Applied Intelligence (2023) 53:272–288
https://doi.org/10.1007/s10489-022-03361-2

A novel multi-module integrated intrusion detection system for high-dimensional imbalanced data

Jiyuan Cui[1] · **Liansong Zong**[1] · **Jianhua Xie**[1] · **Mingwei Tang**[1]

Accepted: 7 February 2022 / Published online: 14 April 2022

Abstract
The high dimension, complexity, and imbalance of network data are hot issues in the field of intrusion detection. Nowadays, intrusion detection systems face some challenges in improving the accuracy of minority classes detection, detecting unknown attacks, and reducing false alarm rates. To address the above problems, we propose a novel multi-module integrated intrusion detection system, namely GMM-WGAN-IDS. The system consists of three parts, such as feature extraction, imbalance processing, and classification. Firstly, the stacked autoencoder-based feature extraction module (SAE module) is proposed to obtain a deeper representation of the data. Secondly, on the basis of combining the clustering algorithm based on gaussian mixture model and the wasserstein generative adversarial network based on gaussian mixture model, the imbalance processing module (GMM-WGAN) is proposed. Thirdly, the classification module (CNN-LSTM) is designed based on convolutional neural network (CNN) and long short-term memory (LSTM). We evaluate the performance of GMM-WGAN-IDS on the NSL-KDD and UNSW-NB15 datasets, comparing it with other intrusion detection methods. Finally, the experimental results show that our proposed GMM-WGAN-IDS outperforms the state-of-the-art methods and achieves better performance.

Keywords Intrusion detection system · Stacked autoencoder · Gaussian mixture model · Wasserstein generative adversarial network · Imbalance processing · Feature extraction · Deep learning

1 Introduction

The explosive growth of network traffic has made the network security situation increasingly severe. Especially in the epidemic of COVID-19 Coronavirus, most enterprises are working remotely, which further highlights the importance of network security. According to the Global Risk Report 2021 [1], cyber attacks are listed as one of the top five sources most likely to pose a serious risk on a global scale. Intrusion detection system (IDS) [2] is an active defense technology that can detect malicious attacks, unauthorized activities, and potential threats in the network. Traditional intrusion detection methods mainly match detection by creating rule bases, which cannot detect unknown attacks. Along with the great progress of machine learning in the fields such as computer vision [3] and human computer interaction [4], machine learning is also applied to intrusion detection. Intrusion detection methods based on machine learning largely depend on manual feature selection. Deep learning technology can take network traffic anomaly detection as a classification problem [5]. At present, the application of deep learning technology to intrusion detection has achieved good results, but there are still some shortcomings.

First, due to the influence of data dimension, the detection rate of the trained model for some attack classes is low, which affects the overall detection rate.

Secondly, data imbalance affects the performance of intrusion detection classifier. For example, in UNSW-NB15 dataset, there are 2,218,761 normal data and 321,283 attack data, of which the rarest worm attack is only 174.

In order to overcome the above shortcomings, this paper proposes a novel multi-module integrated intrusion detection system for high-dimensional imbalanced data, which is called GMM-WGAN-IDS. The main contributions of the proposed system in this paper are as follows.

– We design a feature extraction module based on stacked autoencoder, SAE module. The module is able to obtain

✉ Mingwei Tang
tang4415@126.com

[1] School of Computer and Software Engineering, Xihua University, Chengdu, Sichuan 610039, China

deep feature representations of high-dimensional data, so as to achieve optimal feature extraction.
- By combining the gaussian mixture model-based clustering algorithm (GMM-based clustering algorithm) and the gaussian mixture model-based wasserstein generative adversarial network (GMM-based WGAN), we propose the imbalance processing module, GMM-WGAN. The module can effectively alleviate the problems of data imbalance and insufficient rare attack samples.
- We merge convolutional neural network (CNN) and long short-term memory (LSTM) together to form a classification module, CNN-LSTM. The module fully considers the correlation between data and is more suitable for intrusion detection environment.
- In order to maximize the intrusion detection capability, this paper integrates the above three sub-modules and proposes GMM-WGAN-IDS. To the best of our knowledge, this is the first attempt to integrate three modules together for intrusion detection. To evaluate its performance, experiments are conducted on two datasets and compared with other intrusion detection methods. In addition, ablation study is performed to assess the effectiveness of each module.

The rest of this paper is organized as follows. We briefly introduce the related work in Section 2, including the application of machine learning methods, deep learning methods, feature extraction, and imbalance processing methods in intrusion detection. In Section 3, we outline the background knowledge, including autoencoder (AE) and generative adversarial network (GAN). Section 4 describes the design and implementation of GMM-WGAN-IDS in details. Section 5 presents the experimental procedure and comparative study. Finally, Section 6 concludes the paper.

2 Related work

In recent years, machine learning is widely used in network intrusion detection. Nawir et al. [6] designed an online classifier (AODE) to process dynamic network data. The algorithm enables the classifier to constantly update features in the training phase, and realizes the rapid detection of data. Using the packing method based on genetic algorithm as the search strategy, Khammassi et al. [7] selected the best feature subset of the whole dataset. Three different decision tree classifiers are used to measure the performance of the selected feature subset. The results show that the method improves the classification performance of the system, but the classification accuracy for rare attacks is low. Most machine learning-based intrusion detection methods rely on manual feature selection, which can be a very cumbersome and time-consuming process. In addition, these methods are constrained by their inadequate expression ability on data and have difficulties in dealing with imbalanced data effectively.

For large datasets, deep learning [8] has significant advantages over shallow neural networks and general machine learning algorithms. Kamalakanta et al. [9] proposed a context-adaptive intrusion detection system. The system distributes multiple independent deep reinforcement learning agents in the network to detect complex network attacks. Caminero et al. [10] applied adversarial reinforcement learning to the field of intrusion detection. Experimental results show that the model has good accuracy but does not adequately deal with the data imbalance problem. Tian et al. [11] proposed an intrusion detection approach based on improved deep belief network (DBN). A combined sparsity penalty term based on Kullback- Leibler (KL) divergence and non-mean Gaussian distribution is introduced in the likelihood function of the unsupervised training phase of DBN, thus avoiding the problem of feature homogeneity and overfitting. Li et al. [12] proposed a deep learning approach for intrusion detection using a multi-convolutional neural network (multi-CNN) fusion method. According to the correlation, the feature data is divided into four parts, and then the one-dimensional feature data is converted into a grayscale graph. The experimental results demonstrate that the multi-CNN fusion model provides a classification method with high accuracy and low complexity. Qureshi et al. [13] proposed deep neural network and adaptive self-taught-based transfer learning technique, which exploits the concept of self-taught learning to train deep neural networks for reliable intrusion detection. It is experimentally shown that the proposed approach is robust and offers good generalization.

2.1 Feature extraction

For high-dimensional network data, efficient feature extraction is a hot topic in the field of intrusion detection. Xu et al. [14] proposed a new intrusion detection method LCVAE. The method uses a logarithmic hyperbolic cosine function (log-cosh) to design an efficient loss term balance generation and reconstruction process with CNN for feature extraction and classification. Ieracitano et al. [15] proposed a novel statistical analysis and autoencoder (AE) driven intelligent intrusion detection system (IDS). The IDS combines data analytics and statistical techniques with recent advances in machine learning theory to extract more optimized, strongly correlated features. Al-Turaiki et al. [16] proposed a novel network intrusion detection model based on deep learning. Also, a hybrid two-step preprocessing approach is applied to generate meaningful features. The approach combines dimensionality

reduction and feature engineering using deep feature synthesis. Kasongo et al. [17] studied a novel intrusion detection system using a wrapper based feature extraction unit (WFEU) and feed-forward deep neural network (FFDNN), which performs better in binary classification but poorly in multi-classification. Shams et al. [18] proposed a new context-aware feature extraction method as a preprocessing step for convolutional neural network (CNN)-based multi-class intrusion detection, which reduces feature space and classification time and improves classification accuracy. Liu et al. [19] designed a particle swarm optimization-based gradient descent algorithm (PSO-LightGBM). PSO-LightGBM can extract the features of the data and input them into one-class svm (OCSVM) to identify malicious data. The results show that the method has better accuracy for the detection of small sample data.

When facing a large number of data, the traditional data dimensionality reduction method can only extract incomplete feature representation from the data. It is difficult to extract the deep features of the data by using single-layer AE. Therefore, a stacked autoencoder (SAE) is proposed to achieve dimensionality reduction and reconstruction of high-dimensional data.

2.2 Imbalance processing methods

The imbalanced distribution of training samples can cause classification algorithms to tend to misclassify classes with fewer samples into classes with more samples [20]. Most machine learning and deep learning algorithms are sensitive to imbalanced data, which can lead to low detection rates of rare attacks. To improve the recognition rate of classification algorithms for minority categories, researchers have dealt with data imbalance mainly at the data level. Verma et al. [21] proposed an intrusion detection model based on time series. The detection rate of minority classes is improved by adaptive synthetic sampling (ADASYN). Jiang et al. [22] used the one-side selection (OSS) method to reduce the majority class samples and the synthetic minority over-sampling technique (SMOTE) to increase the minority class samples, so as to alleviate the data imbalance. Bedi et al. [23] proposed an improved siam-ids (I-SiamIDS) to cope with the data imbalance problem. In the first layer, the input samples are hierarchically filtered to identify attack traffic, and then sent to the second layer to identify specific types of attacks. Ma et al. [24] combined reinforcement learning algorithms with SMOTE techniques to design the AESMOTE model, which outperforms the original AE-RL model in several aspects.

Data augmentation algorithms are able to address the imbalance of the training set by artificially creating rare attack samples. However, existing data augmentation methods either ignore the distributional features of the data or ignore the spatial knowledge between features [25]. Therefore, this paper combines gaussian mixture model (GMM) with wasserstein generative adversarial network (WGAN) to adequately deal with the data imbalance problem. Currently, WGAN is rarely used as a data augmentation algorithm for intrusion detection.

3 Background

3.1 Autoencoder

AE [26] consists of an encoder and a decoder, which is an unsupervised neural network. The encoder compresses the original data into low-dimensional data, and the decoder restores the low-dimensional data into the original data. The features extracted from the original data are used as the output of the encoder. Its core function is to mine deep representation of input data, which is mainly used in feature extraction and nonlinear dimensionality reduction [27]. The AE minimizes the error between the input data and the reconstructed output data by learning a mapping function. Finally, the network parameters are adjusted by backpropagating the error [28].

The AE structure is shown in Fig. 1, given the input dataset $(x_1, x_2, x_3, \cdots, x_m)$. h_m denotes the feature representation of the input data after encoding by the hidden layer, and $\hat{x}_i$ denotes the output representation of the reconstructed original data after decoding from the hidden layer. The functional expression of the encoding process is shown in (1).

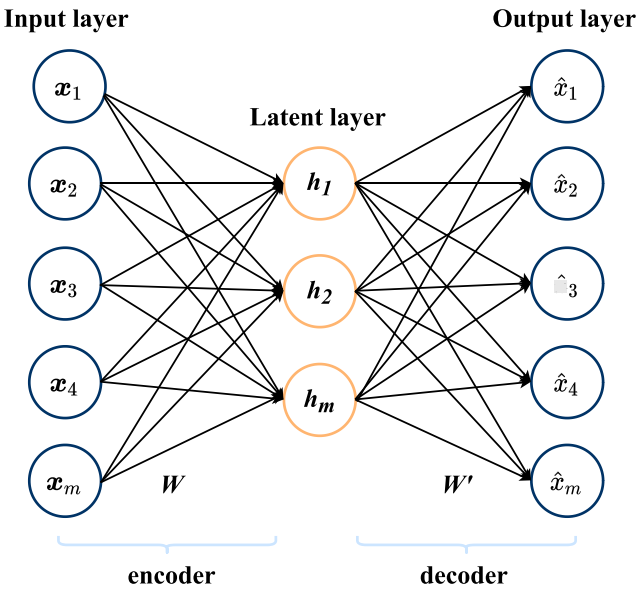


Fig. 1 The structure of autoencoder

$$h\left(x_i\right) = f\left(W^{(1)}x_i + b^{(1)}\right) \tag{1}$$

Where, $W^{(1)}$ is the weight matrix between the input layer and the hidden layer, and $b^{(1)}$ is the bias vector of the input layer. The activation function f is a sigmoid function with the value range of [0,1]. The decoder process is shown in (2).

$$\hat{x}_i = g\left(W^{(2)}h\left(x_i\right) + b^{(2)}\right) \tag{2}$$

Where, $W^{(2)}$ is the weight matrix between the hidden layer and the output layer, and $b^{(2)}$ is the bias vector of the hidden layer, and g is the decoding activation function sigmoid. The parameter matrix of the AE is optimized to minimize the reconstruction error, as shown in (3).

$$\phi\left(\Theta\right) = arg_{\theta,\theta'}min\frac{1}{m}\sum_{i=1}^{m}L\left(x_i, \hat{x}_i\right) \tag{3}$$

Where L represents the error loss function, as shown in (4).

$$L\left(x, \hat{x}\right) = \|x - \hat{x}\|^2 \tag{4}$$

3.2 Generative adversarial network

GAN is a probabilistic generative model proposed by Goodfellow et al. [29], which is inspired by the "two-person zero-sum game" in game theory. GAN can learn the distribution characteristics of sample data and generate data with similar characteristics.

As shown in Fig. 2, GAN consists of a generator (G, Generator Model) and a discriminator (D, Discriminative Model). The task of G is to generate data $G(z)$ that matches the real data distribution Pr by the input randomly distributed noise z. The task of D is to discriminate whether the input sample is the real data x from the dataset or the generated data $G(z)$. The ultimate goal of G is to maximize the probability that D discriminates incorrectly, and the ultimate goal of D is to maximize the probability that it discriminates correctly, i.e., $D(G(z))$ is as close as possible to 0 and $D(x)$ is as close as possible to 1. The performance of generator G and discriminator D is mutually improved by adversarial training. The model optimization function is given by (5).

$$\min_G \max_D V\left(D, G\right) = E_{x-P_r(x)}\left[\ln_e D\left(x\right)\right] + E_{z-P_g(z)}\left[\ln\left(1 - D\left(G\left(z\right)\right)\right)\right] \tag{5}$$

Where $Pr(x)$ is the distribution of real samples and $Pg(z)$ denotes the distribution of noise. The training of the GAN is performed by G and D alternately. In the ideal state, the model finds a globally optimal solution, that is, D cannot determine whether the input data is the real data x or the data $G(z)$ generated by G. At this time, the data generated by the generator G is highly similar to the real data.

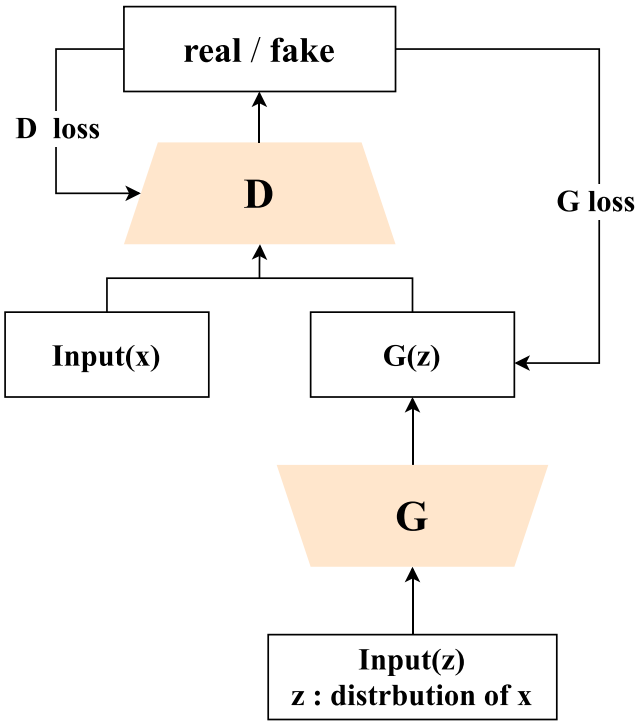


Fig. 2 Structure of generative adversarial network

4 Proposed methodology

4.1 Frame

The architecture of GMM-WGAN-IDS proposed in this paper is shown in Fig. 3, which consists of three main modules: feature extraction module, imbalance processing module, and classification module.

4.2 Feature extraction module

AE consists of a simple three-layer network. When the dimension of data is high, the feature extraction function can not meet the requirements, especially for minority class data. Therefore, this paper constructs an SAE-based feature extraction module (SAE module) to perform feature dimensionality reduction on the data and mine the deep representation of the data.

After the AE training is completed, the output features of its hidden layer are used as the input of the next AE. And so on, multiple AEs are stacked layer by layer to form SAE [30]. By learning multiple hidden layers, the low-level features are combined to obtain the high-level sample representative features without learning redundant features. The structure of the SAE is shown in Fig. 4.

As the first module of GMM-WGAN-IDS, whether the best features can be extracted or not will affect the performance of the whole system, so it is crucial to

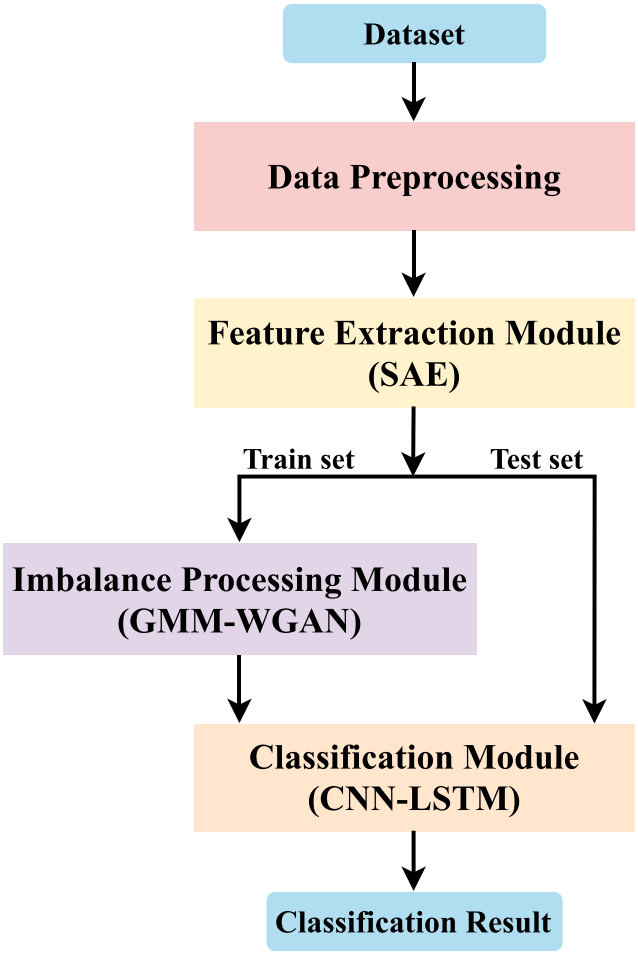


Fig. 3 Structure of GMM-WGAN-IDS

train the module to achieve optimal feature extraction. The training of SAE consists of unsupervised pre-training and supervised parameter fine-tuning. In the unsupervised pre-training phase, the output of each layer is used as the input of the next layer, and training is performed independently layer by layer. In the supervised fine-tuning phase, all layers are considered as a whole and fine-tuning of parameters such as weights and biases of the whole network is done by training again using labeled data and gradient descent. This two-stage training mode optimizes the network structure, reduces the output error, improves the learning efficiency, which enables SAE to have powerful dimensionality reduction and feature expression capability.

4.3 Imbalance processing module

The imbalance processing module is divided into two parts. For the majority class, GMM-based clustering algorithm is used to under-sample the redundant samples. For the minority class, GMM-based WGAN is used to generate rare samples, thereby expanding the labeled sample set. When performing multi-classification tasks, this module is used to assist training, alleviate data imbalance and improve the detection capability of the model.

4.3.1 Majority class under-sampling of GMM-based clustering algorithm

GMM is a parametric probability distribution model that represents a linear combination of multiple gaussian distribution functions. Probability distributions based on GMM use sampling algorithms to produce samples that are consistent with the distribution of real data. A common method for GMM parameter estimation is the expectation-maximization (EM) algorithm [31]. Assuming that all samples are from multiple gaussian distributions with different parameters, samples belonging to the same distribution are grouped into the same cluster and the GMM returns the probability that the samples belong to different clusters. As in (6), the GMM can be viewed as a mixture of K gaussian distributions at certain proportions, with each gaussian component determined by the mean μ and the covariance matrix Σ.

$$p(x) = \sum_{k=1}^{K} p(k)\, p(x|k) = \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k), \sum_{k=1}^{K} \pi_k = 1 \quad (6)$$

GMM is an extension of the gaussian model, which uses a combination of multiple gaussian distributions to portray the data distribution. As shown in Fig. 5, the GMM-based clustering algorithm can obtain the clustering distribution state of each class of samples. The majority class samples are under-sampled according to the clustering distribution state, and only the most representative core samples of the majority class are retained. The algorithm not only reduces redundant data, but also achieves accurate under-sampling.

4.3.2 Minority class generation for GMM-based WGAN

Although the classical GAN uses zero-sum game theory to define a new generative model, it measures the distance between the real sample and the generated sample through the jensen-shannon divergence (JS). This leads to gradient disappearance [32] when optimizing the objective function (5). In addition, some problems also occur in the training process, such as instability, model collapse and so on. WGAN [33] is a modification of GAN that uses wasserstein distance as a distance metric and translates it into an optimization problem. By using wasserstein distance to compare data distributions, the wasserstein value can well represent the distance between two data samples. Even if there is no overlap between the two data distributions, the training of the model is more stable. This method basically solves the problems of model collapse and gradient disappearance. The optimized objective function is (7).

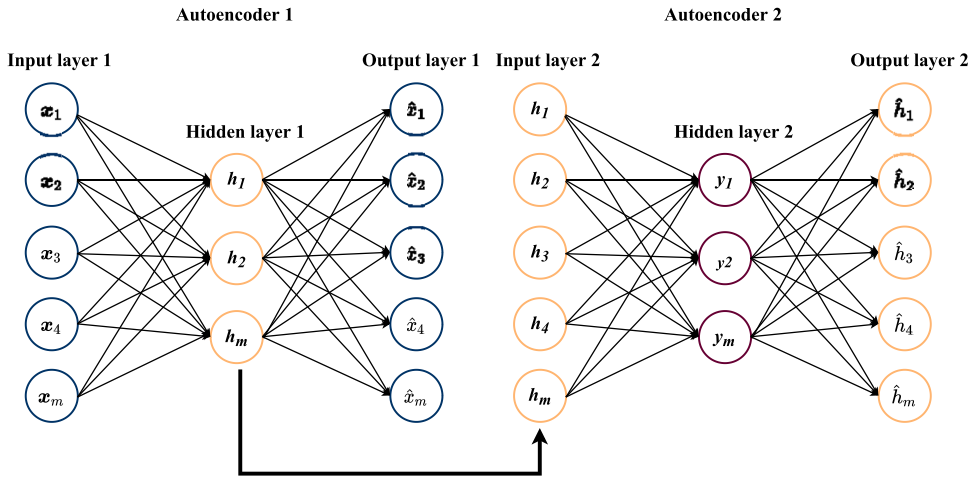


Fig. 4 The structure of stacked autoencoder

$$L = E_{x\sim P_{data}(x)}\left[D\left(x\right)\right] - E_{z\sim P_z(z)}\left[D\left(G\left(z\right)\right)\right] \tag{7}$$

According to (7), the loss functions of the discriminator and generator can be derived as (8) and (9).

$$D_{loss} = E_{z\sim P_z(z)}\left[D\left(G\left(z\right)\right)\right] - E_{x\sim P_{data}(x)}\left[D\left(x\right)\right] \tag{8}$$

$$G_{loss} = -E_{z\sim P_z(z)}\left[D\left(G\left(z\right)\right)\right] \tag{9}$$

Where x is the input real sample, $P_{data}\left(x\right)$ is the real sample distribution, and z is the input noise, $P_z\left(z\right)$ is the distribution of the noise.

Although WGAN makes the training process more stable, there are still some problems. When the number of training samples is small, bad samples will be generated, which makes it difficult for WGAN to converge. When G and D reach the equilibrium point, the data generated by G still has high repeatability. The generator describes the distribution of training samples with a single distribution, which is difficult to reflect the feature diversity of samples. Due to the single characteristics of the generated samples, the purpose of data enhancement cannot be achieved.

The essence of GMM is to use multiple normal distributions to portray the diverse characteristics of the samples as a whole. By constructing multiple gaussian distributions of minority class samples, a mixed distribution model consisting of multiple normal distributions is built. On the one hand, the mixed model of multiple distributions can better portray the diversity characteristics of the samples. On the other hand, the diversity of data features is constrained by each distribution, which makes the new samples generated by the hybrid model not only have diversity, but also maintain the similarity with the characteristics of the original samples. Through the integration of GMM into WGAN, GMM-based WGAN can better deal with each gaussian distribution of minority classes to generate more real minority class samples. The structure of the GMM-based WGAN is shown in Fig. 6.

Multiple distribution noise is used as the input of the generator to generate fake data with features of individual class and multiple distributions. The discriminator identifies the real data from the generated data and then gives the

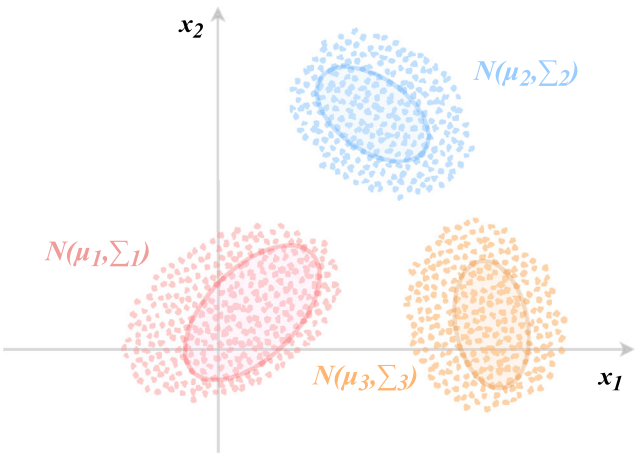


Fig. 5 Under-sampling of GMM-based clustering algorithms

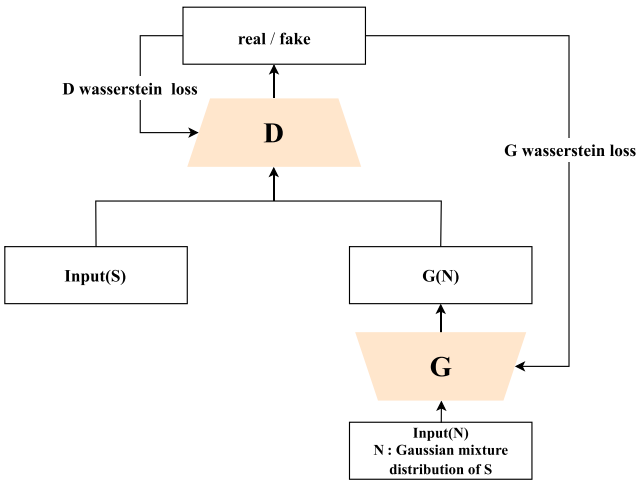


Fig. 6 Structure of GMM-based WGAN

wasserstein loss to update the parameters of the discriminator and generator. The minority classes generation algorithm for GMM-based WGAN is written as Algorithm 1.

Algorithm 1 Generation algorithm of GMM-based WGAN.

Input: Minority class samples $S = \{x_1, x_2, x_3, \ldots, x_n\}$
Output: Generated data $G(N)$

1: Obtain the clustering distribution $Distribution(S)$ of S by the GMM-based clustering algorithm.
2: According to $Distribution(S)$, gaussian mixed noise N consisting of multiple gaussian distributions is generated.
3: G receives the Gaussian mixture noise N, fits the multiple Gaussian distributions separately, and outputs the generated data $G(N)$.
4: D discriminates between the generated data $G(N)$ output by G and the real data S.
5: If D cannot correctly distinguish the generated data $G(N)$ from the real data S, the algorithm terminates and returns the generated data; otherwise, repeat steps (2) to (4).

In this paper, we use GMM-based WGAN as a generative model to generate minority class data. GMM-based WGAN can better fit multiple distributions of minority class samples and alleviate the problem of insufficient rare attack samples.

4.4 Classification module

The traditional intrusion detection technology has low detection accuracy when performing multi-classification tasks. Because the shallow learning model lacks the ability to deal with high-dimensional complex data, it cannot learn the feature representation well. To this end, based on solving the problem of feature redundancy and data imbalance, this paper designs CNN-LSTM [34] that is more suitable for intrusion detection tasks, i.e., a classification module based on CNN and LSTM. The module merges CNN and LSTM together to maximize the performance of intrusion detection.

CNN [35] consists of alternating stacks of convolutional and pooling layers. The convolutional layers are used to extract features and the pooling layers are used to enhance the generality of the features.

LSTM [36] is suitable for processing and predicting important events with long intervals and delays in time series, which solves the problems of gradient disappearance and gradient explosion. It can effectively deal with the attack memory overflow problem in intrusion detection.

CNN-LSTM module can automatically learn the representation of data without manual extraction of complex features. It has a strong potential when facing complex high-dimensional massive data. The structure of CNN-LSTM module is shown in Fig. 7.

5 Experiment and analysis

In this section, the performance of GMM-WGAN-IDS is experimentally evaluated. The evaluation dataset, experimental environment, and experimental procedures are described in detail, compared with classical methods, imbalance processing algorithms, and state-of-the-art intrusion detection methods. In addition, an ablation study is performed.

5.1 Dataset

There are 41 features and one label in NSL-KDD [37] dataset, which solves the problem of data redundancy in KDD Cup 99 dataset. The original training set KDDTrain+ contains 125,973 data and the original test set KDDTest+ contains 22,544 data. There are four main types of attacks: Dos, Probe, R2L, and U2R, as shown in Table 1.

The UNSW-NB15 [38] dataset was created by the Australian Centre for Cyber Security (ACCS) in 2015. It covers a large number of low occupancy intrusion and deep structure network transmission information. It represents the modern network traffic mode, so it is more suitable to simulate the current complex network environment. It has 47 features and one specific attack category label. As shown in Table 1, the dataset contains 2,540,044 samples and 9 different attack types, which are Fuzzers, DoS, Analysis, Reconnaissance, Exploit, Shellcode, Worm, Backdoor, and Generic.

NSL-KDD and UNSW-NB15 are typical high-dimensional imbalanced datasets, which have the characteristics of high feature dimension and large data volume. Among them, most of the data are normal network data and contain only a small amount of attack data. The redundant features and imbalanced data will reduce the detection accuracy and increase the training and detection time. There are also unknown attacks in the test set, which test the generalization ability more.

5.2 Evaluation metrics

Due to the highly imbalanced nature of the dataset, using accuracy alone as an evaluation metric is not sufficient. Because the sample size of the majority class is much larger than that of the minority classes, the classification algorithm can still have high classification accuracy even if all the minority classes are misclassified. Therefore, we adopt Accuracy, Precision, Recall, and F1 Score as evaluation metrics to comprehensively evaluate GMM-WGAN-IDS.

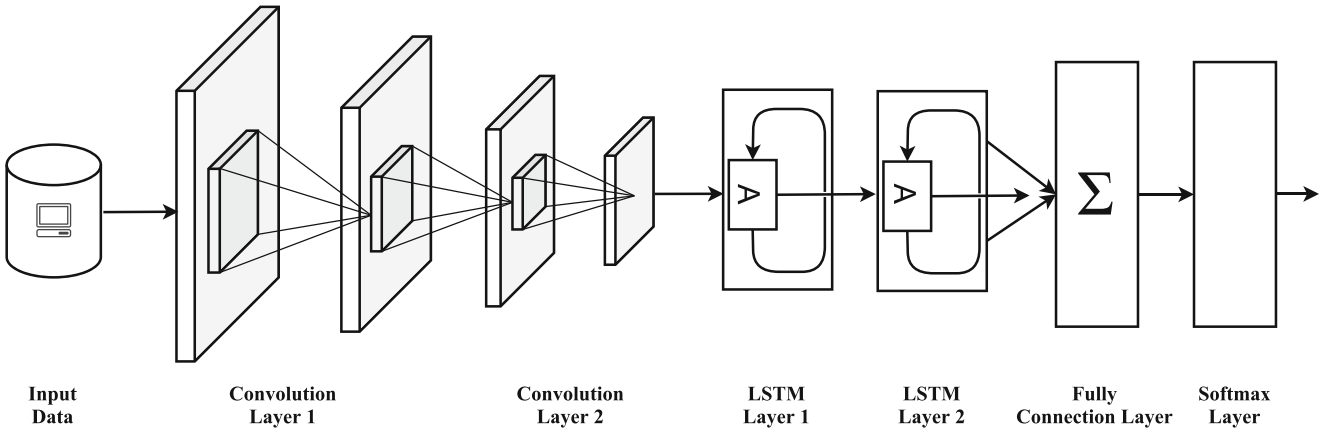


Fig. 7 Structure of CNN-LSTM

Accuracy (ACC): the number of correctly classified samples as a percentage of the total number of samples.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{10}$$

Precision (P): percentage of correctly identified attack data out of all predicted attack data.

$$Precision = \frac{TP}{TP + FP} \tag{11}$$

Recall (R): percentage of correctly identified attack data out of all actual attack data.

$$Recall = \frac{TP}{TP + FN} \tag{12}$$

F1 Score (F1): it is the harmonic mean of Recall and Precision, which is a combined evaluation of Recall and Precision.

$$F1 = \frac{2 * P * R}{P + R} \tag{13}$$

Where True Positive (TP) is attack data correctly identified as attack data by the classifier. False Positive (FP) is normal data incorrectly identified as attack data by the classifier. True Negative (TN) is normal data correctly identified as normal data by the classifier. False Negative (FN) is that the attack data is incorrectly identified as normal data by the classifier.

Table 1 Distribution of the datasets

Dataset	Class	Samples	Percentage
NSL-KDD	Normal	77,054	51.8822 %
	DoS	53,385	35.9453%
	Probe	14,077	9.4783%
	R2L	3749	2.5242%
	U2R	252	0.1696%
UNSW-NB15	Normal	2,218,761	87.35%
	Generic	215,481	8.4833%
	Exploits	44,525	1.7529%
	Fuzzers	24,246	0.9545%
	DoS	16,353	0.6438%
	Reconnaissance	13,987	0.5506%
	Analysis	2677	0.1053%
	Backdoor	2329	0.0916%
	Shellcode	1511	0.0594%
	Worms	174	0.0068%

Table 2 Parameters of GMM-WGAN-IDS

Module	Parameter settings
SAE	Batch size=100
	learning rate=0.001
	Epoch=100
	Activation=Relu
	Optimizer=Rmsprop
GMM-WGAN	Hidden nodes=256
	Batch size=68
	Learning rate=0.00005
	Optimizer=Rmsprop
	activation=tanh
CNN-LSTM	Batch size=100
	Epochs=20
	Optimizer=Adam
	Dropout=0.5
	Activation=Relu

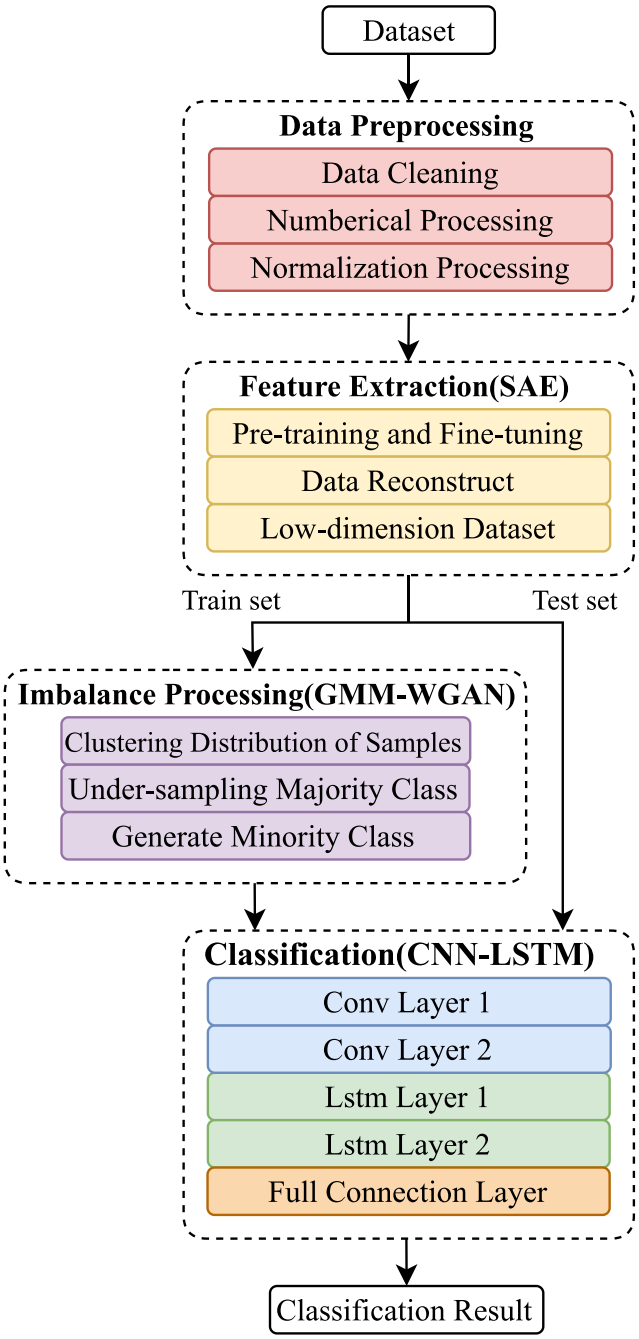


Fig. 8 The flow chart of GMM-WGAN-IDS

5.3 Parameters setting

All experiments in this paper are conducted on an Ubuntu 16.04.10 LTS (64-bit) system with an Intel®Core™i7 processor, 64GB of RAM and a 1TB hard drive. Python is chosen as the programming language. In addition, third-party libraries such as Numpy and Pandas are used in the experiments. Since there is no automatic parameter optimization algorithm available, the parameters and

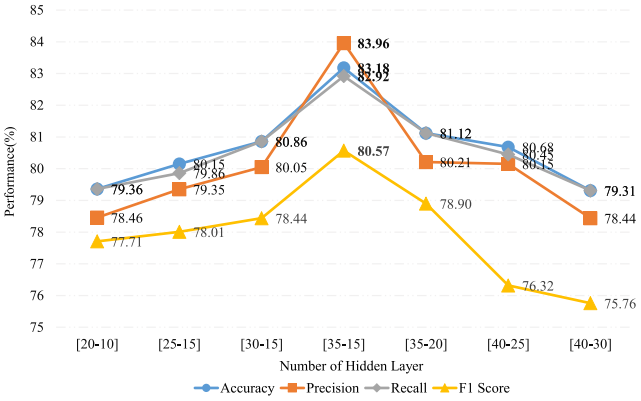


Fig. 9 Performance comparison by number of neurons in hidden layer on the NSL-KDD dataset

structure of the module can only be tuned through iterative experiments or chosen based on previous experience. We conduct a large number of parameter tuning experiments and cross-validation to compare and analyze the module performance under different parameters to determine the final parameters. The main parameters of SAE, GMM-WGAN, and CNN-LSTM are shown in Table 2.

5.4 Experimental procedure

As shown in Fig. 8, the preprocessed data is used as input. The preprocessing includes data cleaning, numerical coding, and normalization.

First, the optimal features are extracted using the SAE module and fed into the classification module CNN-LSTM to evaluate the effect of feature extraction.

Secondly, the GMM-WGAN module is used to alleviate the imbalance of the data, and the original features are used as input to evaluate the imbalance processing effect of the GMM-WGAN module and its various parts.

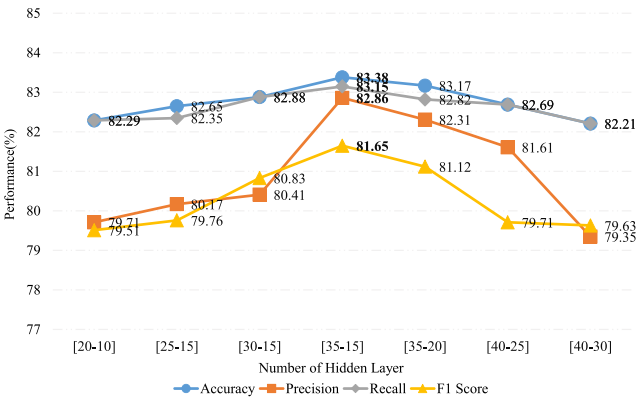
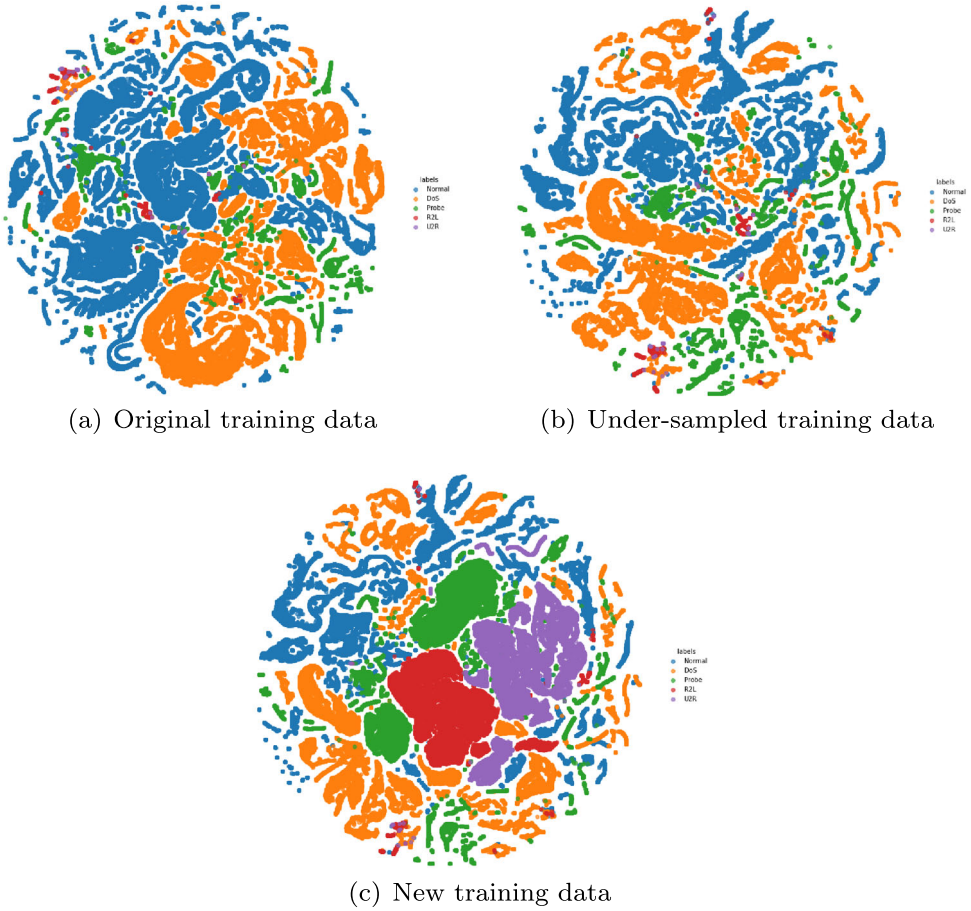


Fig. 10 Performance comparison by number of neurons in hidden layer on the UNSW-NB15 dataset



(a) Original training data (b) Under-sampled training data

(c) New training data

Fig. 11 t-SNE visualization of training data based on the NSL-KDD dataset

Finally, the CNN-LSTM module is used for intrusion detection to validate the overall performance of GMM-WGAN-IDS, while an ablation study is performed to evaluate the effectiveness of each module.

5.4.1 Optimal low-dimensional feature extraction based on SAE

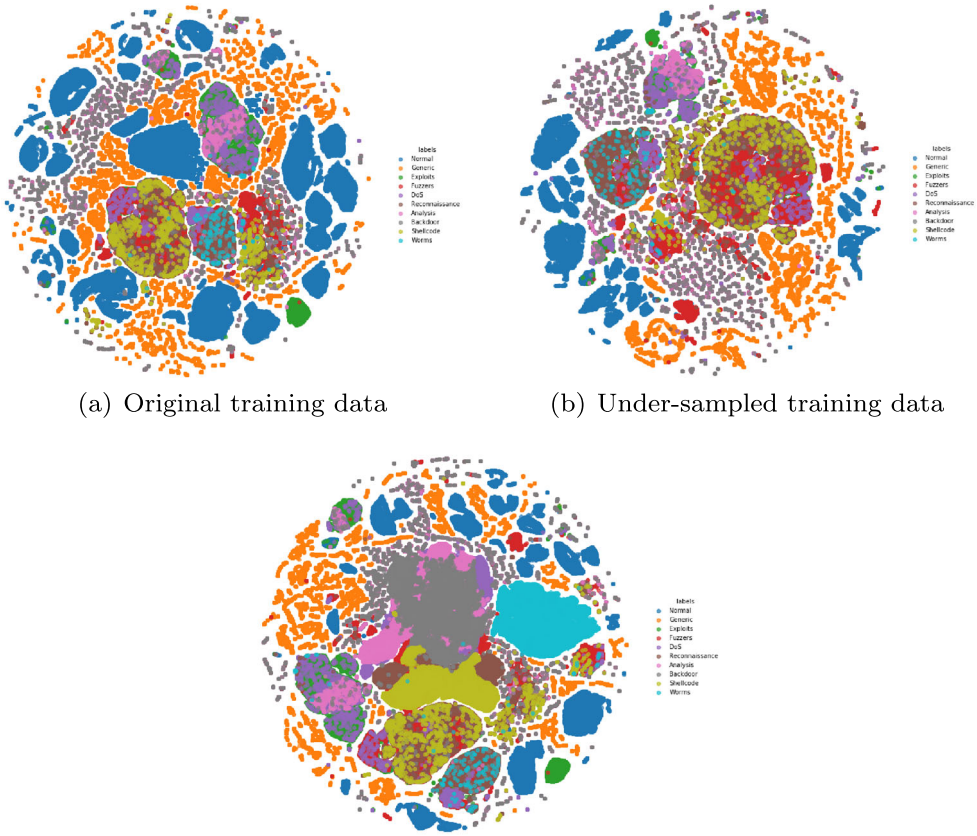
To deeply extract the optimal low-dimensional features from the dataset, two layers of AE stacking are considered to form the SAE. Firstly, the SAE is trained to tune it to the optimal state. The first stage performs bottom-up layer-by-layer unsupervised pre-training. The second stage performs top-down fine-tuning using labeled data. At the same time, noise is added to the original data to assist the training of SAE in order to prevent overfitting during the training process. Finally, the pre-processed data are input to the SAE module for data reconstruction to extract the optimal features and obtain the optimal low-dimensional dataset.

In this paper, the best features are found by compressing the number of hidden layer neurons layer by layer. The extracted optimal features are then input to the CNN-LSTM module for multi-class detection to evaluate the performance of the SAE module. After extensive cross-validation, the best performance of the CNN-LSTM module is achieved when the number of neurons in the two hidden layers of SAE is 35 and 15, respectively, which means that the SAE module extracts the optimal low-dimensional features, as shown in Figs. 9 and 10.

5.4.2 GMM-WGAN based imbalance processing

For imbalance processing, the extracted optimal low-dimensional dataset is divided into training and test sets, and only the training set is processed. When training the GMM-based WGAN, it is not enough to train by discriminator D and generator G alternately against each other, but whether the generated samples conform to the data distribution of the real samples should also be considered. To show the clustering under-sampling effect and the distribution of the generated samples more visually, the t-SNE (t Distributed Random Neighbor Embedding) [39] method is used to visualize the data.

In the original dataset, the data distribution is seriously imbalanced and uneven, as shown in Figs. 11a and 12a. First, the clustering distribution status of each class of



(a) Original training data (b) Under-sampled training data

(c) New training data

Fig. 12 t-SNE visualization of training data based on the UNSW-NB15 dataset

samples in the training set is derived by the GMM-based clustering algorithm. Under-sampling of the majority class samples is performed based on the distribution status. As shown in Figs. 11b and 12b, the redundant data of the majority class samples are reduced and the most representative core data are retained. Secondly, the optimal low-dimensional features extracted by SAE are used as input, and the minority class samples are generated by the GMM-based WGAN according to the multiple distributions of the minority classes. Finally, the majority class core samples are merged with the minority class generated samples to form a new training set. As shown in Figs. 11c and 12c, the generated minority class samples have similarity and maintain diversity with the original minority class samples, which demonstrates the effectiveness of the GMM-WGAN module in alleviating the problems of insufficient rare attack samples and data imbalance.

To further evaluate the performance of the GMM-WGAN module and its various parts, the original high-dimensional features are used as input and subjected to imbalance processing, which is then tested using the CNN-LSTM module. In addition, to better represent the advantages of the GMM-based WGAN in alleviating the shortage of minority class samples, the performance comparison with the original WGAN is also added. The results are shown in Figs. 13 and 14.

5.4.3 CNN-LSTM based intrusion detection and ablation study

The new training set is used as the input in the CNN-LSTM module. Since CNN processes two-dimensional data, we convert the input data into a matrix for intrusion detection. The module contains two convolutional layers and two LSTM layers. The LSTM layer implements the analysis of the data sequence, enabling the classification module to examine the data according to the previously received packets and to perform memory checks. The final layer of the classification module is the fully connected layer, whose performance is evaluated on the test set. Meanwhile, an ablation study of GMM-WGAN-IDS is also performed to verify the effectiveness of each module.

(1) CNN-LSTM Only: We only keep the CNN-LSTM module to perform intrusion detection, which will evaluate the classification performance of the CNN-LSTM module.

(2) w/o GMM-WGAN: We remove the GMM-WGAN module from GMM-WGAN-IDS but keep the SAE module

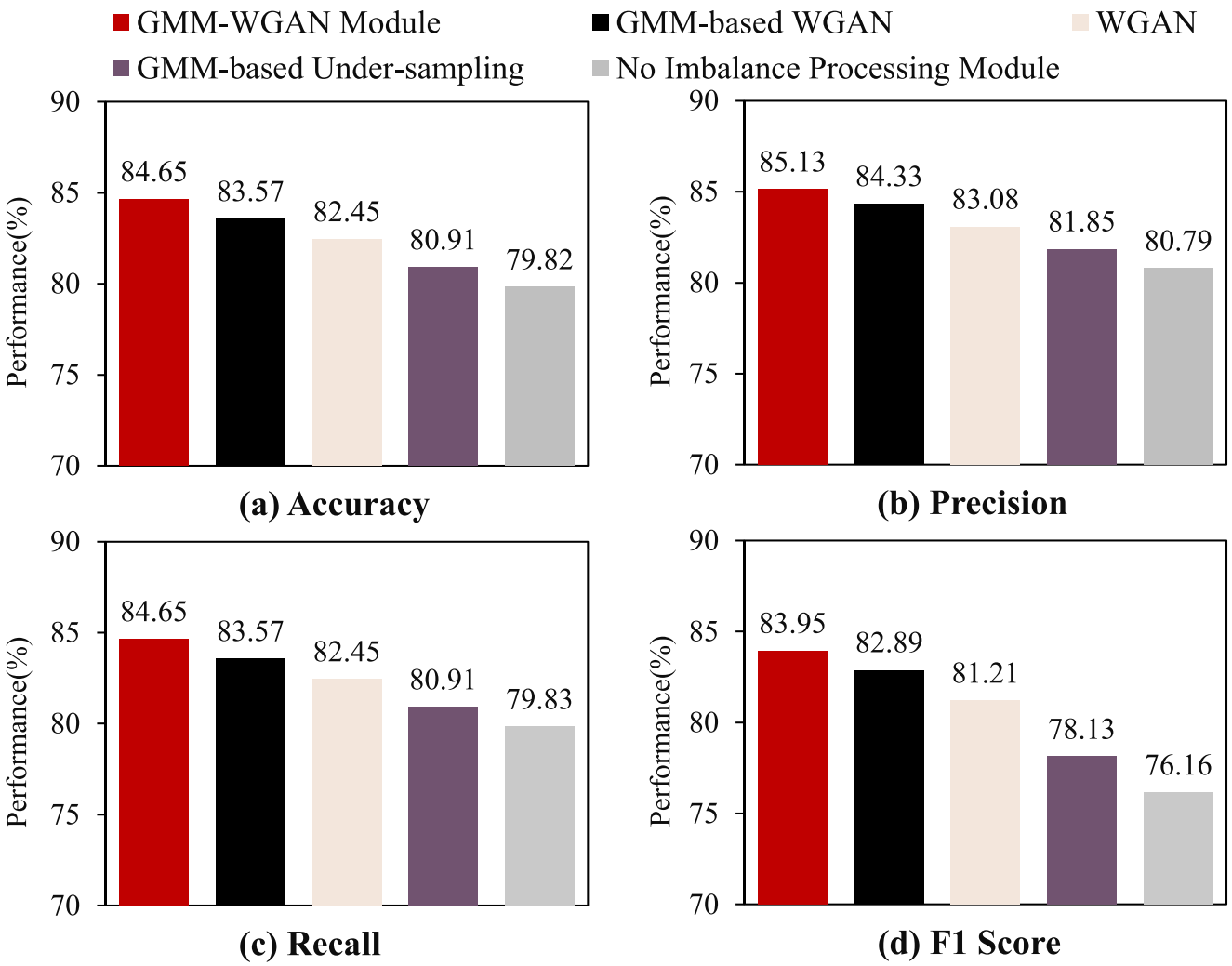


Fig. 13 Classification results after imbalance processing of the NSL-KDD dataset by the GMM-WGAN module and its various parts

and the CNN-LSTM module, which will evaluate the feature extraction capability of the SAE module.

(3) w/o SAE: The SAE module is removed from GMM-WGAN-IDS, which will evaluate the imbalance processing capability of the GMM-WGAN module in the face of high-dimensional data.

The results of the ablation study are shown in Table 3. Each module of GMM-WGAN-IDS shows good performance, especially model (3), which strips off the SAE module to use the original high-dimensional features as the input, and still performs the intrusion detection task well. This is because the GMM-WGAN module has powerful data imbalance processing capability, it can filter the majority class redundant information, and capture the minority class information more widely, to improve the minority class detection accuracy. Comparing Model (2) and Model (1), it can be seen that the SAE module can well achieve optimal feature extraction from high-dimensional data, minimize the interference of redundant features to the classifier, and alleviate the data imbalance phenomenon, to improve the classifier detection performance. When all modules are applied together, the better performance can be obtained. Because the GMM-WGAN module uses the extracted best features as the input instead of the original features. This allows better fitting of the data distribution and more efficient imbalance processing. Therefore, it is necessary to use SAE as the first module of GMM-WGAN-IDS.

5.5 Comparative study

In this section, in order to verify the performance of the GMM-WGAN module in terms of imbalance processing, it is compared with the most representative methods and imbalance processing algorithms. Also, to demonstrate the superiority of GMM-WGAN-IDS, it is compared with the state-of-the-art intrusion detection methods.

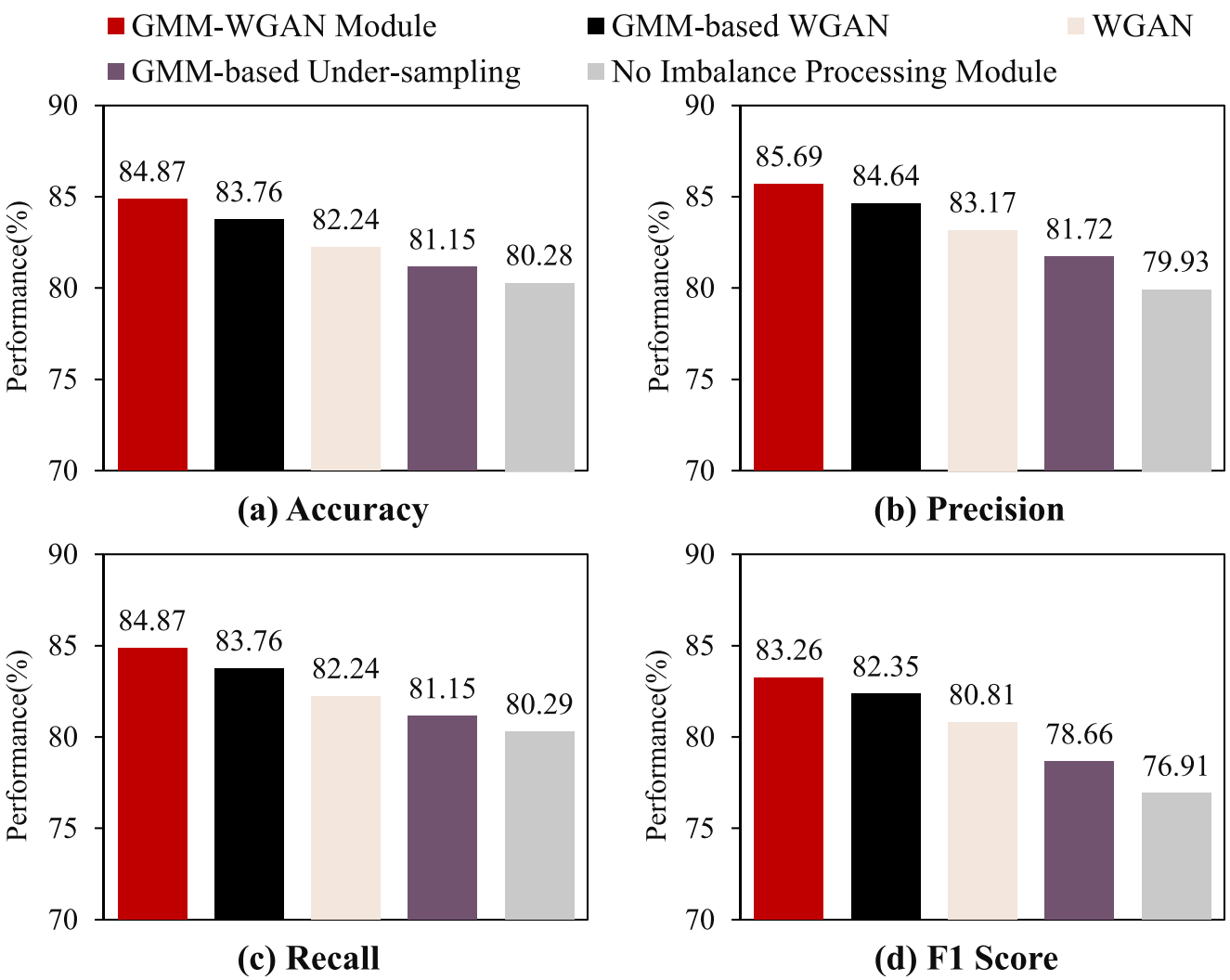


Fig. 14 Classification results after imbalance processing of the UNSW-NB15 dataset by the GMM-WGAN module and its various parts

5.5.1 Comparison with classical methods and imbalance processing algorithms

As shown in Table 4, the GMM-WGAN module is experimentally compared with classical intrusion detection methods and imbalance processing algorithms. The classification performance of classical models such as random forest (RF) [40], support vector machine (SVM) [41] and multilayer perceptron (MLP) [42] are compared on the NSL-KDD dataset. The accuracy of MLP and F1 Score are higher, 77.31% and 75.29% respectively. MLP is the basic model of deep learning and has better classification performance. Second, imbalance processing algorithms such as random under-sampling (RUS) [43], random over-sampling (ROS) [44] and synthetic minority over-sampling technique (SMOTE) [45] are combined with the classical model to

Table 3 Results of the ablation study (%)

Model	Module			NSL-KDD				UNSW-NB15			
	SAE	GMM-WGAN	CNN-LSTM	Acc	Pre	Recall	F1	Acc	Pre	Recall	F1
GMM-WGAN-IDS	✓	✓	✓	**86.59**	**88.55**	**86.59**	**86.88**	**87.70**	**88.46**	**87.70**	**85.44**
(1) CNN-LSTM Only	-	-	✓	79.82	80.79	79.83	76.16	80.28	79.93	80.29	76.91
(2) w/o GMM-WGAN	✓	-	✓	83.18	83.96	82.92	80.57	83.38	82.86	83.15	81.65
(3) w/o SAE	-	✓	✓	84.65	85.13	84.65	83.95	84.87	85.69	84.87	83.26

Table 4 comparison results between GMM-WGAN module and different methods(%)

Model	NSL-KDD				UNSW-NB15			
	Acc	Pre	Recall	F1	Acc	Pre	Recall	F1
RF [40]	74.25	80.21	74.25	69.98	74.35	75.55	74.34	72.38
SVM [41]	73.55	73.76	73.54	69.54	68.49	71.35	68.49	65.13
MLP [42]	77.31	78.52	77.31	75.29	78.32	80.12	78.32	75.98
RUS + RF	76.58	81.20	76.58	72.98	76.66	78.62	76.66	73.97
RUS + SVM	73.46	74.89	73.46	70.22	67.16	72.54	67.15	70.45
RUS + MLP	76.62	78.32	76.62	74.58	77.27	78.25	77.26	75.21
ROS + RF	75.02	80.11	75.02	70.89	77.67	80.49	77.67	75.01
ROS + SVM	74.78	80.23	74.77	73.05	68.32	73.64	68.32	70.00
ROS + MLP	78.10	80.12	78.10	75.18	76.13	79.68	76.12	76.20
SMOTE + RF	74.15	79.96	74.14	69.63	77.93	80.34	77.92	74.95
SMOTE + SVM	74.46	79.34	74.45	72.38	71.50	74.89	71.50	71.77
SMOTE + MLP	78.85	80.65	78.85	76.36	78.59	80.66	78.59	77.10
GMM-WGAN Module	**84.65**	**85.13**	**84.65**	**83.95**	**84.87**	**85.69**	**84.87**	**83.26**

compare the classification performance. The model with the highest accuracy and F1 Score is SMOTE+MLP with 78.85% and 76.36%, respectively. The model with the highest precision is RUS+RF with 81.20%. On the UNSW-NB15 dataset, the accuracy and F1 Score of MLP still perform well relative to the other classical models with 78.32% and 75.98%, respectively. When combined with the imbalance processing algorithm, SMOTE+MLP maintains the best performance with 78.59% and 77.10%. This indicates that the deep learning model has stronger generalization ability and scalability compared to the traditional machine learning model.

On the other hand, imbalance processing algorithms such as RUS, ROS, and SMOTE improve the classification ability of the model less and even degrade the classification performance. That is because they simply add or remove samples without considering the real data distribution. Under-sampling will lose useful information, over-sampling will generate duplicate samples, while SMOTE will increase the possibility of overlap between different classes in the data and make the boundaries of adjacent classes more blurred. GMM-WAGN module can accurately under-sample majority class samples and generate minority class samples based on the data distribution, which reduces redundant samples and avoids the generation of duplicate samples. On the NSL-KDD and UNSW-NB15 datasets, the GMM-WGAN module improves at least 5.8%, 3.9%, 5.8%, 7.5% and 6.2%, 5%, 6.2%, 6.6% in accuracy, precision, recall, and F1 Score.

5.5.2 Comparison with state-of-the-art methods

In addition, GMM-WGAN-IDS is compared with the state-of-the-art intrusion detection methods in recent years. As shown in Tables 5 and 6, GMM-WGAN-IDS achieves the best performance. On the NSL-KDD dataset, the accuracy, precision, recall, and F1 Score of GMM-WGAN-IDS are 86.59%, 88.55%, 86.59%, and 86.88%, which are slightly

Table 5 Comparison results of different detection models on the NSL-KDD dataset (%)

Model	Year	Acc	Pre	Recall	Fl
AE-RL [10]	2019	80.16	79.74	80.16	79.40
DQN [9]	2020	81.80	/	/	/
multi-CNN [12]	2020	81.33	/	/	/
DST-TL [13]	2020	84.60	/	/	/
AE [15]	2020	/	87.85	82.04	81.21
AESMOTE [24]	2020	82.09	/	/	82.43
LCVAE [14]	2021	85.51	/	68.90	80.78
CAFE-CNN [18]	2021	83.34	85.35	83.44	82.60
I-SiamIDS [23]	2021	80.00	/	/	68.34
GMM-WGAN-IDS		**86.59**	**88.55**	**86.59**	**86.88**

Table 6 Comparison results of different detection models on the UNSW-NB15 dataset (%)

Model	Year	Acc	Pre	Recall	F1
DQN [9]	2020	85.09	/	/	/
KG-DBN [11]	2020	86.49	/	/	/
WFEU-FFDNN [17]	2020	77.16	/	/	/
CNN-BiLSTM [22]	2020	77.16	82.63	79.91	81.25
MCNN-DFS [16]	2021	80.51	81	81	81
PSO-LightGBM [19]	2021	86.68	/	/	/
GMM-WGAN-IDS		**87.70**	**88.46**	**87.70**	**85.44**

higher than the literature [14] in terms of accuracy and slightly higher than the literature [15] in terms of precision. On the UNSW-NB15 dataset, the accuracy, precision, recall, and F1 Score are 87.70%, 88.46%, 87.70%, and 85.44%, respectively, which are higher than other intrusion detection methods. It demonstrates the superiority of GMM-WGAN-IDS compared with the state-of-the-art intrusion detection methods.

Overall, the above comparison results fully prove that GMM-WGAN-IDS can improve the detection accuracy of rare attacks and has good generalization ability for the detection of unknown attacks. It shows excellent performance in the face of high-dimensional, complex, and imbalanced data, which achieves intelligent and efficient intrusion detection.

6 Conclusion

In the paper, for high-dimensional imbalanced data, a novel multi-module integrated intrusion detection system GMM-WGAN-IDS is proposed. The system improves the detection performance in the case of feature redundancy and data imbalance, especially the detection rate of rare attacks. Firstly, we propose the feature extraction module, SAE module. The module can obtain optimal low-dimensional features of high-dimensional data by stacking multiple AEs. Secondly, we propose an imbalance processing module, GMM-WGAN. The module reduces the imbalance of the training dataset by majority class under-sampling based on GMM-based clustering algorithm and minority class generation by GMM-based WGAN. Finally, we propose the classification module, CNN-LSTM. The module maximizes the intrusion detection performance by merging CNN and LSTM together. The experimental results show that GMM-WGAN-IDS improves the detection accuracy of overall and rare attack classes and achieves good intrusion detection results. Meanwhile, an ablation study is conducted to verify the effectiveness of each module, which fully demonstrates the powerful detection capability of GMM-WGAN-IDS for high-dimensional imbalanced data. In the future, convolutional generative adversarial network will be designed and applied to intrusion detection to obtain better conclusions. Moreover, the attention mechanism may be applied to feature extraction to make feature extraction more effective.

Acknowledgements We would like to thank our anonymous reviewers for their valuable comments and suggestions. This work is supported by the Scientic Research Funds project of Science and Technology Department of SichuanProvince (No. 2016JY0244, 2017JQ0059, 2019GFW131, 2022JY**), Funds Project of Chengdu Science and Technology Bureau (No. 2017-RK00-00026-ZF), the National Natural Science Foundation of China (No. 61902324), and Sichuan Youth Science and technology innovation research team(2022**).

Declarations

Conflict of Interests There is no conflict of interest between the authors.

References

1. McLennan M (2021) The global risks report 2021 16th edition
2. Garcia-Teodoro P, Diaz-Verdejo J, Maciá-Fernández G, Vázquez E (2009) Anomaly-based network intrusion detection: Techniques, systems and challenges. Comput Secur 28(1-2):18–28. https://doi.org/10.1016/j.cose.2008.08.003
3. Gao Z, Guo L, Guan W, Liu A-A, Ren T, Chen S (2020) A pairwise attentive adversarial spatiotemporal network for cross-domain few-shot action recognition-r2. IEEE Trans Image Process 30:767–782
4. Gao Z, Xuan H-Z, Zhang H, Wan S, Choo K-KR (2019) Adaptive fusion and category-level dictionary learning model for multiview human action recognition. IEEE Internet Things J. 6(6):9280–9293
5. Javaid A, Niyaz Q, Sun W, Alam M (2016) A deep learning approach for network intrusion detection system. In: Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS), pp 21–26
6. Nawir M, Amir A, Yaakob N, Lynn OB (2018) Multi-classification of unsw-nb15 dataset for network anomaly detection system. Journal of Theoretical & Applied Information Technology, 96(15)
7. Khammassi C, Krichen S (2017) A ga-lr wrapper approach for feature selection in network intrusion detection. Comput Secur 70:255–277. https://doi.org/10.1016/j.cose.2017.06.005
8. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539

9. Sethi K, Rupesh ES, Kumar R, Bera P, Madhav YV (2020) A context-aware robust intrusion detection system: a reinforcement learning-based approach. Int J Inf Secur 19(6):657–678. https://doi.org/10.1007/s10207-019-00482-7
10. Caminero G, Lopez-Martin M, Carro B (2019) Adversarial environment reinforcement learning algorithm for intrusion detection. Comput Netw 159:96–109. https://doi.org/10.1016/j.comnet.2019.05.013
11. Tian Q, Han D, Li K-C, Liu X, Duan L, Castiglione A (2020) An intrusion detection approach based on improved deep belief network. Appl Intell 50(10):3162–3178. https://doi.org/10.1007/s10489-020-01694-4
12. Li Y, Xu Y, Liu Z, Hou H, Zheng Y, Xin Y, Zhao Y, Cui L (2020) Robust detection for network intrusion of industrial iot based on multi-cnn fusion. Measurement 154:107450. https://doi.org/10.1016/j.measurement.2019.107450
13. Qureshi AS, Khan A, Shamim N, Durad MH (2020) Intrusion detection using deep sparse auto-encoder and self-taught learning. Neural Comput & Applic 32(8):3135–3147. https://doi.org/10.1007/s00521-019-04152-6
14. Xu X, Li J, Yang Y, Shen F (2020) Towards effective intrusion detection using log-cosh conditional variational autoencoder. IEEE Internet Things J., https://doi.org/10.1109/JIOT.2020.3034621
15. Ieracitano C, Adeel A, Morabito FC, Hussain A (2020) A novel statistical analysis and autoencoder driven intelligent intrusion detection approach. Neurocomputing 387:51–62. https://doi.org/10.1016/j.neucom.2019.11.016
16. Al-Turaiki I, Altwaijry N (2021) A convolutional neural network for improved anomaly-based network intrusion detection. Big Data 9(3):233–252. https://doi.org/10.1089/big.2020.0263
17. Kasongo SM, Sun Y (2020) A deep learning method with wrapper based feature extraction for wireless intrusion detection system. Comput Secur 92:101752. https://doi.org/10.1016/j.cose.2020.101752
18. Shams EA, Rizaner A, Ulusoy AH (2021) A novel context-aware feature extraction method for convolutional neural network-based intrusion detection systems. Neural Comput & Applic, pp 1–19, https://doi.org/10.1007/s00521-021-05994-9
19. Liu J, Yang D, Lian M, Li M (2021) Research on intrusion detection based on particle swarm optimization in iot. IEEE Access 9:38254–38268. https://doi.org/10.1109/ACCESS.2021.3063671
20. ZHAI Y, WANG SP, MA N, YANG BR, ZHANG DZ (2014) A data mining method for imbalanced datasets based on one-sided link and distribution density of instances. ACTA ELECTONICA SINICA 42(7):1311. https://doi.org/10.3969/j.issn.0372-2112.2014.07.011
21. Verma AK, Kaushik P, Shrivastava G (2019) A network intrusion detection approach using variant of convolution neural network. In: 2019 International Conference on Communication and Electronics Systems (ICCES), IEEE, pp 409–416
22. Jiang K, Wang W, Wang A, Wu H (2020) Network intrusion detection combined hybrid sampling with deep hierarchical network. IEEE Access 8:32464–32476. https://doi.org/10.1109/ACCESS.2020.2973730
23. Bedi P, Gupta N, Jindal V (2021) I-siamids: an improved siam-ids for handling class imbalance in network-based intrusion detection systems. Appl Intell 51(2):1133–1151. https://doi.org/10.1007/s10489-020-01886-y
24. Ma X, Shi W (2020) Aesmote: Adversarial reinforcement learning with smote for anomaly detection. IEEE Transactions on Network Science and Engineering, https://doi.org/10.1109/TNSE.2020.3004312
25. Bauder R, Khoshgoftaar T (2018) Medicare fraud detection using random forest with class imbalanced big data. In: 2018 IEEE international conference on information reuse and integration (IRI), IEEE, pp 80–87
26. Kunang YN, Nurmaini S, Stiawan D, Zarkasi A et al (2018) Automatic features extraction using autoencoder in intrusion detection system. In: 2018 International Conference on Electrical Engineering and Computer Science (ICECOS), IEEE, pp 219–224
27. Chen Y, Lin Z, Zhao X, Wang G, Gu Y (2014) Deep learning-based classification of hyperspectral data. IEEE J Sel Top Appl Earth Obs Remote Sens 7(6):2094–2107. https://doi.org/10.1109/JSTARS.2014.2329330
28. Lu H, Li Y, Chen M, Kim H, Serikawa S (2018) Brain intelligence: go beyond artificial intelligence. Mobile Networks and Applications 23(2):368–375. https://doi.org/10.1007/s11036-017-0932-8
29. Goodfellow I, Pouget-Abadie J, Mirza M (2014) Nips. Generative Adversarial Nets 2014:2672–2680
30. Yuqing Z, Ying D, Caiyun L, Kenan L, Hongyu S (2018) Situation, trends and prospects of deep learning applied to cyberspace security. Journal of computer research and development 55(6):1117. https://doi.org/10.7544/issn1000-1239.2018.20170649
31. Jin BS, Han JJ, Ding S, Miao BQ (2018) Em algorithm of the truncated multinormal distribution with linear restriction on the variables. Acta Mathematicae Applicatae Sinica, English Series 34(1):155–162. https://doi.org/10.1007/s10255-018-0733-2
32. Cao Y-J, Jia L-L, Chen Y-X, Lin N, Yang C, Zhang B, Liu Z, Li X-X, Dai H-H (2018) Recent advances of generative adversarial networks in computer vision. IEEE Access 7:14985–15006. https://doi.org/10.1109/ACCESS.2018.2886814
33. Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. In: International conference on machine learning, PMLR, pp 214–223
34. Ding L, Fang W, Luo H, Love PED, Zhong B, Ouyang X (2018) A deep hybrid learning model to detect unsafe behavior: Integrating convolution neural networks and long short-term memory. Automation in construction 86:118–124
35. Krizhevsky A, Sutskever I, Hinton GE (2017) Imagenet classification with deep convolutional neural networks. Commun ACM 60(6):84–90. https://doi.org/10.1145/3065386
36. Zazo R, Nidadavolu PS, Chen N, Gonzalez-Rodriguez J, Dehak N (2018) Age estimation in short speech utterances based on lstm recurrent neural networks. IEEE Access 6:22524–22530. https://doi.org/10.1109/ACCESS.2018.2816163
37. Tavallaee M, Bagheri E, Lu W, Ghorbani AA (2009) A detailed analysis of the kdd cup 99 data set. In: 2009 IEEE symposium on computational intelligence for security and defense applications, IEEE, pp 1–6
38. Moustafa N, Slay J (2015) Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In: 2015 military communications and information systems conference (MilCIS), IEEE, pp 1–6
39. Van der Maaten L, Hinton G (2008) Visualizing data using t-sne. Journal of machine learning research, 9(11)
40. Breiman L (2001) Random forests. Machine learning 45(1):5–32. https://doi.org/10.1023/A:1010933404324
41. Cortes C, Vapnik V (1995) Support vector machine. Machine learning 20(3):273–297. https://doi.org/10.1007/BF00994018
42. Moradi M, Zulkernine M (2004) A neural network based system for intrusion detection and classification of attacks. In: Proceedings of the IEEE international conference on advances in

intelligent systems-theory and applications, IEEE Lux-embourg-Kirchberg, Luxembourg, pp 15–18
43. Tahir MA, Kittler J, Yan F (2012) Inverse random under sampling for class imbalance problem and its application to multi-label classification. Pattern Recogn 45(10):3738–3750. https://doi.org/10.1016/j.patcog.2012.03.014
44. Liu A, Ghosh J, Martin CE (2007) Generative oversampling for mining imbalanced datasets. In: DMIN, pp 66–72
45. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research 16:321–357. https://doi.org/10.1613/jair.953

Jiyuan Cui received the B.S. degree from the School of Computer and Information Engineering, Anyang Normal University, in 2019. He is currently pursuing the M.S. degree with the School of Computer and Software Engineering, Xihua University. His research interests include intrusion detection and deep learning.

Liansong Zong is a associate professor in the school of computer and software engineering, Xihua University. She received the master degree in 2007. Her research interests include machine learning and natural language processing.

Jianhua Xie was born in Linyi, Shandong province. In 2018, he graduated from Weifang University with a bachelor's degree in network Engineering. He is currently studying for a master's degree in software engineering at Xihua University. His research interest covers Chinese related natural language processing.

Mingwei Tang is a professor in the school of computer and software engineering, Xihua University. He received the Ph.D. degree in the school of computer science and Engineering from University of Electronic Science and technology of China in 2012. His research interests include machine learning and natural language processing.