A data generation framework for extremely rare case signals

Thasorn Chalongvorachai, Kuntpong Woraratpanya *

Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Department and Organisation, 1 Chalong Krung 1 Alley, Lat Krabang, 10520, Bangkok, Thailand

**HIGHLIGHTS**

- Our implemented data generation framework can increase from a single sample to thousands of high-quality data.
- Unlike data augmentation, this framework uses signal comparison techniques to ensure the quality of output data.
- With the power of latent space and our data picker, this framework can yield numerous anomaly samples.

**ARTICLE INFO**

**Keywords:**
- Data generation
- Anomaly detection
- Signal processing
- Deep learning
- Imbalanced dataset
- Data augmentation

**ABSTRACT**

Unlike data augmentation, data generation for extremely rare cases is an approach that can spawn a significant number of high-quality samples based on very few original data. This could be useful in anomaly detection and classification tasks that have the limitation of publicly available datasets for research purposes. Though some other approaches have attempted to solve this problem, such as data augmentation techniques, there was nothing to ensure the characteristics of synthesized samples. Previously, we initiated a framework, called Data Augmentation and Generation for Anomalous Time-series Signals (DAGAT), that was in cooperation with important components: Data Augmentation, Variational Autoencoder (VAE), Data Picker (DP), Signal Fragment Assembler (SFA), and Quality Classifier (QC). And then, an upgraded framework, called An Advanced Data Generation for Anomalous Signals (ADGAS), was introduced to eliminate the limitations of DAGAT; those are uncontrollable outputs and the possibility of bad data included in a training set. By reforming DAGAT architecture, ADGAS achieves a better outcome of generated samples. Nonetheless, ADGAS could be improved through better SFA, DP, and QC. Hence, this paper proposed a Data Generation Framework for Extremely Rare Case Signals. The proposed framework is achievable in generating reliable data for various objectives. We challenged this framework by using the 1D-CNN to serve as the performance evaluator in multi-class abnormal classifications and using the water treatment and water distribution testbed (SWAT and WADI) as the real-world anomaly datasets. The result shows that it surpasses other baseline methods of anomaly data augmentation and data generation techniques.

1. Introduction

In the real world, some anomaly data are rare to find or collect, such as the high-value machinery breakdowns, the rare-to-find symptoms, and the dangerous event alarms. In addition, some data contain private information, thus making it inaccessible publicly. This situation is one of the causes that rare case incidents are always collected in the form of imbalanced datasets. Many studies [1, 2, 3, 4] demonstrate an approach to solve the imbalanced dataset problem by using clustering or outlier detection techniques. In a word, the approach learns anomaly data from the majority class and assumes any data, which does not match with normal characteristics as an anomaly class. Although this approach may fit with some applications, it has limitations in classifying multi-classes of anomalous incidents and in learning the real characteristics of anomalous events through the real data.

Besides, as mentioned above, another interesting approach to work with the imbalanced dataset problem was based on deep learning/machine learning like SGM-CNN [5] and the input doubling method [6]. The SGM-CNN was designed in such a way that convolutional neural network (CNN) was combined with the synthetic minority oversampling technique (SMOTE) and under-sampling for clustering based on the Gaussian mixture model (GMM). Another one was the input dou-
bling method using the support vector regression (SVR) with RBF kernel for applying to the limited number of medical field data.

However, one of the most powerful solutions to work with a deep learning model under the imbalanced dataset situation is data augmentation. Um et al. [7] have proposed a learning model in conjunction with various augmentation techniques for time series data in medical fields, specifically for Parkinson’s disease monitoring. The results revealed that the increased number of samples, by applying the rotation, rotation with permutation, and time-warping techniques, can significantly improve the classification efficacy of the CNN model. Nevertheless, the data augmentation approach may have a risk of inadequate quality of data, according to the following reasons: (i) there is space of possibility of synthetic data that have not been discovered for generating great varieties of syntactic samples and (ii) no methods can measure the similarity of an outcome to ensure its quality. Therefore, the augmented data from the augmentation methods might not contain any characteristics of the original input or, conversely, may have to be too similar. This can result in terrible synthetic data that cause the low accuracy of deep learning classification models. This phenomenon happened where the amount of anomalous data is very small, compared to normal samples, and leads to a bias during the learning process [8]. An example of this effect will be described in Section 2.

According to the mentioned issues and the best of our knowledge, there were no methods for generating rare datasets. Therefore, we have proposed an alternative way to solve the problem of anomaly data in rare cases. We initiated a framework, called DAGAT (Data Augmentation and Generation for Anomalous Time Series Signals) [9]. The outcomes of this framework are different from the traditional data augmentation in that the DAGAT can efficiently synthesize data for various and reliable output samples. This study investigated applying data augmentation on various domains to generate data via latent space of trained variational autoencoder (VAE), and at the same time, using a statistical method to evaluate the quality of generic data. This method can remarkably increase the quantity of data. However, the DAGAT still suffered from the possible bad quality data that can be included in a VAE’s training set, an uncontrollable number of outputs, and unable to integrate multiple augmentation characteristics into a single sample.

Our recent work, called ADGAS (Advanced Data Generation for Anomalous Signals) [10], was proposed. ADGAS can eliminate the limitations of DAGAT by reassembling the signal fragments with assembler (SFA) function before performing VAE training, creating a loop of data picker (DP) method, and providing an additional quality classifier. The evidence of our experiments revealed that ADGAS can control the number of outputs, unlike DAGAT, and can outperform the baseline augmentation techniques.

Notwithstanding that the ADGAS could be a strong framework for anomalous data synthesizing, it still has room for improvement to increase the quality of the rare datasets. Therefore, in this paper, we proposed a framework for extremely rare case data generation. The implemented model is inherited from the ADGAS concept. It was improved by dividing data into more portions, including additional methods for the quality classifier (QC), and redesign the data picker. This implemented model can achieve better experimental results in a multi-class anomaly classifier.

The contributions of this paper can be summarized as follows:

- The proposed method is used SFA with a four-fragment partition instead of a two-fragment partition as used in the ADGAS model. This modification achieves a great variety of generated samples.
- DP of the proposed method is redesigned by randomly selecting a pair of points in a latent space, and then a new sample is generated with linear interpolation of that pair.
- QC of the proposed method is implemented with weighted metric and signal comparison technique. This improvement makes it adaptable to various tasks.

The remainder of this paper is organized as follows: Section 2 describes problem formulation. Section 3 explains how this framework functions. The experimental results are reported and discussed in Section 4. Lastly, we conclude our work in Section 5.

2. Problem statement

To the best of our knowledge, datasets for extremely rare case signals suffered from the unavailability and inaccessibility, especially breakdown signals of the high-value machinery, signals of the dangerous event alarms, and even rare-to-find signals of the symptoms of diseases. This issue becomes a barrier for the research and development community and inspires us to surpass the barrier. The following subsections describe scenarios of datasets and tools from the problem to the feasible solution.

2.1. Anomaly incident problems

In the real world, there are various types of sensors installed on modern machinery, such as power plants for technical analysis, and even humans, such as cardiometers for monitoring patients. In some cases, these sensors can detect anomalous incidents that might happen due to the breakdown of machinery, the error of systems, some types of diseases, or the specific Internet attacks [11, 12]. Most of the time, these anomaly events will lead to the loss of machinery or systems. In a worst-case scenario, a serious system failure can lead to the loss of people’s lives. Hence, many researchers are interested in reducing or preventing these anomaly events as much as possible by detecting or classifying them for further purposes.

2.2. Outlier detection and clustering problems

One of the most used approaches for anomaly detection is outlier detection: a process to detect and exclude any data that are dissimilar from the majority data. This approach is useful for anomalous detection since normal data usually have a great number of accessible samples, while abnormal one has very few [13, 14, 15]. Another approach to solving the problem of the imbalanced dataset and anomaly detection is clustering methods [16, 17, 18]. However, most outlier detection and clustering still suffer from the inability to classify between multiple types of anomaly incidents and unable to learn directly through characteristics of data. Especially, some irregularity classes do not express the peak behavior that is easy to detect but do express a small change of shape or behavior. Further, some anomaly data are very rare to find and might happen only once, thus making this event challenging for these methods to solve the problem. Hence, some types of anomaly incidents sometimes do not suit outlier detection and clustering approaches.

2.3. Imbalanced dataset problems

Since the outlier detection and clustering cannot provide efficient solutions to problems, the deep learning approach is an alternative way to solve the classification effectively [19]. However, naturally, the behavior of data in the real world comes with an imbalanced form, especially the anomaly data since they rarely happen and are difficult to collect. These problems lead to the classification issue, an imbalanced dataset, where the amount of data in some classes are very few when compared to others. The imbalanced dataset problem can affect the performance of the deep learning model because there is not enough training set for the model to learn and minimizes errors [20, 21, 22].

2.4. Time series data augmentation problems

Since some anomaly detection or classification tasks are unable to use the outlier detection and clustering approaches, another way to solve this problem is by increasing the number of samples to balance
between normal data and abnormal data. One of them is called data augmentation. It is not only beneficial for increasing the amount of data for better classification performance but also helps reduce and prevent the issue of overfitting of deep learning model [23, 24]. Overfitting can affect the learning process and model accuracy. It makes deep learning models tending to have high performance since there are a few quantities of data for training. This issue results in a model that does not have any robustness against the real test dataset that the model never learns before, thus making outcome accuracy very poor [25, 26].

Many recent augmentation techniques tried to eliminate the problem of the limited amount of data [27, 28, 29, 30], and one of the great examples is from Um’s paper [7]. This paper proposed various augmentation techniques for Parkinson’s disease classification using a convolutional neural network (CNN). He showed that increasing the amount of data through his augmentation techniques can significantly improve the accuracy performance. Nevertheless, this approach can be suffered from the absolute arbitrary when perturbing the data, thus resulting in an output that might contain very few to no characteristics of original data and affect the performance of the deep learning model for classification purposes. In addition, there is still a space of possibility for augmentation process that can increase varieties of data for further usage.

2.5. DAGAT solution and its limitations

Data augmentation and generation for anomalous time-series signals (DAGAT) [9] is our initiated framework for the data generation purpose. As mentioned in Subsection 2.4, typical data augmentation still has some limitations. DAGAT is a framework that using augmentation techniques on the various domains and then trained those augmented data with variational autoencoder (VAE). A reason behind using VAE is that the VAE model can be represented as a latent space that is beneficial for data exploration and generation. To increase the number of generated outputs, a signal fragment assembler (SFA) is used to increase the quantity and include multiple characteristics of augmentation inside each sample. Lastly, the process that makes data generation outstanding from other typical data augmentation is the quality classifier (QC). In the last step, this process compares generated sample with the original one to filter any inadequate quality sample out from generated sample. The overall framework of DAGAT is illustrated in Fig. 1.

This framework can perform well in generating simple data cases; however, there are problems with the functional design. It is unable to control the output number. Additionally, there is a possibility that bad quality data may be trained in VAE, producing bad latent space. Also, it misses an opportunity to include multiple characteristics for each sample in the latent space. Hence, in 2021, ADGAS was proposed to overcome these issues.

2.6. ADGAS solution and its limitations

As shown in Fig. 2, an advanced data generation for anomalous time-series signals (ADGAS) is an improved version of DAGAT. By modifying the placement of SFA and providing a quality classifier before proceeding to VAE, this modification can integrate multiple characteristics from various augmentation techniques into a sample [10]. To allow this framework to be able to control the number of outputs, DP is placed along with additional QC in a loop so that if the quality is not qualified, this function still works and finds another point for a new sample on latent space again, until it reaches the determined number.

According to the scenario of data generation problems and feasible solutions, we can summarize that overall problems can be stated as follows:

- Outlier detection and clustering sometimes are not suitable for classifying multi-class anomaly, the extremely limited number of anomaly events, and very small changes in the shape of signals.
- The limited amount of the available real anomaly can affect the performance of deep learning model due to the imbalance datasets.
- Existing data augmentation methods might result in random data that has no characteristics of original data and still has the limitation of new possible augmented data.
- DAGAT cannot control the number of generated outputs, unable to include multiple augmentation techniques in the latent space, and possibly unacceptable quality data in the VAE training.
- The signal fragment assembler, data picker technique, and quality classifier of ADGAS should be improved for better data variety and
quality on the latent space to achieve satisfactory outcome performance.

3. Proposed method

An overall framework for extremely rare case data generation is shown in Fig. 3, succeeding from an ADGAS framework with an improvement. The following describes each phase of the proposed framework as functional operations.

3.1. Data generation preparation phase

In this phase, applying time window slicing (TWS) as the first step is to crop time series data from a long record and turn it into a determined length by trimming only interesting parts of anomaly incidents. This function helps increase the initial substance of samples before augmenting on various domains to widen the possibility of augmentation instead of on time-domain only. Especially, it will be beneficial for raising more samples in the case that initiated data have very few. The result of applying the TWS process, called anomaly incident seeds, is transformed by three different transformations: upsampling-downsampling (see Subsection 3.1.1), fast Fourier transform (see Subsection 3.1.2), and time series decomposition (see Subsection 3.1.3). At this point, we have four different anomaly incident seeds, i.e., three transformed seeds and one direct seed of anomaly incidents, as shown in Fig. 3. These seeds are modified slightly by a time series data augmentation technique (see Subsection 3.1.4), and then signal fragment assembler (SFA) (see Subsection 3.1.5) provides a new set of time series anomaly incident samples. The following subsections describe the functional operations of each process.

3.1.1. Upsampling - downsampling

Upsampling - Downsampling [31] is one of the simple methods that can widen the sample space by varying the sampling rate. This method helps increase variations and possibilities of augmentation on different domains. Fig. 4 shows the augmentation process of upsampling and downsampling. Upsampling is an expansion of in-between original samples by augmenting the newly interpolated data, while downsampling is a contraction of samples; therefore, the upsampling and downsampling can help in transforming and perturbing samples differently.

3.1.2. Fast Fourier transform and inverse fast Fourier transform

Another domain transform approach is fast Fourier transform (FFT) and its inverse (IFFT). The one-dimensional FFT and IFFT are used in this work, similar to DAGAT [9] and ADGAS [10]. The FFT and IFFT can be expressed in Eqs. (1) and (2).

\[ Y(k) = \sum_{j=1}^{n} x(j)W_n^{j-1(k-1)} \]  
\[ x(j) = \frac{1}{n} \sum_{k=1}^{n} Y(k)W_n^{-(j-1)(k-1)} \]

where \( W_n = e^{i2\pi/n} \), \( n \) is the length of the signal, \( Y(k) \) is the frequency-domain signal and \( x(j) \) is the time-domain signal [32]. Fig. 7 shows an example of data augmentation on the frequency domain. After transforming normal data into the frequency domain using Eq. (1), various augmentation methods are applied and then inverting it back to the time domain using Eq. (2). This procedure can increase variations of samples.

3.1.3. Time series decomposition and re-composition

Normally, a time domain signal can be decomposed into three components: Trend \((T_i)\), Seasonal \((S_i)\), and Residual or noise \((R_i)\), as mentioned in Eqs. (3) and (4). This method can achieve a result in additive model \((y_{ia})\) or multiplicative model \((y_{im})\), depending on data behaviors and tasks [33].

\[ y_{ia} = T_i + S_i + R_i \]  
\[ y_{im} = T_i \times S_i \times R_i \]

Figs. 5 and 6 illustrate the time series decomposition, when the original data is decomposed using Eqs. (3) and (4), respectively. It randomly chooses different components of \( T_i, S_i, \) or \( R_i \), and then augments those components before recomposing back to the original domain.

Note that Figs. 5(b) and (d) and Figs. 6(b) and (d) show the missing values that occurred at the head and tail of signals. This phenomenon happens since the decomposition process using the symmetric moving average for statistic calculation.
3.1.4. Time series data augmentation

At this stage, we have four different anomaly incident seeds, as shown in Fig. 1. Here, data augmentation [7] is used by randomly choosing techniques among Jittering, Scaling, Magnitude Warping, Rotation, Permutation, and Time Warping. The number of parameters can be static or arbitrary, depending on the tasks. Note that after the augmentation process is complete, a set of anomaly incident seeds is transformed back to the time domain again.

3.1.5. Improved signal fragment assembler

Signal Fragment Assembler (SFA) can be viewed as an additional augmentation technique that leverages augmentation methods from Subsections 3.1.1 to 3.1.4 by including multiple ways to perturb time series in each sample. It arbitrarily divides each different sample into two parts and then randomly reassemble them into the new different orders; however, with an increased rate of division, we can provide more new possible different orders. Here, an improved signal fragment assembler (I-SFA) can divide up to four fragments as expressed in Eq. (5).

\[ n_s = i_s || j_s || k_s || l_s \] (5)

\[ n_s = i_s || j_s || k_s || l_s \] (5)

where \( i_s, j_s, k_s, l_s \) are the splitted data from different augmented samples, and \( n_s \) is the resulted sample that is completely assembled.
3.2. First quality classifier

Quality classifier (QC) is a process that selects good quality samples from the new different order sets of generated anomaly incident signals/samples. Here, the first QC (QC1) is implemented with a combination of mean absolute error (MAE) as denoted in Eq. (6) and histogram area scoring as expressed in Eqs. (7) and (8). The weighted technique is used to make the evaluation process determinable and adjustable as defined by Eq. (9). Furthermore, the QC1 can capture the first, middle, and last time window samples of original data and obtain the highest possible score from them as denoted in Eqs. (10) and (11). This approach can make QC1 more consistent against randomness while picking a sample for comparison.

\[
S_{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{N}
\]

(6)

where \( S_{MAE} \) is the MAE score, \( N \) is the total number of samples, \( y_i \) is the augmented time series sample, and \( x_i \) is the original time series sample.

A reason behind MAE instead of mean squared error (MSE) or root mean squared (RMSE) is that MAE is a completely absolute error in the comparison process, unlike MSE and RMSE that square the number of errors, resulting in insensitiveness against the outlier [34]. In this case, we seek an evaluation function that can be more strict when compared to the original data but not too similar. Hence, MAE is used instead for a better outcome.

\[
A = \sum_{i=1}^{N} \text{freq\_density} \times \text{class\_width}
\]

(7)

\[
S_{\text{histogram}} = (1 - |A_{\text{original}} - A_{\text{generated}}|) \times 100
\]

(8)

where \( S_{\text{histogram}} \) is the result score in percentage, \( A_{\text{original}} \) is the total histogram area calculated from the original data, and \( A_{\text{generated}} \) is the total histogram area calculated from the generated data.

\[
T = (S_{\text{MAE}} \times W_{\text{MAE}}) + (S_{\text{histogram}} \times W_{\text{histogram}}) \times 100
\]

(9)

where \( T \) is the total score calculated in percentage, \( S_{\text{MAE}} \) is the MAE score from Eq. (6), \( W_{\text{MAE}} \) is the weight of the MAE score, \( S_{\text{histogram}} \) is the histogram score from Eq. (7), and \( W_{\text{histogram}} \) is the weight of the histogram score.

\[
H_{\text{Highest}} = \max(T_{\text{first}}, T_{\text{middle}}, T_{\text{last}})
\]

(10)

where \( T_{\text{first}}, T_{\text{middle}} \), and \( T_{\text{last}} \) are the total scores of data, when compared with the first time-window, the middle time-window, and the last time-window of the original data, respectively.

\[
f(H_{\text{Highest}}) = \begin{cases} 1 & : H_{\text{Highest}} \leq H_{\text{Threshold}} \\ 0 & : \text{otherwise} \end{cases}
\]

(11)

where 1 implies the acceptable sample, selected to proceed in the data generation provider phase, 0 implies the unacceptable sample, rejected from the augmentation phase, \( H_{\text{Threshold}} \) is the highest acceptable threshold, and \( L_{\text{Threshold}} \) is the lowest acceptable threshold.

3.3. Data generation provider phase

Data generation provider is the cooperation of variational autoencoder (VAE) and data picker (DP). VAE learns generated samples, selected by QC1, and then represents them in a latent space. After that, DP randomly selects two points on the latent space and obtains the possible and available samples that are in-between for generating a new sample. The following subsections describe their functional operations.

3.3.1. Variational autoencoder

Variational autoencoder (VAE) is a network that learns signals from the previous part and represents them in latent space. It is a core of this phase to generate numerous samples with the controllable ability through the latent space, which can represent the learned data and find another available sample with multiple characteristics within. One of the ways to obtain the mentioned latent space is through the learning of VAE [35], as defined in Eq. (12). Fig. 8 demonstrates the latent space of observed compressed sample data projection on specific or determined dimensions. The importance of VAE is that the latent space gains the interpolation capability through undefined sample point (b) in between representation points (a and c) as highlighted with a dashed ellipse in Fig. 8. This action is useful for data generation since it can give us numerous amounts of samples from available areas with high variations. Here, the VAE learns all the samples, with good quality, selected by QC1.

\[
\log P(X) - D_{KL}(Q(z|X)||P(z|X)) = E[\log P(X|z)] - D_{KL}(Q(z|X)||P(z)]
\]

(12)

3.3.2. Data picker

Data picker (DP) of the ADGAS framework used the circular radius method for picking up some samples from latent space to select generated samples; however, this method may miss an opportunity to include variations of data that can help output more generalization for anomalous detection. Hence, we modify the previous DP by finding two random points in the latent space and obtaining the possible and available samples that are in-between. This modification helps DP explore the latent space further and widen the chance of including multiple characteristics of the data perturbation. The distance of two points is calculated by Eqs. (13) and (14), while finding obtainable point within the interval is defined in Eqs. (15) and (16).

\[
x_{\text{distance}} = \frac{x_2 - x_1}{\text{points} + 1}
\]

(13)

\[
y_{\text{distance}} = \frac{y_2 - y_1}{\text{points} + 1}
\]

(14)

\[
\text{position}_z = x_1 + \text{Current\_pace} \times x_{\text{distance}}
\]

(15)

\[
\text{position}_z = y_1 + \text{Current\_pace} \times y_{\text{distance}}
\]

(16)

3.4. Second quality classifier

In the last phase, the second quality classifier (QC2) takes a role as the double checker of this framework. Similar to the quality classifier (QC1) in Subsection 3.1, QC2 provides to ensure the quality of output samples from Subsection 3.3. Ones can modify the weight of each metric in QC2 as well to meet the requirement of data generation for their tasks. Fig. 9 illustrates an example of the finally generated samples using our proposed method. The model is synthesized from the WADI dataset [36], as explained in Subsection 4.1. The observation shows that this model can perturb original data and create slightly different outputs.
4. Experimental results

4.1. Dataset explanation

Here, we clarify two datasets used in our experiments, consisting of Secure Water Treatment System (SWaT) [37] and Water Distribution System (WADI) [36]. Both datasets were recorded from the researcher’s testbed, in terms of anomalous incidents that were simulated as cyber-attacks by researchers. The purpose of this information is to demonstrate the effect of cyber-attacks on the water treatment system.

4.1.1. Raw water tank’s sensor

The Raw water tank’s sensor (RAW) is part of the Secure Water Treatment system (SWaT) dataset. In the RAW dataset, sensors recorded various attacks that attempt to interrupt or modify the raw water tank component. Fig. 10 illustrates three scenarios of attacks: (i) increasing the water level in the tank at 1 mm/second to damage the components, as shown in the upper-right of Fig. 10, (ii) setting the water level at 700 mm constantly to cause the underflow of the water tank, as shown in the lower-left of Fig. 10, and (iii) decreasing the water level less than the minimum value to cause the overflow of the water tank, as shown in the lower-right of Fig. 10. For scenario (ii), attackers can successfully cause the underflow of the water as expected, while, for scenarios (i) and (iii), the sensor can record the attack in the system, but that attack did not damage any components or cause the overflow as intended.

4.1.2. Reverse osmosis water tank’s actuator

Like the previous dataset, the Reverse Osmosis water tank’s actuator (RO) is an actuator dataset, part of SWaT. RO recorded anomalous attacks on the reverse osmosis water tank system. In this dataset, researchers demonstrated attacks of tank overflow by (i) setting the level sensor that detects the level of liquid at 600 mm for a brief duration, (ii) adjusting the level of detected water lower than the lowest level. One case is a successful attack, and one case is not, and (iii) setting the water level at 1,000 mm constantly to cause the underflow of the tank.

4.1.3. Water Reservoir’s sensor

Water Reservoir’s sensor (WR) is a dataset from WADI [36]. Here, we observed the behavior of anomalous attacks on various sensors. The malicious cyber-attack on different components can affect a part or even the whole testbed. For this experiment, WR was selected for experiments to show the consequences of the system assaults. The attack scenarios are consist of: (i) stealthy strike that manipulates draining and filling tank level to drain an elevated reservoir, (ii) turn off various consumers valves to make WR level dropped drastically, and (iii) damage a motorized valve and water pump to drain an elevated reservoir.

Note that, in the ADGAS experiment, RAW and RO datasets were used for performance evaluation as well. However, both were tested in binary classification. Here, we were challenging our newly proposed method, ADGAS, and baseline augmentation techniques in multi-class classification and tested on the RAW, RO, and WR datasets to prove the performance.

4.2. Experimental setup

To evaluate the performance of each data generation model, a one-dimensional convolutional neural network (1D-CNN) [38] was implemented as a data generation performance tester. The 1D-CNN model comprises 7 convolution layers. The first to the seventh layers contain 16, 32, 64, 64, 64, 64, and 64 features, respectively. The $3 \times 1$ ker-
Fig. 9. A set of the generated anomalous samples, generated by our proposed method based on WADI anomaly dataset.

Fig. 10. An example of the original Raw Water Tank's anomaly signal and normal signal.
nel, 2 strides, rectified linear unit (ReLU) activation function, similar to the model’s setting as reported in [7]. Additionally, dropout at a rate of 0.1, and batch normalization [39] were administered on each layer. Flattened features are connected to a fully connected layer (FC) of 5,312 nodes, plus the ReLU as an activation function. The last layer, providing the result of the anomaly classification class, is the Softmax activation function for multi-categories of anomalous events. Furthermore, this experiment applied early stopping to prevent an incident, where the loss at the last epoch may tremendously increase and affect the classification performance [40, 41].

For typical augmentation baseline setting, datasets were generated by means of rotation (R), rotation and permutation (R + P), and rotation with permutation and time warping (R + P + TW) techniques with arbitrary values of arguments, similar to the experimental setting as reported in [7]. Note that the lack of ability to determine the output number as mentioned in Section 2.5, as a consequence, makes it impossible for the DAGAT model to test in the controlled environment experiment.

For ADGAS, we duplicated the setting from previous experiments. Jittering, scaling, magnitude warping, time warping, and permutation were used as data pre-processing with random values of frequency, amplitude, and phase. The data pickers chooses generated samples from the latent space within 5 units circle radius. The score for every QC is in-between 80%-95%.

For the proposed model setting, the augmentation process was similar to ADGAS. For SFA, we randomly chose two types of division (two-part division and four-part division) to create the high varieties of output samples. To maintain the diversity of generated samples, \( W_{\text{histogram}} \) was set to 0.7, while \( W_{\text{MAE}} \) score weight was set to 0.3 in the QCI process. The improved DP chose two random points on the latent space within 100 units. The QC2’s score weight setting for \( W_{\text{MAE}} \) was 0.3 and \( W_{\text{histogram}} \) was 0.7. This can increase the similarity of the generated samples when compared to the original samples.

To ensure the performance evaluation, 10-fold cross-validation was applied in every experiment. Each dataset contains 4,000 samples consisting of 800 samples of anomaly events in 3 classes and 1,600 samples of normal events with 1 class. Every data generation method exported data with 3 hours time periods (10,800 records) and 1 second for each time window slicing. The anomaly data for the training phase are all from the generated models, while normal data are from real normal samples. For the testing phase, both anomalous and normal are the real samples.

4.3. Results and discussion

Here, all test models, including R, R + P, R + P + TW, ADGAS, and our proposed model, were evaluated in the form of multi-class classification. Such test models were implemented with 1D-CNN as described in Subsection 4.2. R, R + P, R + P + TW models were trained with training sets that were generated by rotation, rotation and permutation, and rotation with permutation and time warping augmentation techniques, respectively. ADGAS model was trained with a training set, generated from the ADGAS framework in Fig. 2. In the same way, our proposed model was trained with a training set, generated from the framework in Fig. 3. The performance of all test models was evaluated in terms of accuracy, precision, recall, and F1-score, as defined by Eqs. (17), (18), (19), and (20), respectively.

\[
\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100 \tag{17}
\]

\[
\text{Precision} = \frac{TP}{TP + FP} \tag{18}
\]

\[
\text{Recall} = \frac{TP}{TP + FN} \tag{19}
\]

\[
F1 - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{20}
\]

where \( T \) is the correct classification of anomaly incidents, \( TN \) is the correct classification of normal classes, \( FP \) is the incorrect classification of anomaly incidents, and \( FN \) is the incorrect classification of normal classes.

Consequently, as can be seen in Tables 1 and 2, our proposed model outperforms all baseline methods in every measurement aspect for both datasets, 91.85% and 94.26% of accuracy on RAW and RO, respectively. The proposed model also achieves much better results in precision, recall, and F1-score in all cases, especially on the RO dataset.

As expected, when comparing our proposed model with ADGAS, as reported in Tables 1 and 2, the main achievement comes from (i) applying a concept of blending multiple characteristics to an I-SFA process (see Fig. 3), so that a new synthetic sample can be generated from various augmentation techniques, (ii) screening any too dissimilar samples out from the initial dataset by QCI (see Fig. 3), so that the latent space contains good quality data, and (iii) an improved DP was implemented with interpolation approach that can be beneficial in selecting numerous characteristics of data that are in between two random points on the latent space (see Figs. 3 and 8).

Not only does generalization of data improve a learning model of 1D CNN, but also the robustness to the randomness of data is part of this success. Fig. 3 shows the use of QC2 as a double checker by setting the weight of MAE’s QC2 higher than that of MAE’s QCI, yielding outputs with better characteristics. This modification can improve the model robustness, reflected as a lower SD-score of the multi-class classification results.

As listed in Table 3, the overall performance of the proposed model achieves higher scores when compared to other data augmentation techniques, R, R + P, R + P + TW, and when compared to ADGAS in terms of mean values, the accuracy of the proposed model is better than that of ADGAS by 15.50%. Notice that the accuracy, precision, recall, and F1-score in terms of SD score of the proposed model slightly higher than the ADGAS model. This result implies that the robustness of the proposed model was slightly lower than that of the ADGAS model. Although the proposed model used the QCI and QC2 with additional signal measurement procedures for quality control, sometimes, the randomness for generalization in synthesizing procedure might affect the output quality, similar to the typical augmentation approach. To avoid the robustness drop down in the future, the set of variables in the framework should adjust flexibly, since the same setting might not suitable for all datasets and scenarios.

5. Conclusion

In this paper, we have proposed a data generation framework for extremely rare case signals. The proposed model, an implemented framework, succeeds from the ADGAS framework with modification. The main achievements are from: (i) eliminating a problem of synthesizing data that might contain very few characteristics of the original one and (ii) levitating data variety and quality on a latent space for the better-generated data. The three main contributions to support the above achievements are listed below:

- SFA is upgraded by modifying the division of the sample from two parts to four parts. This modification helps generate a single sample by integrating multiple various characteristic samples by means of various augmentation techniques.
- DP is improved by applying a linear interpolation between two random points on a latent space. This improvement helps achieve a great variety of generated samples.
- QC is implemented with a signal comparison technique in conjunction with controllable weighted metrics. This technique helps provide high various synthesized samples that contain attributes of original data.
As a result of our contributions, we achieve the generalization of data samples, thus leading to an improvement of model robustness and performance, as obviously reported in Tables 1, 2, and 3.

However, some important parameters of DP and QC were manually set as described in the experimental setup. It can be viewed as a manual mode, which was not adaptable for further datasets. Therefore, in future work, all parameters should be self-adjustment; that is, we will extend our proposed method to work with an automatic/adaptive mode.

Declarations

Author contribution statement

T. Chalongvorachai: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper. K. Woraratpanya: Conceived and designed the experiments; Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data; Wrote the paper.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

Data included in article supplementary material/referenced in article.

Declaration of interests statement

The authors declare no conflict of interest.

Additional information

No additional information is available for this paper.

References

[1] J. Zhang, M. Zulkernine, Anomaly based network intrusion detection with unsupervised outlier detection, in: 2006 IEEE International Conference on Communications, vol. 5, 2006, pp. 2388–2393.
[2] N.D. Singh, A. Dhall, Clustering and learning from imbalanced data, CoRR, arXiv: 1811.00972.
[3] L. Cao, H. Shen, Imbalanced data classification using improved clustering algorithm and under-sampling method, in: 2019 20th International Conference on Parallel and Distributed Computing, Applications and Technologies (PDCAT), 2019, pp. 358–363.
[4] M. Kamaladevi, K. Sekar, V. Venkataraman, K. Kannan, Outlier detection in imbalanced data classification, Int. J. Recent Trends Eng. Technol. 7 (2019) 972–975.
[5] H. Zhang, L. Huang, C. Wu, Z. Li, An effective convolutional neural network based on smote and Gaussian mixture model for intrusion detection in imbalanced dataset, Comput. Netw. 177 (2020) 107315.
[6] I. Ioinon, R. Tkachenko, M. Gregus, K. Zuh, N. Lotoshynka, Input doubbling method based on SVR with RBF kernel in clinical practice: focus on small data, Proc. Comput. Sci. 184 (2021) 606–613.
[7] T.T. Um, F. Pilster, D. Pichler, S. Endo, M. Lang, S. Hirsch, U. Fietzek, D. Kulic, Data augmentation of wearable sensor data for Parkinson’s disease monitoring using convolutional neural networks, in: Proceedings of the 19th ACM International Conference on Multimedia Interaction, 2017, pp. 216–220.
[8] S. Kotsiantis, D. Kanellopoulos, P. Intelas, Handling imbalanced datasets: a review, GSTS Int. Trans. Comput. Sci. Eng. 30 (2005) 25–36.
[9] T. Chalongvorachai, K. Woraratpanya, DAGAT: data augmentation and generation for anomalous time series signals, in: 9th Joint Symposium on Computational Intelligence (JSCI9), Bangkok, Thailand, 2020, pp. 1–4.
[10] T. Chalongvorachai, K. Woraratpanya, ADGAS: an advanced data generation for anomalous signals, in: The 17th International Conference on Computing and Information Technology, 2020, pp. 1–10.
[11] D. Cieslak, N. Chawla, A. Striegel, Combating imbalance in network intrusion datasets, in: 2006 IEEE International Conference on Granular Computing, 2006, pp. 732–737.
[12] B. Brzezinski, D.J. Deppu, D. Njo, Real-Time, Personalized Anomaly Detection in Streaming Data for Wearable Healthcare Devices, Springer International Publishing, Cham, 2017, pp. 403–426.
[13] K. Singh, S. Upadhya, Outlier detection: applications and techniques, Int. J. Comput. Sci. 9 (3) (2012) 307–323.
[14] P. Gogoi, D.K. Bhattacharyya, B. Borah, J. Kalita, A survey of outlier detection methods in network anomaly identification, Comput. J. 54 (2011) 570–588.
[15] R. Domingues, M. Filippone, J. Zouauoi, A comparative evaluation of outlier detection algorithms: experiments and analyses, Pattern Recognit. 74 (2018) 406–421.
[16] I. Syafri, A. Prugel-Bennett, G. Wills, Unsupervised clustering approach for network anomaly detection, in: R. Benlamri (Ed.), Networked Digital Technologies, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 135–145.
[17] L. Portnoy, E. Eskin, S. Stolfo, Intrusion detection with unlabeled data using clustering, in: Proceedings of ACM CCS Workshop on Data Mining Applied to Security (DMSA), 2001, pp. 5–8.
[18] Y. Yasami, S. Khorasandi, S.P. Mozaffari, A. Jalalian, An unsupervised network anomaly detection approach by k-Means clustering & ID3 algorithms, in: 2008 IEEE Symposium on Computers and Communications, 2008, pp. 398–403.
[19] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, P.-A. Muller, Deep learning for time series classification: a review, Data Min. Knowl. Discov. 33 (2019) 917–963.
[20] R. Potorel, C. Lennara, A comprehensive study of the effect of class imbalance on the performance of classifiers, in: Proceedings of the 13th International Conference
on Enterprise Information Systems - Volume 2: ICEIS, INSTICC, SciTePress, 2011, pp. 14–21.
[21] S. Visa, A. Ralescu, Issues in mining imbalanced data sets – a review paper, in: Proc. 16th Midwest Artificial Intelligence and Cognitive Science Conference, Dayton, USA, 2005, pp. 67–73.
[22] J. Johnson, T. Khoshgoftaar, Survey on deep learning with class imbalance, J. Big Data 6 (2019) 27.
[23] X. Ying, An overview of overfitting and its solutions, J. Phys. Conf. Ser. 1168 (2019) 022022.
[24] S. Salman, X. Liu, Overfitting mechanism and avoidance in deep neural networks, arXiv:1901.06566 [abs].
[25] W. Zhao, Research on the deep learning of the small sample data based on transfer learning, AIP Conf. Proc. 1864 (1) (2017) 020018.
[26] R. Rodelo, V. Shankar, B. Recht, S. Fridovich-Keil, M. Hardt, J. Miller, L. Schmidt, A meta-analysis of overfitting in machine learning, in: NeurIPS, 2019.
[27] A. Mikolajczyk, M. Grochowski, Data augmentation for improving deep learning in image classification problem, in: 2018 International Interdisciplinary PhD Workshop (IIPSW), 2018, pp. 117–122.
[28] K.M. Phyu, K. Woraratpanya, Car damage detection and classification, in: Proceedings of the 11th International Conference on Advances in Information Technology, IAIT2020, 2020.
[29] F. Wang, S. Zhong, J. Peng, J. Jiang, Y. Liu, Data augmentation for EEG-based emotion recognition with deep convolutional neural networks, in: MMM, 2018, pp. 82–93.
[30] Z. Zhang, F. Duan, J. Solé-Casals, J. Dinarés-Ferran, A. Cichocki, Z. Yang, Z. Sun, A novel deep learning approach with data augmentation to classify motor imagery signals, IEEE Access (2019) 15945–15954.
[31] K. Karioja, J. Laurila, Signal upsampling in machine diagnostics - an experimental approach, in: The Thirteenth International Conference on Condition Monitoring and Machinery Failure Prevention Technologies, 2016.
[32] E.O. Brigham, R.E. Morrow, The fast Fourier transform, IEEE Spectr. 4 (12) (1967) 63–70.
[33] E.B. Dagum, Time series modelling and decomposition, Statistica 70 (2010).
[34] T. Chai, R. Draxler, Root mean square error (RMSE) or mean absolute error (MAE)?– arguments against avoiding RMSE in the literature, Geosci. Model Dev. 7 (2014) 1247–1250.
[35] D.P. Kingma, M. Welling, Auto-encoding variational bayes, CoRR, arXiv:1312.6114 [abs].
[36] Singapore University of Technology and Design (SUTD), Dataset characteristics, https://itrust.sutd.edu.sg/itrust-labs_datasets/dataset_info/, 2017. (Accessed 17 June 2021).
[37] J. Goh, S. Adepu, K.N. Junejo, A. Mathur, A dataset to support research in the design of secure water treatment systems, in: Critical Information Infrastructures Security, Springer International Publishing, Cham, 2017, pp. 88–99.
[38] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, D.J. Inman, 1d convolutional neural networks and applications: a survey, Mech. Syst. Signal Process. 151 (2021) 107396.
[39] S. Ioffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in: Proceedings of the 32nd International Conference on International Conference on Machine Learning, Vol. 37, ICML’15, JMLR.org, 2015, pp. 448–456.
[40] R. Caruana, S. Lawrence, L. Giles, Overfitting in neural nets: backpropagation, conjugate gradient, and early stopping, in: Proceedings of the 13th International Conference on Neural Information Processing Systems, NIPS’00, MIT Press, Cambridge, MA, USA, 2000, pp. 381–387.
[41] L. Prechelt, Early Stopping · But When?, Springer Berlin Heidelberg, 2000, pp. 53-67.