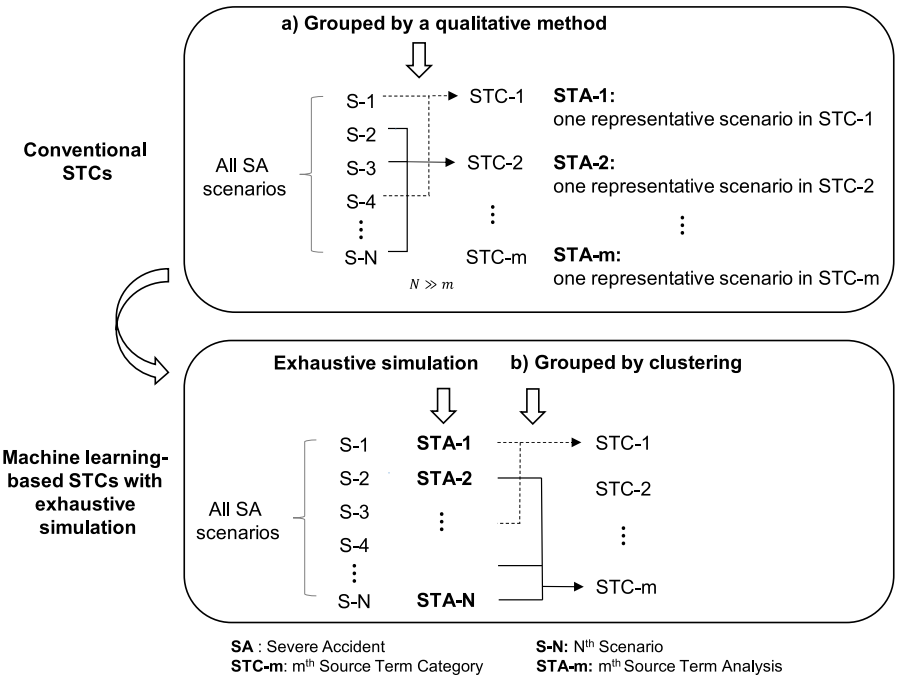


Fig. 1. Comparison between conventional STC grouping and the proposed STC grouping based on the exhaustive simulation approach.

term category (STC). Accident consequence analysis is likewise limited to one case because each STC has just one source term data.

For this reason, the categorization method and its performance are significantly important for risk quantification. Clearly, the higher the similarity between the scenarios within an STC, the more advantageous the method. However, as shown in Fig. 1, the current categorization scheme has a contradiction in that SA scenarios are grouped before analyzing their source terms. While the objective of the categorization is to make SA scenarios within the same category have similar source term characteristics, the problem is that there is no available data about the source terms while categorizing because the source term analysis is carried out after the categorization. If dissimilar scenarios end up in the same group from incorrect logical trees or expert judgments, the results of risk quantification will involve significant uncertainties. Unfortunately, no evaluation has been made on the grouping performance of these quantitative methods.

Recently, the Korea Atomic Energy Research Institute (KAERI) proposed an exhaustive simulation approach [7] to carry out source term analysis not for a limited number of SA scenarios but for all scenarios (Fig. 1, lower panel). For this purpose, a program called Module for Exhaustive Scenarios-based Severe Accident input Generation (MESSAGE) was developed. The MESSAGE program can automatically read Level 2 PSA event tree models regardless of their size and generate a large number of input files for a source term analysis code such as Modular Accident Analysis Program (MAAP5) [8]. With MESSAGE, we can now analyze the source terms for numerous SA scenarios instead of just one representative scenario.

Thanks to such exhaustive simulation, a source term database for a large number of SA scenarios can be constructed. Hence, it is now possible to quantitatively categorize SA scenarios into several groups having similar characteristics. In Level 3 PSA, categorizing source terms is also beneficial for analyzing and interpreting a large amount of data more efficiently. With the source term database and quantitative categorization method, it is expected that more realistic results of risk assessment can be derived because similar source term characteristics are likely related to similar accident consequences.

Therefore, with the source term database, this paper proposes a machine learning-based categorization method to improve grouping performance and handle a large number of data efficiently (Fig. 1, lower panel). Because the source terms have no specific distinction for categorization, a clustering approach, which is a typical unsupervised learning method in the field of machine learning [9,10], was employed. Clustering is a method of grouping by determining the similarity between data when data labeling is not given. To improve the quality of clustering, an autoencoder structure [11–13] was introduced to perform a dimensionality reduction of the time series data before clustering.

This paper is organized as follows. Section 2 describes the limitations of conventional categorization and introduces the machine learning-based categorization method. Section 3 shows the application results of the proposed method to the Optimized Power Reactor 1000 (OPR1000) Level 2 PSA model. Section 4 and Section 5 present our discussion and conclusion, respectively.

2. Categorization of source terms in level 2 PSA

2.1. Limitations of the conventional categorization method

In general, the risk of NPPs can be quantified using the frequency and consequence of each SA scenario as follows:

$$R_{Total} = \sum_{i=1}^{N} F_i \times C_i, \tag{1}$$

where R_{Total} is the total risk of the target NPP, and F_i and C_i are the frequency and consequence of the i-th SA scenario, respectively. SA scenarios can be identified from the core damage scenarios in Level 1 PSA considering SA progression. The number of SA scenarios N is generally in the thousands to several hundred thousand. Once the SA scenarios are identified, F_i can be evaluated using the probability of SA phenomena and the given core damage frequency.

Prior to the quantification of consequences, source terms for a given SA scenario are first evaluated through a source term analysis code. If the types and amount of radioactive material released to the environment are determined in Level 2 PSA, C_i can be estimated by considering how the hazardous materials are dispersed and how the public will evacuate in Level 3 PSA.

As mentioned in the Introduction, conventional PSA has employed the concept of grouping similar scenarios for saving analysis costs. In other words, source term analysis for a limited number of SA scenarios has been performed. For N SA scenarios, the conventional categorization method constructs a qualitative logical tree to assign similar scenarios to specific STCs based on the SA process and expert judgements, as shown in Fig. 2. In Ref. [14], the key attributes in specifying STCs are summarized. Here, the numbers of logics and categories completely depend on the experts.

Each STC in Fig. 2 can include at least one SA scenario. After this categorization, the total N SA scenarios are grouped into m STCs. The scenario with the highest frequency of each STC is generally selected as the representative scenario and their source terms are evaluated. Therefore, m analyses of source terms are performed. In this case, the approximated risk of NPPs can be estimated using Eq. (2):

$$R_{Total} = \sum_{k=1}^{m} (C^k \times \sum_{j=1}^{v_k} F_j^k), \tag{2}$$

where v_k is the number of SA scenarios in the k-th STC, therefore $\sum_{k=1}^{m} v_k = N$. C^k is the consequence of the representative scenario in the k-th STC, and F_j^k is the frequency of the j-th SA scenario in the k-th STC. This categorization approach has the advantage of being able to obtain an approximation of the risk with a smaller number of source terms/consequence analyses if the scenarios with similar accident consequences are well characterized. Accordingly, the categorization method has played an important role in reducing the uncertainties in risk quantifications.

Despite its advantages, the conventional categorization method shown in Fig. 2 involves a lot of uncertainties for risk quantification via Eq. (2) because it labels scenarios in a qualitative way without quantitative source term data. The categorization is therefore inevitably limited because each logic explains only a few discrete branches. If the logic itself is designed incorrectly, the similarity among SA scenarios within an STC cannot be guaranteed.

According to exhaustive simulation [7], it was discovered that the conventional categorization method caused significant differences in the release amount of Cs-137 within the same category. Cs-137 is known as a major contributor to off-site consequences [15], and therefore it can be said that the results of risk quantification using Eq. (2) with the conventional method are uncertain. In order to visualize the uncertainties in the conventional categorization method, the time-dependent release fractions of a radioactive material, which is closely related to the Cs-137 release amount, are represented in Fig. 3 using the source term data derived from exhaustive simulation.

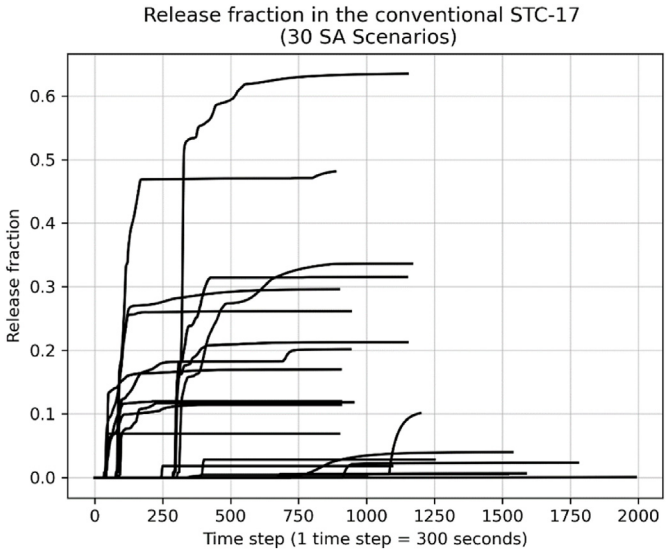


Fig. 3. Time-dependent release fraction characteristics in the conventional STC-17.

Fig. 3 shows the release fractions of the SA scenarios belonging to the last category out of a total of 17 categories. A total of 30 SA scenarios are included in STC-17. As shown in Fig. 3, it turned out that the conventional method incorrectly labeled the SA scenarios. Comparing the end points between the scenarios within STC-17, one can see that they are totally different. Moreover, they also differ in the initial release time, which is another key factor in determining off-site consequences [2]. If the analysis of off-site consequences is performed in such a situation where there is a large deviation in the major attributes within a group, it is clear that the final risk assessment result using Eq. (2) may contain considerable uncertainty.

2.2. Construction of a source term database

2.2.1. Source term database using the exhaustive simulation

To construct a source term database for a large number of SA scenarios, KAERI proposed the concept of exhaustive simulation [7]. This approach aims to support source term analysis through an automated program called MESSAGE, which can read and interpret Level 2 PSA models and automatically generate numerous input

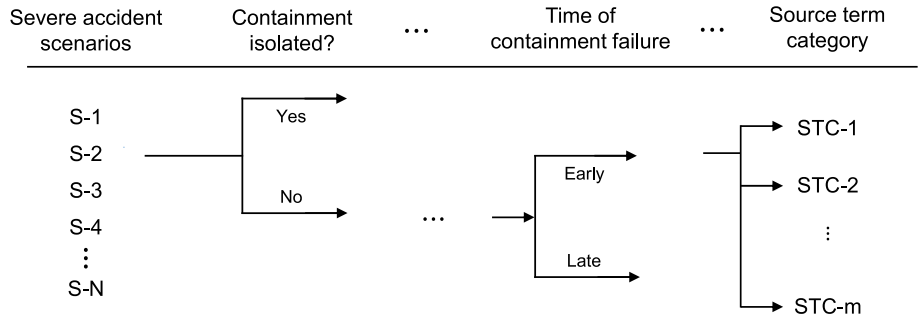


Fig. 2. Example of the conventional categorization of source terms (N ≫ m).

Table 1
Summary of the source term database from exhaustive simulation.

No. SA scenarios	$N = 690$
Notation of *i*-th SA scenario	S_i, $i = 1, 2, 3, \ldots, N$
Type of data	Time-dependent release fraction, x_{t_i}
No. variables	$d = 18$
Sequence length	Variable length (t_i) (Min: 574, max: 2194)

$$\boldsymbol{S_i} = \begin{bmatrix} \boldsymbol{x}_1^{(1)} & \cdots & \boldsymbol{x}_1^{(18)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}_{t_i}^{(1)} & \cdots & \boldsymbol{x}_{t_i}^{(18)} \end{bmatrix}$$

files for MAAP5 suitable for each SA scenario. The MESSAGE program is composed of the following: (i) an input library module that reads the Level 2 PSA model, (ii) a scenario identification module that sorts out the SA scenarios with non-zero frequency, and (iii) an input generation module that generates input files based on the SA scenarios.

From exhaustive simulation with the MESSAGE code, KAERI constructed a source term database for the OPR1000. Specifically, the source terms for a total of 690 SA scenarios were analyzed. The reason why the number of analysis scenarios was selected as 690 is that these selected SA scenarios occupied 99% of total accident frequency. In other words, SA scenarios with very low frequency that seldom contribute to risk quantification were excluded. The selected SA scenarios were analyzed for 72 h, with the source term data of each scenario consisting of the time-dependent release fraction of 18 radioactive materials. It should be noted that each scenario has a different sequence length. Table 1 shows a summary of the source term database constructed using exhaustive simulation.

2.2.2. Consequence analysis with the source term database

While the source terms for numerous SA scenarios can be evaluated through the exhaustive simulation approach, for risk quantification, it is essential to perform accident consequence analysis with the obtained source term data. It seems desirable to analyze the consequences for numerous SA scenarios to reduce the uncertainty of the quantification results as much as possible. For this purpose, KAERI developed automation programs for consequence analysis in Level 3 PSA called Multi Unit Source Term (MUST) converter [16] and Mr. (Multi-run) manager [17], similar to the MESSAGE code developed for Level 2 PSA. With these programs, more accurate risk can be quantified using Eq. (1).

Although consequence analysis for a large number of scenarios can now be accomplished with the dedicated programs, it becomes more difficult to efficiently gain risk insights from the massive amounts of data as the number of SA scenarios increases. In this context, categorizing similar scenarios is still useful to carry out comprehensive PSA incorporating Level 3 PSA. If the similarity of the accident consequence between scenarios is guaranteed, the uncertainty in Eq. (2) can be reduced. For that reason, this paper proposes a quantitative categorization method by introducing a machine learning technique to enhance the grouping performance and effectively handle large amounts of data.

2.3. Machine learning-based source term categorization

In order to quantitatively group SA scenarios using the source term database, two methods can be considered: classification or clustering. Classification, which is a typical supervised learning method, is used when the number of classes is known. On the other hand, if data labels are not predefined, an unsupervised learning method such as clustering should be employed. Clustering enables grouping based on similarity between data without specific distinctions.

However, as described in Table 1, the source term data is multivariate and time-dependent. It is difficult to discriminate such complex data using conventional qualitative methods such as a logical tree. Even though a quantitative grouping method is now available, the high dimensionality of the data may lead to the degradation of grouping performance. This is known as the curse of dimensionality.

This paper employs a clustering algorithm since the source term data have no specific labels (i.e., it is not known which category the nth SA scenario should belong to). In addition, a dimensionality reduction technique is also incorporated into the proposed framework. Specifically, an autoencoder, known as an unsupervised learning method, is introduced before clustering. Fig. 4 shows the proposed framework for machine-learning-based source term categorization.

When SA scenarios are successfully categorized by the proposed method, the similarity of their accident consequence can also be guaranteed. If so, the uncertainty in risk quantification using Eq. (2) can be reduced.

2.3.1. Autoencoder for dimensionality reduction and feature extraction

The source term data is composed of the time-dependent release fraction of the radioactive materials. In order to perform clustering using such time series data, two major methods can be applied. The first is to utilize the time series data as it is to determine the similarity between scenarios and group them through a clustering algorithm based on this similarity. The Euclidean distance and dynamic time warping (DTW) [18] methods are widely used to calculate the similarity between time series data. In particular, DTW is commonly used when the sequence length is different. In this case, the similarity can be calculated without data loss; however, it is difficult to calculate the similarity between data when the time series data is of high dimensionality. This is known as the curse of dimensionality. In addition, calculating the similarity of data with high dimensionality via DTW takes a considerable amount of time.

One way to avoid the curse of dimensionality is to reduce the data dimension and extract the major features from the time series data. Clustering is then performed on this compressed data. Although this method is likely to involve some information loss while compressing the data, the similarity can be simply calculated using the Euclidean distance, and it is more efficient to gain good clustering results rather than using high dimensionality data as it is [19].

In this paper, a dimensionality reduction method is employed to effectively analyze the complex source term data. It should be noted that there are various dimensionality reduction methods, such as principal component analysis (PCA). Since this paper does not focus on finding the optimal dimensionality reduction method but rather on proposing a source term categorization framework, an autoencoder structure, which is widely used to reduce the dimensions of multivariate time-series data, is simply adopted. An autoencoder is an ANN-based structure composed of an encoder and decoder. While most applications of ANN-based structures are developed for supervised learning, an autoencoder is a typical unsupervised learning method because the decoder reconstructs the original input data. Through unsupervised learning, it encodes the input time series data and decodes it to reconstruct the encoded data. After training the autoencoder, only the encoder is used to generate the compressed data. Dimensionality reductions and feature extractions are implemented in the encoder. Fig. 5 shows an example of an autoencoder with ANN structure.

In the autoencoder in Fig. 5, the input layer receives time series data (S_i) as input. This sequential data is compressed while passing through the hidden layers. It is common to set the number of nodes

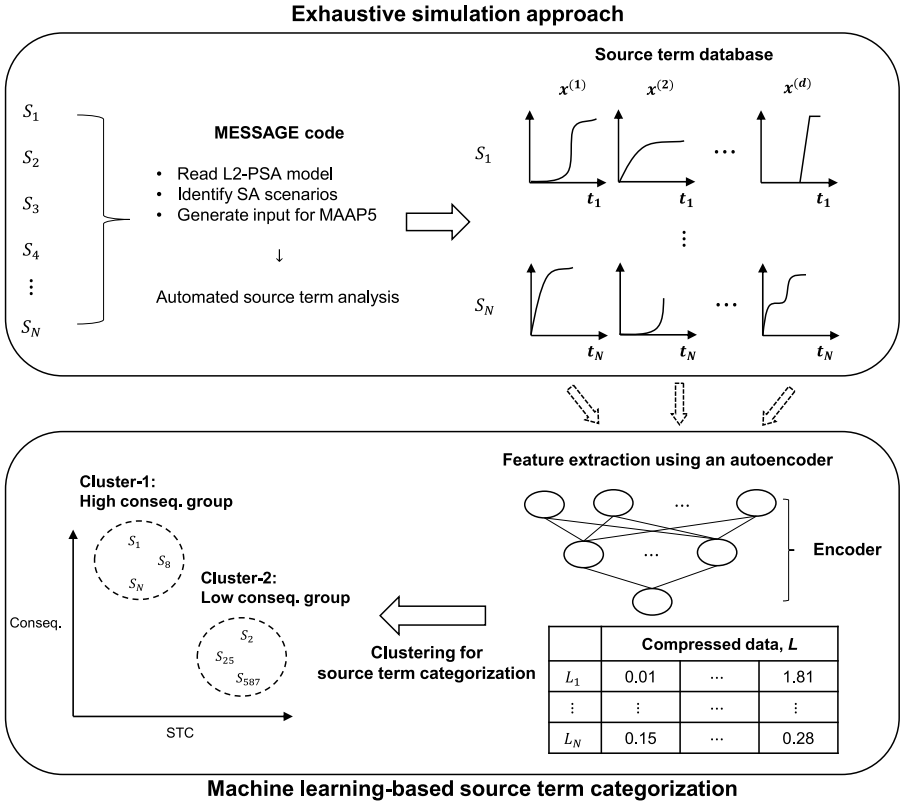


Fig. 4. Framework of machine learning-based source term categorization.

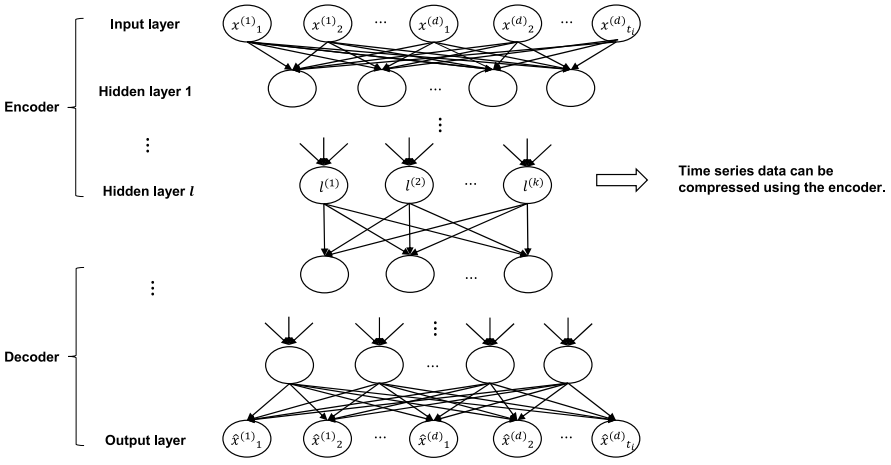


Fig. 5. Example of the autoencoder structure.

in the current hidden layer to a smaller number than that of the previous layer. Then in the final hidden layer of the encoder, the number of nodes k is defined by the desired dimension. The dimension of the original time series data $(N \times t \times d)$ can be reduced to $(N \times k)$ after encoding. The compressed data can be represented as follows:

$$L = \begin{bmatrix} l_1^{(1)} & \cdots & l_1^{(k)} \\ \vdots & \ddots & \vdots \\ l_N^{(1)} & \cdots & l_N^{(k)} \end{bmatrix} \quad (3)$$

The decoder receives the dimensionally reduced data L, which is the output of the encoder, and reconstructs the inputs $\widehat{x_1^1}, \widehat{x_2^1}, \ldots, \widehat{x_{t_i}^d}$. The number of nodes and layers in the decoder is usually symmetric with the encoder.

The autoencoder can compress time series data and reduce the dimension while learning the weights of the layers so that the input and 006Futput are derived identically. The number of layers and nodes, which are hyper-parameters, in the encoder and decoder is naturally determined while training the model. It should be noted that the compressed data L contains the key features of the time

series data. In other words, it is sufficient to describe the source term data of the 1st scenario S_1 in terms of $L_1 = \{l_1^{(1)}, l_1^{(2)}, \ldots, l_1^{(k)}\}$ because the key characteristics of the time series data, such as the final release fraction or the initial release point, are well summarized in L_1. Detailed descriptions on how to train the autoencoder model are omitted in this paper because it is equal to fitting an ANN structure, which is well established [10,20].

2.3.2. Clustering for categorization of source terms

Once the dimension of the time series data is successfully reduced using the autoencoder, SA scenarios with similar behavior can be easily characterized through the clustering method. Clustering is a way of grouping similar data based on their similarity or dissimilarity without a specific distinction. There are various methods of clustering according to how to calculate the similarity and how to group the similarity data [21,22]. For example, hierarchical clustering, which is a connectivity model, and the k-means algorithm as a partition model are widely used for clustering. K-medoids, also known as partitioning around medoids (PAM), is also popular because it is not sensitive to outliers. The PAM clustering algorithm used in Section 3 of this paper is briefly introduced as follows.

PAM is a partitional clustering method based on medoids. While the k-means algorithm uses the mean distance of the data points within a cluster as a centroid, PAM selects a real data point as a medoid. When the number of clusters n_c is determined, n_c data points are randomly selected as the medoids. After calculating the similarity between the data points and the medoids, each data point is assigned to its nearest medoids. Next, new medoids, which are not identical to the previous medoids, are selected. The similarities between the data points and the new medoids are calculated again and the cost of each medoid is evaluated. These steps are repeated until the difference in the cost does not decrease. The clustering procedure with the compressed data L and PAM is summarized in Table 2.

As shown in Table 2, most clustering algorithms require n_c before performing clustering because of unsupervised learning. In the same context, it is generally necessary to apply various clustering methods to the same data to determine the optimal clustering method, although only the PAM algorithm is introduced here. There are various measures or methods to find out the optimal clustering algorithm [23]. For example, the elbow method is widely used to qualitatively find the optimal number of clusters. The Davies–Bouldin index or Dunn index can also be used to quantitatively measure clustering performance. Another approach is to evaluate the silhouette score [24]. This index measures the similarity of data points for the cluster to which they belong compared to other clusters. The silhouette score of i-th data Sil_i can be quantified by:

$$Sil_i = \frac{b_i - a_i}{max(b_i, a_i)}, \tag{4}$$

where b_i is the minimum average distance of the i-th data point to other clusters and a_i is the average distance within the cluster to which it belongs. This score is between $[-1, 1]$. The closer the value is to 1, the better the clustering is. On the other hand, a score of -1 indicates that the data belongs to the wrong cluster. The average score of all Sil_i can be interpreted as the performance of clustering.

Table 2
PAM clustering with compressed data L obtained from the autoencoder.

Step	PAM algorithm
I	Determine n_c
II	Randomly select n_c data points as medoids $U = \{u_1, u_2, \ldots, u_{n_c}\}$
III	Calculate the similarities between L and U Assign L to its nearest medoid based on the similarities
IV	Select new medoids U_{new} Calculate the similarities between L and U_{new} Assign L to its nearest medoid based on the similarities
V	Calculate the cost[1], C If $C_U - C_{U_{new}} < 0$, then U is replaced with U_{new}
VI	Repeat (IV) to (V) until the medoids do not change

1) The cost in step V can be calculated by summing the similarity of each medoid.

3. Application results for the OPR1000

This section describes the major results of an application of both the conventional grouping approach and the clustering approach to an OPR1000 full-power internal event PSA model. The PSA models are taken from multi-unit PSA research by KAERI [6,25]. Section 3.1 summarizes the results from the conventional grouping approach, and Section 3.2 describes the application results of the clustering approach suggested in this study.

3.1. Conventional grouping approach

Note that the results described in this section were taken from Refs. [7]. Fig. 6 summarizes the STC information using the conventional grouping approach. A total of 17 STCs and 7 containment failure modes were identified: ECF, LCF, BMT, CFBRB, NOISO BYPASS, and ISLOCA. The conditional probability of containment failure given core damage is 41%. Among the containment failure modes, LCF, CFBRB, and BYPASS are the highest contributors to the containment failure at 17%, 13%, and 8%, respectively.

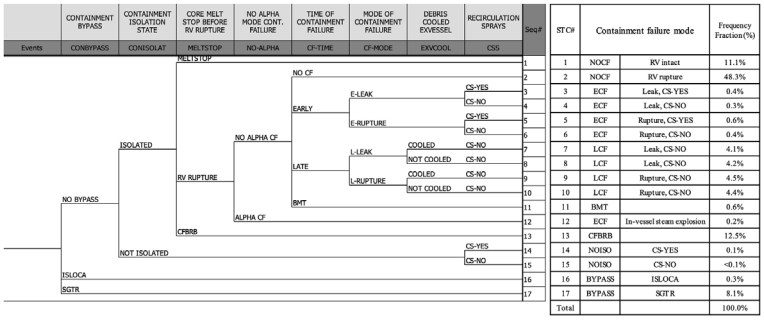
Using the exhaustive simulation results, distributions of the amount of Cs-137 release for the 17 STCs were obtained as shown in Fig. 7. Considering that there were significant differences among the data from STCs 6, 11, and 17, it can be said that the conventional grouping approach does not combine sufficiently similar scenarios.

3.2. Clustering for categorization of source terms

In this section, the application results of source term categorizations for OPR1000 PSA models are described based on the proposed method with the source term database summarized in Table 1. It should be noted that a total of 658 SA scenarios were used in this paper, excluding 32 SA scenarios that are difficult to use for clustering (e.g., most of them have a value of zero).

Before reducing the dimensionality of the source term data with the autoencoder, pre-processing was carried out so that the lengths of the time series data are identical. The sequence length of all scenarios was preprocessed by padding the last release fraction value of each scenario to the sequence of the maximum length because the release fraction always increases over time. Therefore, all t_i after preprocessing have a maximum length of 2194.

According to Refs. [15,26], Cs-137 is known as a major contributor to accident consequence analysis. Variables 2, 6, and 16 among the 18 variables in the source term data are closely related with the release amount of Cs-137; these variables include the release fractions of CsI, CsOH, and Cs_2MoO_4, respectively [26]. Therefore, instead of using all variables, this paper used these three variables to make the clustering result more highly associated with not only source terms but also consequences. This approach can also lead to a simplification of the autoencoder structure. Accordingly, the number of nodes in the input layer of the autoencoder was 6,582 (2194 x 3).



STC#	Containment failure mode		Frequency Fraction (%)
1	NOCF	RV intact	11.1%
2	NOCF	RV rupture	48.3%
3	ECF	Leak, CS-YES	0.4%
4	ECF	Leak, CS-NO	0.3%
5	ECF	Rupture, CS-YES	0.6%
6	ECF	Rupture, CS-NO	0.4%
7	LCF	Leak, CS-NO	4.1%
8	LCF	Leak, CS-NO	4.2%
9	LCF	Rupture, CS-NO	4.5%
10	LCF	Rupture, CS-NO	4.4%
11	BMT		0.6%
12	ECF	In-vessel steam explosion	0.2%
13	CFBRB		12.5%
14	NOISO	CS-YES	0.1%
15	NOISO	CS-NO	<0.1%
16	BYPASS	ISLOCA	0.3%
17	BYPASS	SGTR	8.1%
Total			100.0%

Fig. 6. STC results using the traditional grouping approach.
NOCF: no containment failure, ECF: early containment failure, LCF: late containment failure, BMT: basement melt-through, CFBRB: containment failure before reactor vessel breach, NOISO: containment isolation failure, BYPASS: containment bypass, ISLOCA: interfacing system loss of coolant accident (taken from Refs. [7]).

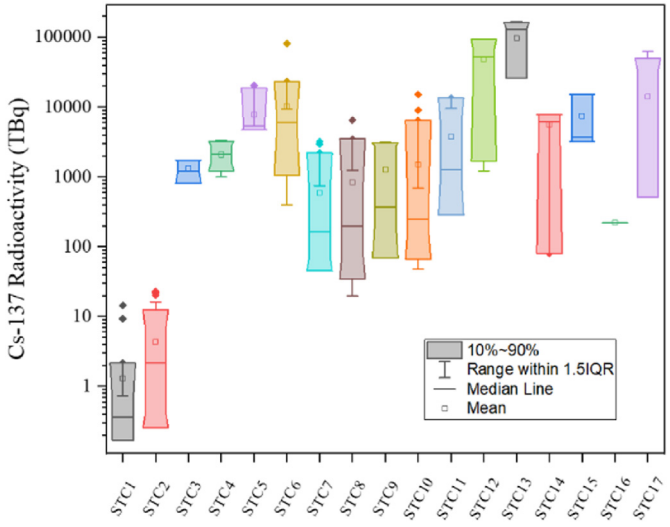


Fig. 7. Box plot results for the amount of Cs-137 release for 17 STCs defined from conventional grouping approach (taken from Refs. [7]).

Table 3
Average silhouette scores by clustering method.

	Hierarchical clustering	K-means clustering	PAM clustering
Avg. silhouette score	0.77	0.69	0.79

Table 4
Configurations and hyper-parameters used in the proposed method.

	Configurations or hyper-parameters
Encoder	Input layer: 6,582 nodes Hidden layer 1: 3,000 nodes Hidden layer 2: 500 nodes Hidden layer 3: 5 nodes (Output layer of the encoder)
Decoder	Hidden layer 4: 5 nodes Hidden layer 5: 500 nodes Hidden layer 6: 3,000 nodes Output layer: 6,582 nodes (Reconstruction of inputs)
Dimension of L	5
Optimization method	Adam (learning rate = 5.0E-05)
Clustering method	PAM algorithm
Similarity	Euclidean distance
Number of clusters	17

When compressing the input data to obtain the key features of each scenario, it is essential to determine the dimension size k of the compressed data L. If k is too high, the performance of clustering may deteriorate due to the curse of dimensionality. Conversely, the compressed data cannot capture the key features of the time series data when k is too low. Therefore, k was empirically determined in this paper so that the output layer of the encoder has 5 dimensions.

In terms of Eq. (2), uncertainties of risk quantification definitely decrease as the number of categories increases. Therefore, it can be assumed that finding the optimal number of clusters in a given problem is not critical to the results of categorization; the higher the number of clusters, the better the risk quantification results. For this reason, the number of clusters n_c was simply set to 17, which is equal to the number of conventional STCs for a comparison of the results. A sensitivity analysis depending on n_c is discussed in Section 4. In order to determine the optimal clustering method, the silhouette scores for three clustering methods were evaluated at $n_c = 17$.

Based on the results shown in Table 3, the PAM algorithm was employed as the clustering method in the current study. It should be noted that the optimal clustering method can vary depending on the number of clusters or the means to reduce the dimension of the data. Table 4 summarizes the key configurations and hyper-parameters used for the applications.

Before examining the similarity of accident consequences within a group, the similarities of the raw data were confirmed. Fig. 8(a)-(c) show the release fraction of variable 2 (CsI) over time for machine learning-based STC-d, g, h, respectively. It should be noted that while both methods have the same number of categories, there are no relationships between the categories in the conventional method and the proposed method. For this reason, the proposed STCs were denoted in alphabetical order to clearly compare the results.

The numbers in Fig. 8 indicate the scenario number for a total 658 scenarios. Compared to Fig. 3, it can be seen that the grouping was well performed as the scenarios show a similar behavior for each STC. The scenarios belonging to each STC have a similar release amount (end point) and initial release point. In addition, the source term behaviors among SA scenarios are clearly discriminated. To quantitatively compare the categorization results, the average of the silhouette scores for the total 658 SA scenarios was evaluated using the conventional STC and proposed STC method through Eq. (4), as shown in Table 5. Note that the silhouette value was scored from the compressed data L for both methods.

From the result in Table 5, it can be concluded once again that the conventional method should be improved since its averaged silhouette score is negative. Compared to the conventional method,

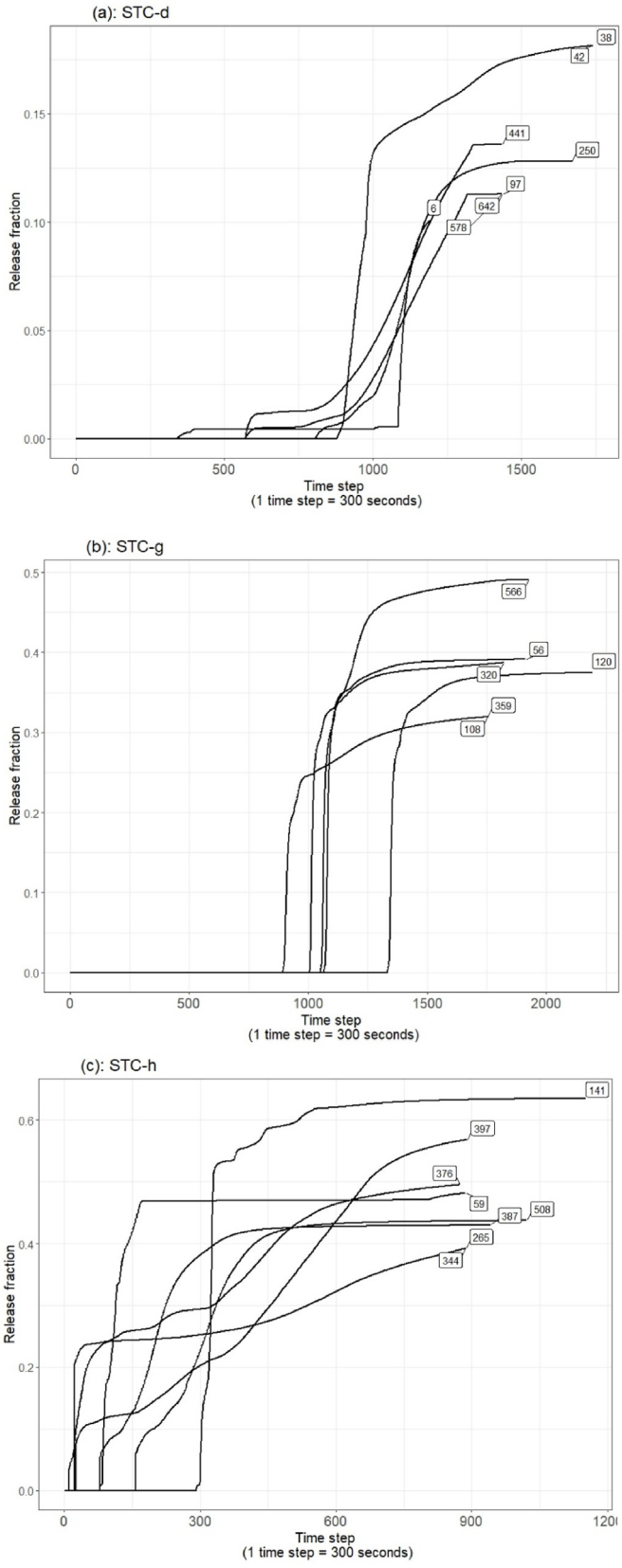


Fig. 8. Results of machine learning-based STC: release fract ion of variable 2 in (a) STC-d, (b) STC-g, (c) STC-h.

the proposed method well categorized the source term data. It can be concluded that the proposed method provides enough resolution to discriminate the source term behaviors since the average silhouette score is about 0.8. Based on the categorization results, the similarities of the accident consequences were confirmed, as presented in the next section.

3.3. Validation of clustering results with consequence analysis

Although it was confirmed that similar source terms were well characterized in Section 3.2, it is also necessary to verify the similarity of the accident consequences within the groups to reduce the uncertainty of risk quantification through Eq. (2). For verification, five scenarios for each category were randomly selected and their accident consequences were analyzed through MACCS [27]. In cases where the number of scenarios belonging to each category was less than five, accident consequence analysis was performed for the number of scenarios of that category.

Reference site installing OPR1000 was considered in the MACCS analyses incorporating emergency responses such as sheltering, evacuation, and dose-dependent relocation. Long-term health effects using CHRONC module were also evaluated and included in the estimation of total consequence results. Similar to the consequence metrics of state-of-the-art consequence analyses (SOARCA) project [28], population weighted individual risk (PWIR) of cancer fatality is employed as a consequence metrics since early fatality seldom occurs in the offsite consequence analysis of OPR1000.

Fig. 9 shows the results of the similarity of accident consequences for the 17 STCs depending on the categorization method, i.e., conventional STC and the proposed method. In the figure, the values of the PWIR are normalized between zero to one; they are expressed as red points along with the scenario number assigned to each STC, with the minimum and maximum values also indicated.

The numbers in Fig. 9 indicate the scenario number for a total 658 scenarios. As Fig. 9 confirms, the conventional STC method does not guarantee similarity of the accident consequences within a group. Especially, STC-6, 12, 13, and 17 include large deviations of consequences, which means that evaluating risks using a representative scenario with Eq. (2) can lead to significantly uncertain results. On the other hand, the proposed method categorized similar scenarios with less deviation than the conventional one. The reason why there are overlap regions in the results of the proposed method is that categorization was not based on the accident consequence but source terms. In order to confirm the uncertainty of both categorization methods, the results of risk quantification through Eq. (2) are compared in the next section based on the results in Fig. 9.

4. Discussions

In order to quantitatively compare the grouping performance between STC methods, a metric named grouping convergence index (GCI) is newly defined in this paper using Eq. (2) as follows:

$$GCI = \frac{\sum_{k=1}^{m}\sum_{j=1}^{\nu_k} F_j^k \times C_{max}^k}{\sum_{k=1}^{m}\sum_{j=1}^{\nu_k} F_j^k \times C_{min}^k}, \tag{5}$$

where ν_k is the number of SA scenarios in the k-th STC, F_j^k is the frequency of the j-th SA scenario in the k-th STC, C_{min}^k and C_{max}^k is the minimum and maximum consequence in the k-th STC as the representative scenario, respectively. If GCI indicates 1, it can be said that the result of risk quantification using Eq. (2) does not include uncertainty due to grouping because there is no deviation in the accident consequence within the group. In other words, the larger the GCI value, the worse the grouping performance. Table 6

Table 5
Average silhouette score by STC approach.

	Conventional STC	Proposed STC
Avg. silhouette score	−0.44	0.79

shows the estimation of GCI depending on the STC methods with the normalized PWIR results derived in Section 3.3 to compare the grouping performance.

As Table 6 shows, the results are significantly sensitive to the representative scenario when the risks were approximated based on the conventional STC method. The risk using the maximum consequence was approximately 14 times the risk using the minimum consequence. On the other hand, the proposed method provided less sensitive results. In this case, the GCI was calculated as 1.44. The two GCI differ by about 10 times. Although the approach still employs categorization to evaluate the approximated risk, the proposed method can obtain more robust and less uncertain results than the conventional method.

As described in Section 3.2, it is obvious that the higher the number of clusters, the better the risk quantification results. To confirm this, the risk quantification results depending on the number of clusters were calculated and tabulated in Table 7. Note that only the number of clusters was changed in the calculation with the same procedure as in Table 6.

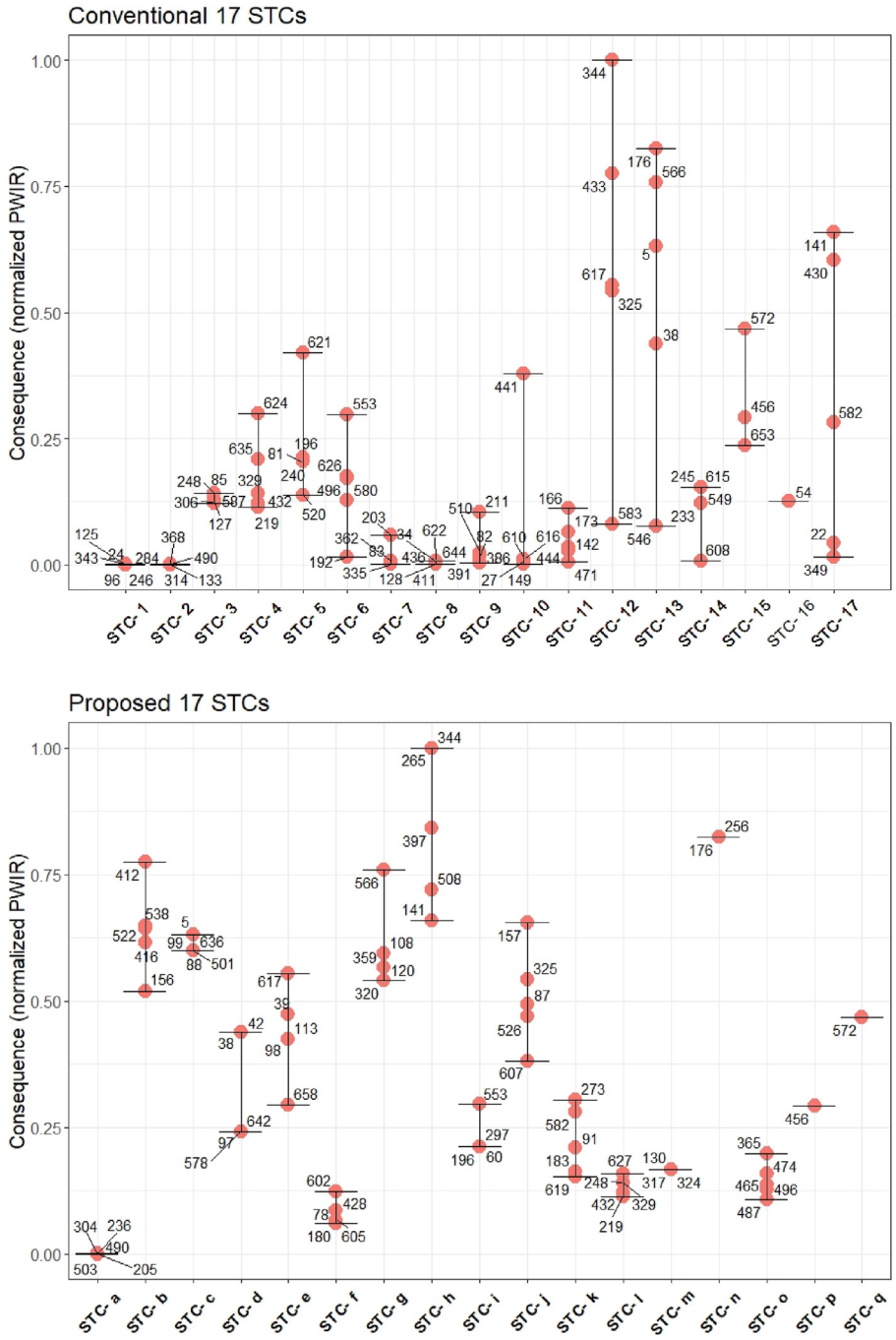


Fig. 9. Results of consequences analysis grouped by the conventional method (upper) and the proposed method (lower).

Table 6
GCI to compare the grouping performance by STC method.

	Scenario index (S_i)		GCI
Conventional STC	$C^1_{max}, C^2_{max}, \ldots, C^{17}_{max}$	24, 368, 85, 624, 621, 553, 203, 622, 211, 441, 166, 344, 176, 245, 572, 54, 141	14.0
	$C^1_{min}, C^2_{min}, \ldots, C^{17}_{min}$	96, 314, 127, 219, 520, 192, 335, 128, 391, 149, 471, 583, 546, 608, 653, 54, 349	
Machine learning-based STC	$C^a_{max}, C^b_{max}, \ldots, C^q_{max}$	304, 412, 5, 38, 617, 602, 566, 344, 553, 157, 273, 627, 317, 256, 365, 456, 572	1.44
	$C^a_{min}, C^b_{min}, \ldots, C^q_{min}$	503, 156, 88, 97, 658, 180, 320, 141, 196, 607, 619, 219, 317, 256, 487, 456, 572	

Table 7
Average silhouette score and GCI by number of clusters n_c.

No. clusters	7	17	27
Avg. silhouette score	0.78	0.79	0.79
GCI	2.59	1.44	1.37

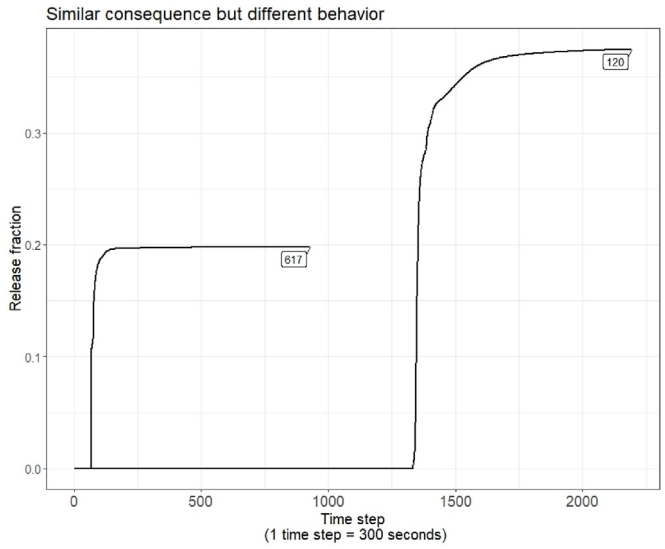


Fig. 10. Plot of scenarios that have similar consequence but different behavior.

It can be seen that the averaged silhouette score did not significantly change as the number of clusters increased. However, the GCI plainly decreased. If n_c is 658 which is equal to the number of SA scenarios, the GCI will definitely indicate 1. While it is desirable to allocate a large number of clusters to reduce uncertainty due to grouping, it is also important to select an appropriate number of clusters to reduce the computational burden.

On the other hand, it is a well-known fact that it is difficult to interpret and analyze the results as the amount of data increases. The proposed method may contribute to this legacy issue. One of the advantages of using the proposed method is that intuitive and useful information can be extracted from a large amount of data by identifying the features of each category. For example, the characteristics of the source terms of each category in Fig. 8 are apparently discriminative. It is easy to know that the SA scenarios in STC-g will release more source terms rather than those in STC-d. While the accumulated release amounts of the source terms in STC-g and h are similar, SA scenarios belonging to STC-h will initially release the source terms.

Furthermore, in Fig. 9, take for example that S_{617} in STC-e and S_{120} in STC-g belong to different categories but they have similar consequence results. The release fraction of variable 2 (CsI) of each scenario can be plotted in Fig. 10.

As shown in Fig. 10, there may be cases where the behavior of the source term itself is different even if its accident consequences are similar. This is not meaningful in terms of risk quantification, but it should be noted that the ultimate purpose of PSA is not only to quantify the risk but also to find vulnerabilities for the risk reduction of NPPs. In this context, Fig. 10 can provide important information. For example, based on the fact that S_{617} releases the source terms at a much faster time than S_{120}, results can be used for decision-making to devise accident responses or to prepare mitigation measures. Using clustering, important information can be extracted through the intra-cluster characteristic analysis or inter-cluster characteristic analysis of a large amount of data.

5. Conclusion

This paper proposed a quantitative source term categorization method based on machine-learning techniques with the source term database constructed by exhaustive simulation. The proposed method employed an autoencoder structure to reduce the data dimension and extract the key features from the time series data. In addition, different data lengths were preprocessed to have the same sequence length, and three variables closely related with Cs-137 were sorted out to reduce the size of the autoencoder. Finally, the severe accident scenarios were categorized by the PAM clustering method for the feature data compressed through the encoder, and it was confirmed that grouping by scenario with similar source term behavior was well performed in Section 3.2.

To confirm that the proposed method successfully guarantees not only the similarity of the source term behavior but also that of the accident consequences, five scenarios for each STC were randomly selected and their consequences were analyzed using MACCS code. Fig. 8 showed the similarities of source term behavior within a category by the proposed method. The similarities of accident consequences within a category by STC method were qualitatively compared in Fig. 9. In addition, the grouping performance was quantitatively estimated by introducing GCI. As a result, it was confirmed that the proposed approach has low deviation between accident consequences within the same group. In terms of GCI, the proposed method showed about 10 times higher grouping performance than the conventional method. Furthermore, it was confirmed that grouping approach can provide important information through the intra or inter-cluster characteristic analysis of a large amount of data.

In this paper, dimensionality reduction was performed only through the autoencoder. It is necessary to find a more optimized method by comparing with additional techniques such as principle component analysis (PCA) or discrete wavelet transform (DWT). Furthermore, the proposed method should be verified using the source term data for various NPPs and more results of accident consequence analysis.

Declaration of competing 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.

Acknowledgement

This work was supported by the Ministry of Science, ICT, and Future Planning of the Republic of Korea and the National Research Foundation of Korea (NRF-2020M2C9A1061638).

References

[1] S. Kaplan, B.J. Garrick, On the quantitative definition of risk, Risk Anal. 1 (1981).
[2] OECD-NEA, CSNI Technical Opinion Papers No. 9 Level-2 PSA for Nuclear Power Plants, 2007.
[3] IAEA, Applications of Probabilistic Safety Assessment (PSA) for Nuclear Power Plants, IAEA-TECDOC-1200), 2001.
[4] EPRI, PSA Applications Guide, EPRI TR-105396s), 1995.
[5] U.S. NRC, Severe Accident Risks: an Assessment for Five U.S. Nuclear Power Plants, NUREG-1150), 1990.
[6] J. Cho, S.H. Han, D. Kim, H. Lim, Multi-unit Level 2 probabilistic safety assessment : approaches and their application to a six-unit nuclear power plant site, Nucl. Eng. Technol. 50 (2018) 1234–1245, https://doi.org/10.1016/j.net.2018.04.005.
[7] J. Cho, S.H. Lee, Y.S. Bang, S. Lee, S.Y. Park, Exhaustive simulation approach for severe accident scenarios, in: Transactions of the Korean Nuclear Society Virtual Autumn Meeting, 2021.
[8] EPRI, Modular Accident Analysis Program (MAAP) Version 5.03 - Windows, Fauske & Associates, Inc., 2014.
[9] J.A. Hartigan, Clustering Algorithms, Wiley, 1975.
[10] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, second ed., O'Rreilly Media, 2019.
[11] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (2006) 504–507, https://doi.org/10.1126/science.1127647.
[12] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P.-A. Manzagol, Stacked denoising autoencoders : learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408.
[13] D.P. Kingma, M. Welling, An introduction to variational autoencoders, Found. Trends Mach. Learn. 12 (2019) 307–392, https://doi.org/10.1561/2200000056.
[14] IAEA, Development and Application of Level 2 Probabilistic Safety Assessment for Nuclear Power Plants (Specific Safety Guide, No. SSG-4), 2010.
[15] N.E. Bixler, S. Kim, Performing a multi-unit level-3 PSA with MACCS, Nucl. Eng. Technol. 53 (2021) 386–392, https://doi.org/10.1016/j.net.2020.07.034.
[16] S.-Y. Kim, D.-S. Kim, Development of MUST (Multi-Unit source term) converter version 1.0, in: Transactions of the Korean Nuclear Society Virtual Autumn Meeting, 2021.
[17] S.-Y. Kim, Development of Mr. (Multi-run) manager version 1.0, in: Transactions of the Korean Nuclear Society Virtual Autumn Meeting, 2021.
[18] A. Kianimajd, M.G. Ruano, P. Carvalho, J. Henriques, T. Rocha, S. Paredes, A. Ruano, Comparison of different methods of measuring similarity in physiologic time series, IFAC-PapersOnLine 50 (2017) 11005–11010, https://doi.org/10.1016/j.ifacol.2017.08.2479.
[19] H. Zhang, T.B. Ho, Y. Zhang, M.-S. Lin, Unsupervised feature extraction for time series clustering using orthogonal wavelet transform, Informatica 30 (2006) 305–319.
[20] C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2011.
[21] P. Rujasiri, B. Chomtee, Comparison of clustering techniques for cluster Analysis, Kasetsart J-Nat. Sci. 43 (2009) 378–388.
[22] R. Gelbard, O. Goldman, I. Spiegler, Investigating diversity of clustering methods : an empirical comparison, Data Knowl. Eng. 63 (2007) 155–166, https://doi.org/10.1016/j.datak.2007.01.002.
[23] G. Brock, V. Pihur, S. Datta, S. Datta, clValid , an R package for cluster validation, J. Stat. Software (2008).
[24] P.J. Rousseeuw, Silhouettes : a graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65.
[25] D.S. Kim, S.H. Han, J.H. Park, H.G. Lim, J.H. Kim, Multi-unit Level 1 probabilistic safety assessment: approaches and their application to a six-unit nuclear power plant site, Nucl. Eng. Technol. 50 (2018) 1217–1233, https://doi.org/10.1016/j.net.2018.01.006.
[26] J. Cho, S.H. Han, Identification of risk-significant components in nuclear power plants to reduce Cs-137 radioactive risk, Reliab. Eng. Syst. Saf. 211 (2021), 107613, https://doi.org/10.1016/j.ress.2021.107613.
[27] J. Leute, F. Walton, R. Mitchell, L. Eubanks, MACCS (MELCOR Accident Consequence Code System) User Guide-Version 4.0, 2021. SAND2021-8998).
[28] U.S.NRC, State-of-the-Art Reactor Consequence Analyses (SOARCA) Report (NUREG-1935), 2021.