# Variational Few-Shot Learning for Microservice-Oriented Intrusion Detection in Distributed Industrial IoT

Wei Liang, *Member, IEEE*, Yiyong Hu, *Student Member, IEEE*, Xiaokang Zhou, *Member, IEEE*, Yi Pan, *Senior Member, IEEE*, and Kevin I-Kai Wang, *Member, IEEE*

***Abstract*—Along with the popularity of the Internet of Things (IoT) techniques with several computational paradigms, such as cloud and edge computing, microservice has been viewed as a promising architecture in large-scale application design and deployment. Due to the limited computing ability of edge devices in distributed IoT, only a small scale of data can be used for model training. In addition, most of the machine-learning-based intrusion detection methods are insufficient when dealing with imbalanced dataset under limited computing resources. In this article, we propose an optimized intra/inter-class-structure-based variational few-shot learning (OICS-VFSL) model to overcome a specific out-of-distribution problem in imbalanced learning, and to improve the microservice-oriented intrusion detection in distributed IoT systems. Following a newly designed VFSL framework, an intra/inter-class optimization scheme is developed using reconstructed feature embeddings, in which the intra-class distance is optimized based on the approximation during a variation Bayesian process, while the inter-class distance is optimized based on the maximization of similarities during a feature concatenation process. An intelligent intrusion detection algorithm is, then, introduced to improve the multiclass classification via a nonlinear neural network. Evaluation experiments are conducted using two public datasets to demonstrate the effectiveness of our proposed model, especially in detecting novel attacks with extremely imbalanced data, compared with four baseline methods.**

***Index Terms*—Distributed Internet of Things (IoT), few-shot learning, imbalanced data, intrusion detection, out-of-distribution, variational feature representation.**

## I. Introduction

WITH the rapid development of Industrial 4.0, a distributed Internet of Things (IoT) system is becoming a dominant architecture in industrial IoT, which enables elastic interconnection of automation and data analytics across IoT networks [1], [2]. However, security in distributed IoT devices is highly threatened by malicious intruders. These intruders attack the vulnerability of IoT networks, which may break the manufacturing workflow and result in huge economic and reputation losses. To provide highly reliable and flexible services across IoT network, microservice has emerged as a mainstream style of service-oriented software architecture (e.g., Docker Swarm, OpenStack Magnum, Kubernetes, etc.) [3], [4]. The microservice architecture facilitates services deployed on distributed IoT nodes, and may separate complex or intensive computational tasks into lightweight tasks [5]. Empowered by the microservice architecture in distributed IoT systems, the network intrusion detection application can be developed as one kind of microservices, to identify a specific set of malicious intrusions on edge nodes [6], [7].

The extremely large-scale training data required by conventional intrusion detection algorithms lead to large storage space for storing the training data, and high computing resources for training or updating the model [8], [9]. However, distributed IoT nodes can only undertake learning process with small-scale training data, e.g., in conventional deep learning algorithms, the amount of training data may reach the scale of millions, whereas in resource-constrained-IoT nodes, it is limited to less than 10 000. It is impractical to perform such learning scheme on resource-constrained-IoT nodes, because conventional deep learning algorithms will result in poor performance when trained with small-scale data. Therefore, it is essential to design a lightweight learning scheme targeting small-scale training data, to train or update the model more effectively in resource-constrained devices.

In recent years, a range of few-shot learning schemes has been proposed to deal with model training using small-scale dataset [10]. However, these methods mainly focus on binary

Manuscript received June 5, 2021; revised September 4, 2021; accepted September 15, 2021. Date of publication September 28, 2021; date of current version May 6, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 91846301, Grant 71790615, and Grant 62072171 and in part by the National Key R&D Program of China under Grant 2017YFE0117500 and Grant 2019YFE0190500. Paper no. TII-21-2373. *(Corresponding author: Xiaokang Zhou.)*

Wei Liang and Yiyong Hu are with the School of Frontier Crossover Studies, Hunan University of Technology and Business, Changsha 410205, China (e-mail: weiliang@csu.edu.cn; aminkira2019@gmail.com).

Xiaokang Zhou is with the Faculty of Data Science, Shiga University, Hikone 522-8522, Japan, and also with the RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo 103-0027, Japan (e-mail: zhou@biwako.shiga-u.ac.jp).

Yi Pan is with the Department of Computer Science, Georgia State University, Atlanta, GA 30302 USA (e-mail: yipan@gsu.edu).

Kevin I-Kai Wang is with the Department of Electrical, Computer, and Software Engineering, University of Auckland, Auckland 1010, New Zealand (e-mail: kevin.wang@auckland.ac.nz).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TII.2021.3116085.

Digital Object Identifier 10.1109/TII.2021.3116085

class learning problem with balanced dataset [11], which might encounter difficulty when dealing with multiclass classification with imbalanced dataset. Generally, network intrusions are composed of multiple attack classes, and each class may contain multiple types of attack, which leads to following two main challenges: 1) multiclass classification on extremely imbalanced data and 2) out-of-distribution problem caused by emerging attack techniques. Considering a random node in a microservice architecture, classes included in the training set may be significantly imbalanced since limited network packets can be captured by this node. In addition, this node may be attacked by some new attacks, which have never occurred before in the training set. This leads to the out-of-distribution problem in classification tasks, which usually results in poor performance for most existing deep learning algorithms [12], [13].

Although lots of algorithms have been developed for a small-scale imbalanced learning problem, they still suffer from two challenges. First, conventional learning algorithms can hardly tackle the out-of-distribution issue simultaneously in an imbalanced learning problem. These conventional learning algorithms mainly focus on rebalancing the class distributions in a preprocessing procedure, including undersampling in majority classes, oversampling in minority classes, or hybrid of both. These methods can alleviate the imbalance distribution but still suffer from the extremely imbalance ratios (e.g., $> 1:1000$) along with the out-of-distribution problem. Second, existing methods developed for the out-of-distribution problem mainly focus on the binary-class learning task with small out-of-distribution ratios (e.g., $< 20\%$). Simply incorporating existing learning schemes into a few-shot learning framework cannot provide effective solutions to automatically identify new types of attack from small-scale imbalanced training data in distributed industrial IoT systems.

Therefore, this study aims to deal with the specific out-of-distribution issue in few-shot learning with limited imbalanced training data, which can be characterized as follows: 1) extremely imbalanced data with an imbalance ratio greater than 1:1000; 2) a large out-of-distribution ratio greater than 30%; and 3) learning on small-scale training data. In particular, we propose an optimized intra/inter-class structure based variational few-shot learning (OICS-VFSL) model, to enhance the microservice-oriented intrusion detection with imbalanced data in distributed IoT systems. An integrated few-shot learning algorithm is newly designed with variational feature representation. The reconstructed feature embeddings are employed to optimize the intra- and inter-class distance during a variation Bayesian and a feature concatenation process, respectively. An intelligent intrusion detection algorithm is then developed to improve the multiclass classification via a nonlinear neural network. The main contributions are summarized as follows.

1) A lightweight VFSL framework is constructed to deal with the out-of-distribution issue in imbalanced data, which can be applied to improve the microservice-oriented intrusion detection in distributed IoT systems with limited computing resources.
2) An intra/inter-class optimization scheme is designed based on reconstructed feature embeddings, in which the intra-class distance is optimized based on the approximation using KL divergence, while the inter-class distance is optimized based on the maximization of similarities during a feature concatenation process.
3) An intelligent detection algorithm is developed to improve the multiclass classification performance when facing extremely imbalanced data in few-shot learning.

The rest of this article is organized as follows. Section II addresses an overview of related works. Section III introduces the detailed structure and mechanisms of our proposed VFSL model. Experiment and evaluation results are discussed in Section IV. Finally, Section V concludes this article.

## II. Related Work

This section investigates and summarizes the existing works related to this study, including few-shot learning for intrusion detection in distributed IoT, and deep learning techniques for imbalanced data, respectively.

### A. Few-Shot Learning in Distributed IoT

In recent years, more and more research works have focused on deep learning models deployed on resource-constrained edge devices to adapt to dynamic IoT network problems [14], [15]. Due to the constrained resources in edge devices, model training can only use small-scale data, resulting in poor performance of deep learning models.

Few-shot learning [16] is a new deep learning scheme that has recently emerged, which allows the model to obtain good performance without using large-scale training data. Yang *et al.* [17] proposed a few-shot learning model based on the Siamese network, which measured the similarity between two feature embeddings through Mahalanobis distance. This could improve the accuracy and robustness of sentiment analysis on the text from IoT devices. Zhou *et al.* [18] introduced a few-shot learning model with a Siamese convolutional neural network (CNN) structure, which could alleviate the overfitting problem, and improve the accuracy of intelligent anomaly detection in industrial CPS. Current research works have not considered any other complex situations only with small-scale data. To deal with a typical problem of imbalanced data in intrusion detection system, Bedi *et al.* [19] addressed a new type of intrusion detection system based on a Siamese neural network, which could detect minority attacks without using conventional class balancing techniques.

Most of few-shot learning algorithms mainly focus on binary classification in the imbalanced learning problem. The out-of-distribution issue in small-scale data rarely attract their attention. Solving this kind of problem may effectively enhance the applicability of few-shot learning algorithms in edge computing environments.

### B. Deep Learning Techniques for Imbalanced Data

Deep imbalanced learning aims to mitigate model training bias toward majority classes by increasing the importance of minority classes. Existing methods [20] can be divided into algorithm/model level, data level, and cost-sensitive learning level, respectively.

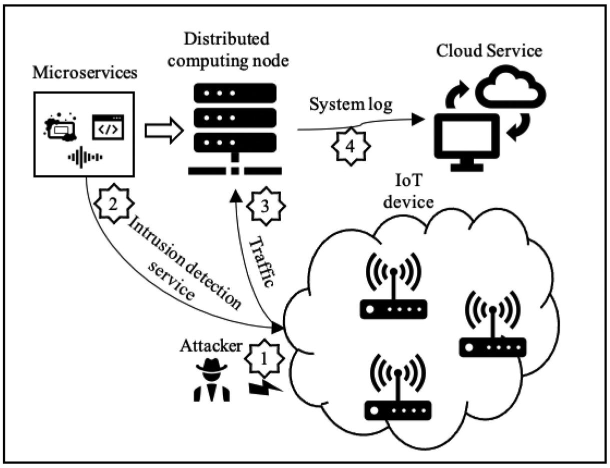


Fig. 1. Overview of microservice-oriented network intrusion detection in distributed IoT systems.

The algorithm/model level methods focus on the modification of training algorithms to obtain better classification performance. Zhou *et al.* [21] developed a variational long short-term memory (LSTM) model to deal with the high-dimensional data, in which three loss functions were quantified and integrated together with a reconstructed hidden variable to improve the anomaly detection performance. In the data level, some sampling-based methods [22], [23], including undersampling, and oversampling, have drawn significant attentions to deal with the imbalanced dataset. Santoso *et al.* [24] investigated different variants of undersampling and oversampling techniques, and claimed that they have different effects in different scenarios.

On the other hand, many researchers [25] have used the generative adversarial network algorithm to amplify specific data distribution, which could reduce the meaningless data or noise generated by the conventional data sampling methods. The cost-sensitive learning [26] is used to reinforce bias against rare but valuable cases of instability. These methods aim to maximize the loss function associated with the dataset to improve the classification performance. However, the actual cost of observing each type of error is usually unknown; thus, they cannot accurately solve the model deviations.

## III. Variational Few-Shot Learning for Intelligent Intrusion Detection

In this section, we introduce the overview of a network intrusion detection framework with microservice architecture in distributed IoT systems. The OICS-VFSL model is then proposed and discussed with its detailed implementation.

### A. Problem Formalization

Fig. 1 shows the overview of microservice-based network intrusion detection in distributed IoT systems. Specifically, IoT devices collect physical data from industrial environments and production processes, and build a bridge between physical and virtual space. Distributed computing nodes are usually responsible for providing complex business services and data analysis based on IoT devices. Typically, attackers may invade distributed IoT systems by sending the malicious code to interrupt the normal operation of IoT devices. As different microservices deployed in distributed computing nodes, the network monitoring service is used to collect the traffic data from IoT devices, and the intrusion detection service is employed to analyze the traffic and feedback the result to IoT devices, in order to ensure network security among all the IoT devices.

Given an intrusion detection problem in distributed IoT systems, two datasets $D_{\text{train}} = \{d_1, d_2, \ldots, d_K\}$ and $D_{\text{test}} = \{d_1, d_2, \ldots, d_K, \ldots, d_N\}$ are taken into consideration as the training set and test set, which contains $K$ types of attacks and $N$ types of attacks, respectively. $K < N$ means that the number of attack types in the test set is larger than that in the training set, which indicates a specific out-of-distribution problem in intrusion detection. $d_K = \{(x_K^{(1)}, y_K^{(1)}), (x_K^{(2)}, y_K^{(2)}), \ldots, (x_K^{(n)}, y_K^{(n)})\}$ is a subset of the $K$th attack type. We select a set of samples from each subset of the training set to form a support set $X_S = \{(x_1^{(1)}, y_1^{(1)}), \ldots, (x_1^{(C)}, y_1^{(C)}), \ldots, (x_K^{(C)}, y_K^{(C)})\}$, and the corresponding query set $X_Q$. It is assumed that $\|X_S\| + \|X_Q\| \ll \|D_{\text{train}}\|$, to demonstrate a few-shot learning scenario, which means the amount of attack data in the few-shot learning model is far less than that in a traditional deep learning model. We randomly select $K$ attack types, and each type with $C$ labeled sample data, to define a K-way-C-shot learning model.

In particular, we employ a one-shot learning model, to tackle the above intrusion detection problem in the distributed IoT network based on a microservice scenario. A generic framework of the proposed OICS-VFSL model is shown in Fig. 2.

The OICS-VFSL model consists of the following two major modules: intra-class distance optimization based on variational feature representation and inter-class distance optimization based on feature concatenation. Specifically, following a variational feature representation, which is used to learn and refine the high-level features from the original data in an LSTM structure, the intra-class distance is optimized through the approximation based on KL divergence during a variation Bayesian learning process. On the other hand, the inter-class distance is optimized during a feature concatenation process, in which we concatenate the input data with the data in support set based on their reconstructed feature embeddings, and maximize their similarity to the correct class using a nonlinear neural network. In summary, based on the intra-class and inter-class distance optimization using reconstructed feature embeddings, our proposed model can efficiently identify the new attack types when dealing with an imbalanced dataset with the out-of-distribution issue in few-shot learning.

### B. Variational Feature Representation Based Intra-Class Distance Optimization

Given $x_i$ as one input data, the feature embedding $w_i$ extracted from the LSTM network can be represented as follows:

$$w_i = f(x_i, \xi) \tag{1}$$

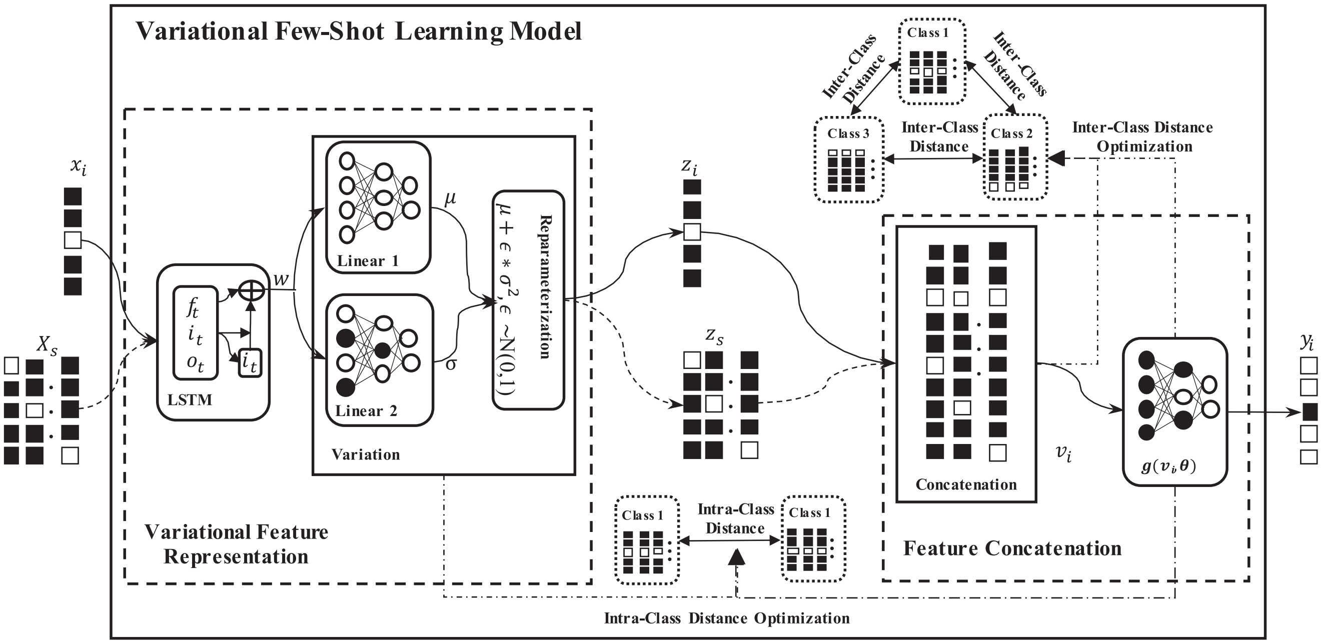


Fig. 2. Structure of a VFSL network.

where $w_i$ indicates the extracted feature embeddings, while $\xi$ denotes the corresponding parameters of $f$.

To obtain the variational feature representation, we consider a joint distribution $p(x, z, y)$, where $x$ is the input data, $z$ denotes the learned feature representation, and $y$ is the corresponding label of data. Due to the large data parameter space, the posterior distribution $p(z|x)$ cannot be directly calculated. Inspired by the variational autoencoder, the variation Bayesian method is used to construct a distribution $q(x, z, y)$ to approximate $p(x, z, y)$. In other words, we use $q(z|x)$ to approximate $p(z|x)$, and the corresponding KL divergence can be described as follows:

$$\begin{aligned}\mathrm{KL}(p(x,z,y)\|q(x,z,y)) &= \sum_y \iint p(z,y|x)\widetilde{p}(x)\\ &\qquad \ln\frac{p(z,y|x)\widetilde{p}(x)}{q(x|z,y)q(z,y)}dzdx\\ &= E_{x\sim\widetilde{p}(x)}[-\log q(x|z)\\ &\quad + \mathrm{KL}(p(y|z)\|q(y))\\ &\quad + \sum_y p(y|z)\log\frac{p(z|x)}{q(z|y)}] \end{aligned} \tag{2}$$

where $z \sim p(z|x)$, $p(z|x)$ obeys to a normal distribution with a mean of $\mu(x)$ and variance of $\sigma(x)^2$.

The reparameterization method is employed to approximate the posterior distribution $p(z|x)$. We introduce a parameter $\epsilon \sim N(0, 1)$, and input the feature embedding $w$ into two different nonlinear neural networks, which are used to calculate $\mu(w)$ and $\sigma(w)^2$, respectively. Finally, we can obtain $q(z|x)$ to approximate $p(z|x)$, which can be calculate as follows:

$$z = \mu(w) + \epsilon * \sigma(w)^2, z \sim q(z|x). \tag{3}$$

We go further to explain (2) in detail. $-\log q(x|z)$ is used to ensure that features in $x$ can be maximally retained in $z$. Importantly, we assume that $y$ obeys to a uniform distribution; thus, the reconstructed $z$ will follow a specific normal distribution correlated to $y$. Then, the distribution of $z$ can be balanced, following the abovementioned feature representation process based on $\mathrm{KL}(p(y|z)\|q(y))$. In addition, $p(z|x)$ represents the probabilistic result after feature extraction from LSTM, and $p(y|z)$ represents the probabilistic result after the classifier; therefore, only $\sum_y \frac{p(y|z)}{q(z|y)}$ needs to be optimized in $\sum_y p(y|z)\log\frac{p(z|x)}{q(z|y)}$.

Theoretically, following the abovementioned feature representation process, the distance between each $z$ belonging to the same class labeled by $y$ can be effectively shortened, while maximally retaining the features extracted from $x$, which can be called intra-class distance optimization in our few-shot learning model.

### C. Feature Concatenation Based Inter-Class Distance Optimization

Traditional methods usually define and calculate the distance between two feature embeddings from the input data $x_i$ and $x_j$ based on the pairwise Euclidean distance, which needs large amount of data during the training process. In contrast, we employ the feature concatenation to optimize the distance between the two feature embeddings in a few-shot learning strategy.

Given an input data $(x_i, y_i)$, a support set $X_S = \{(x_1, y_1), \ldots, (x_K, y_K)\}$, and the corresponding reconstructed $z_i$ and $Z_S = \{z_1, \ldots, z_K\}$, we concatenate $z_i$ with each $z_k$ in $Z_S$, which can be formulated as follows:

$$v_i^k = \mathrm{con}(z_i, z_k) \tag{4}$$

where con denotes a concatenation function, and $v_i^k$ denotes a new vector formed by concatenate $z_i$ and each $z_k$ in $Z_S$.

A nonlinear neural network $g$ takes $v_i^k$ as the input to calculate the distance between $z_i$ and $z_k$. The detailed calculation can be described as follows:

$$p_i^k = g(v_i^k, \theta) = \frac{\exp(\theta * v_i^k)}{\sum_{K=1}^{|Z_S|} \exp(\theta * v_i^K)} \tag{5}$$

where $\theta$ denotes the parameter in $g$, and $p_i^k$ denotes the similarity between $z_i$ and $z_k$.

After calculating all the elements in $Z_S$, a similarity set $p_i = \{p_i^1, p_i^2, \ldots, p_i^K\}$ can be obtained. The predicted label $\hat{y_i}$ can be the biggest similarity value in $p_i$, which is represented as follows:

$$\hat{y_i} = \text{argmax}(p_i). \tag{6}$$

In particular, since we adopt one-shot learning in our model, which means each class only contains one data sample in the support set, and such data sample may be represented as the center of this class for similarity calculation. Following the feature concatenation-based calculation we discussed previously, when we maximize the similarity between each $z$ and the class labeled by its predicted $\hat{y}$, the distance from this class to other classes will be stretched relatively, which can be called inter-class distance optimization in our few-shot learning model.

### *D. Intelligent Intrusion Detection Algorithm*

Based on our model, the distribution of extracted feature embeddings will be optimized to a normal distribution $N(0, I)$ with a mean of 0 and variance of $I$. We further take the KL divergence between $N(\mu, \sigma^2)$ and $N(0, I)$ as an additional loss. The expression of KL divergence loss $\text{KL}(N(\mu, \sigma^2) \| N(0, I))$ can be described as follows:

$$\ell_{\text{KL}} = \frac{1}{2}(-\log(\sigma^2) + \mu^2 - 1). \tag{7}$$

In general, for each class in the support set, the average of feature embeddings in each class will be calculated as the class center. We use the cross information entropy loss to ensure the optimal similarity between the reconstructed feature embedding and the class center. The cross information entropy loss $\ell_c$ is defined as follows:

$$\ell_c = -\sum_{i=0}^{k} y_i \log(p_i) \tag{8}$$

where $y_i$ denotes the label of input data $x_i$.

A scalar $\lambda$ is used to balance the two losses discussed previously. Thus, the final loss function can be described as follows:

$$\ell_{\text{OICS-VFSL}} = \ell_{\text{KL}} + \lambda * \ell_c. \tag{9}$$

The detailed intelligent intrusion detection algorithm is illustrated in Algorithm 1. First, the feature space is mapped to the feature representation space by the LSTM encoder, and the feature representation is optimized via KL divergence during a variation Bayesian-based learning process. Second, the similarities between the reconstructed feature embeddings of the input data and that of the support set are calculated based on the feature concatenation. Finally, the label of the input data can be predicted from the classifier as the output of the few-shot learning.

**Algorithm 1:** Intrusion detection algorithm based on OICS-VFSL.

```
Input: $X_Q$, $X_S$, $D^{\text{test}}$.
Output: A trained intrusion detection model $M$
Initialize the model $M$
Initialize the iteration count $T$, batch size $N$, threshold $\delta$
for q = 1 to T do
  for $x_i$ in $X_Q$ do
    Transfer $x_i$ into embedding vector $w_i$ via LSTM encoder by (1).
    Transfer $w_i$ into a unified feature representation $z_i$ via (3)
    for $x_k$ in $X_S$ do
      Transfer $x_k$ into embedding vector $w_k$ via LSTM encoder by (1).
      Transfer $w_k$ into a feature embeddings $z_k$ via (3)
        Cascade $z_i$ and $z_k$ into a new vector $v_i^k$ by (4)
        Calculate the similarity between $z_i$ and $z_k$ by (5)
    Predict the class label by (6)
  Update $M$ to minimize $\ell_{\text{OICS-VFSL}}$ by (9)
  if $\ell_{\text{OICS-VFSL}} < \delta$ then
    break;
return $M$
```

## IV. Experiment and Analysis

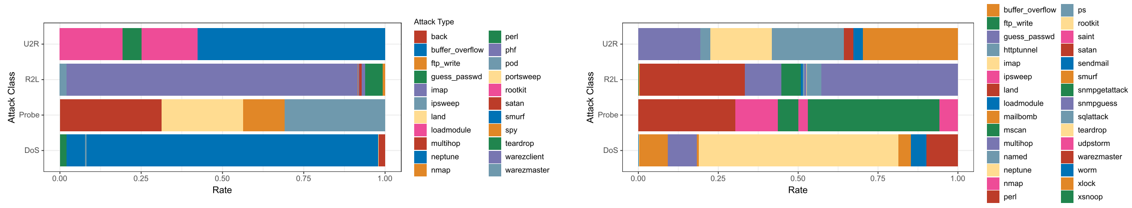
### *A. Dataset and Experiment Design*

Experiments are conducted based on two public datasets: NSL-KDD and CIC-IDS 2017, to evaluate the performance of the proposed method.

As for NSL-KDD, it removes many duplicate and redundant records from the KDD Cup99 dataset, which can be viewed as an imbalanced dataset, and the details can be found in Fig. 3. Attacks are divided into four classes: U2R, R2L, Probe, and DoS, respectively. Each class contains multiple types of attack. The imbalanced ratios of normal data to R2L, U2R, Probe, and DoS are 1:1250, 1:67, 1:6, and 1:2, respectively. It is observed that attack types in the training set are significantly less than those in the test set. The details are shown in Fig. 4. In addition, the ratios of new types of attacks in R2L, U2R, Probe, and DoS are 75%, 19%, 54.3%, and 23%, respectively, which can be viewed as a specific out-of-distribution problem with a large-scale dataset.

CIC-IDS 2017 dataset is proposed by Canadian Institute for Cybersecurity, which is the newest intrusion detection dataset, and cover the necessary conditions for updated attacks in DoS, DDoS, XSS, SQL Injection, Brute Force, Botnet, Infiltration, FTP-Patator, PortScan, SSH-Patator, and HeartBleed. The statistical details are shown in Table I. We noticed that it is a large-scale imbalanced dataset with many novel attacks.

In summary, NSL-KDD is a large-scale and imbalanced dataset with the out-of-distribution problem, while CIC-IDS



(a) (b)

Fig. 3. Attack descriptions in NSL-KDD. (a) Training set. (b) Test set.

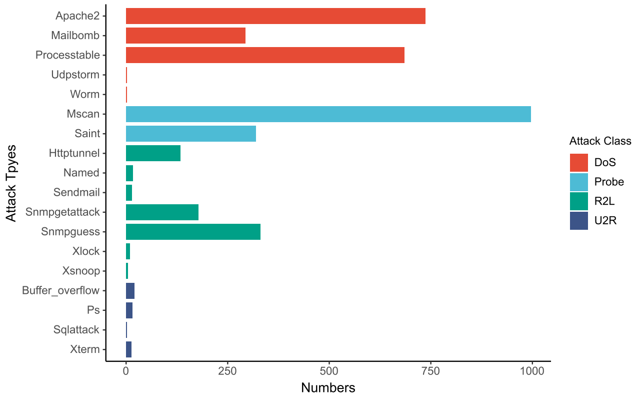


Fig. 4. Out-of-distribution in NSL-KDD.

TABLE I
STATISTICS OF ATTACKS IN CIC-IDS 2017

| *Dataset*<br>*Type* | Training Dataset (Imbalanced Ratio) | Test Dataset |
|---|---|---|
| Normal | 1589800 (-) | 681520 |
| DoS | 176231 (1:10) | 75481 |
| DDoS | 89626 (1:20) | 38399 |
| XSS | 456 (1:5000) | 196 |
| SQL Injection | 14 (1:125000) | 7 |
| Brute Fore | 1066 (1:1430) | 441 |
| Botnet | 1356 (1:1250) | 600 |
| Infiltration | 27 (1:50000) | 9 |
| FTP-Patator | 5558 (1:285) | 2377 |
| PortScan | 111239 (1:14) | 47565 |
| SSH-Patator | 4131 (1:500) | 1766 |
| HeartBleed | 9 (1:200000) | 2 |

2017 is a large-scale and imbalanced dataset but includes many novel attacks. We use these two datasets to demonstrate the performance of our proposed model in solving a common imbalanced issue and a specific out-of-distribution issue, respectively.

Considering the imbalanced issue among minority classes and majority classes in a multiclass classification problem, a set of metrics, including detection rate (DR), false acceptance rate (FAR), and $F_1$, is employed to evaluate the performance of our model. DR is a critical metric to verify whether data are correctly and efficiently classified. FAR is a crucial metric that indicates the proportion of the data that is classified as normal in predicted results but actually is an attack in the test dataset. Four baseline methods used in conventional deep learning, optimization of resource consumption, imbalanced learning, and out-of-distribution problem are selected and summarized as follows.

1) *LuNet [27]:* A hierarchical CNN + RNN neural network model, which used CNN and RNN to learn the network traffic data synchronously with increasing granularity, and extract the spatial-temporal characteristics of the data.
2) *DeRol [28]:* A reinforcement learning model, which mapped the current state of a single learning process to maximize the target function and provided a solution for practical artificial intelligence with limited resources.
3) *Siamese-NN [19]:* An intrusion detection model based on Siamese NN, which detected minority attacks in imbalanced data without using conventional class balancing techniques.
4) *Deep-MCDD [29]:* A multiclass data description method, which could learn the class-conditional distributions and was applied to solve the out-of-distribution problem in multiclass classification.

Implementations of these baseline methods in PyTorch are used in our evaluation experiments.

## B. Analysis on Model Parameter

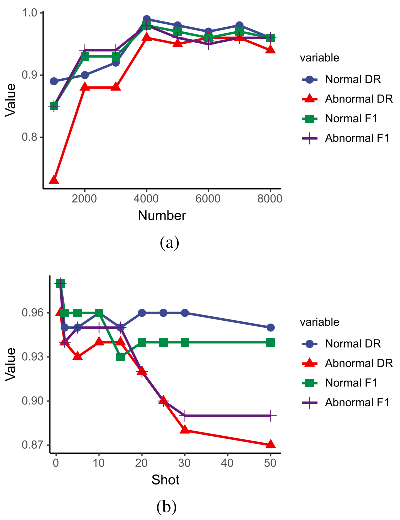
In the few-shot learning methods, the number of training samples and the number of shots are the key parameters to measure the pros and cons of the model. Thus, the selection of these parameters may significantly influence the model. We conduct a range of experiments to test the effect of these parameters in terms of the sensitivity.

First, we examine the impact caused by the number of training data. The result is shown in Fig. 5(a). It is noticed that the performance is significantly improved after increasing the training data. If the amount of data is too small, it cannot provide sufficient information for classification. When using 4000 training data, the model can achieve the best performance and then be stabilized.

Then, we evaluate the influence of the number of shots. As shown in Fig. 5(b), it is observed that our proposed model can achieve the best results with one shot, and the effect will get worse as the number of shots increases.

## C. Performance on Feature Representation Efficiency

Principal component analysis (PCA) is utilized to evaluate the feature representation efficiency based on our proposed model, and DNN and LSTM are employed as the baseline methods for comparison using the NSL-KDD dataset. Performances of feature representation are compared in Fig. 6, in which different colors and shapes describe the clustering results of different



(a)

(b)

Fig. 5. Parameter testing result. (a) Testing on training data. (b) Testing on training shots.

TABLE II
DETECTION PERFORMANCE COMPARISON ON NSL-KDD

| Method | DR | | $F_1$ | | FAR |
|---|---|---|---|---|---|
| | Normal | Attack | Normal | Attack | |
| LuNet | **1.0** | 0.78 | 0.86 | 0.88 | 0.24 |
| DeRol | 0.74 | **1.0** | 0.85 | 0.84 | 0.25 |
| Siamese-NN | 0.98 | 0.80 | 0.87 | 0.88 | 0.21 |
| Deep-MCDD | **1.0** | 0.73 | 0.85 | 0.84 | 0.26 |
| OICS-VFSL | **1.0** | 0.98 | **0.97** | **0.98** | **0.05** |

The boldface highlights the best results.

TABLE III
DETECTION PERFORMANCE COMPARISON ON CIC-IDS 2017

| Method | DR | | $F_1$ | | FAR |
|---|---|---|---|---|---|
| | Normal | Attack | Normal | Attack | |
| LuNet | 0.94 | 0.95 | 0.97 | 0.87 | **0.01** |
| DeRol | 0.96 | 0.90 | 0.97 | 0.87 | 0.05 |
| Siamese-NN | 0.94 | 0.96 | 0.96 | 0.87 | **0.01** |
| Deep-MCDD | 0.96 | 0.98 | 0.98 | 0.92 | 0.05 |
| OICS-VFSL | **0.98** | **0.99** | **0.98** | **0.95** | **0.01** |

The boldface highlights the best results.

TABLE IV
ADVERSARIAL RESULT ON NSL-KDD

| Method | Origin | Adversarial |
|---|---|---|
| DNN | 0.78 | 0.73 |
| LuNet | 0.81 | 0.74 |
| DeRol | 0.86 | 0.80 |
| Siamese-NN | 0.87 | 0.77 |
| Deep-MCDD | 0.85 | 0.80 |
| OICS-VFSL | 0.92 | 0.91 |

classes. We discuss the feature representation performances in terms of the intra-class distance and inter-class distance optimizations, respectively.

First, it can be observed that distributions of the data in the same class are relatively scattered in both DNN and LSTM. Benefiting from the special design in the intra-class distance optimization, data in the same class are compactly distributed close to the class center. Second, it can be easily found that distributions of data in the different classes are overlapping in both DNN and LSTM. Our proposed method has very clear boundaries in different classes because of the inter-class distance optimization.

These results evidently indicate that our model can achieve a better performance on feature extraction, because we design a feature representation operator $\sum_y p(y|z)\log\frac{p(z|x)}{q(z|y)}$ to optimize the intra-class distance, and apply the feature concatenation to optimize the inter-class distance.

### D. Performance on Intrusion Detection

We go further to evaluate the advantages of our model in intrusion detection compared with LuNet, DeRol, Siamese-NN, and Deep-MCDD based on datasets NSL-KDD and CIC-IDS 2017, respectively.

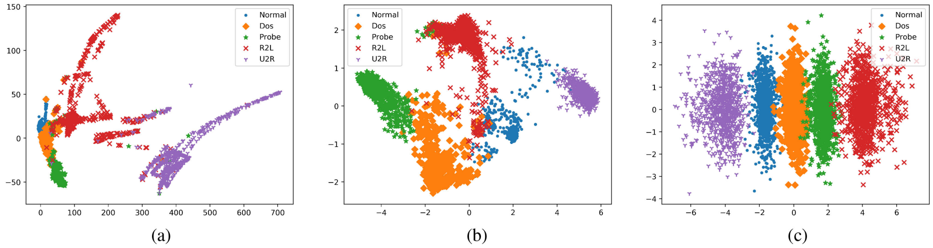
Tables II and III demonstrate the binary classification results for intrusion detection according to DR, FAR, and $F_1$ based on NSL-KDD and CIC-IDS 2017, respectively. Our model outperforms all the other baseline methods in both of the two datasets, especially in NSL-KDD, which is an imbalanced dataset with the out-of-distribution problem.

We further conduct multiclass classification experiments to demonstrate the advantages of our model and explore the reasons for the poor performance of the baseline methods based on NSL-KDD and CIC-IDS 2017, respectively. We choose the experiment data with the imbalance ratio within 1:2000, to investigate the model's performance in large-scale imbalance data.

The multiclass classification results of DR based on CIC-IDS 2017 are shown in Fig. 7. The imbalanced ratios of normal data to SSH Patator, PortScan, FTP Patator, DoS, DDoS, Brute Force, and Botnet are 1:285, 1:14, 1:20, 1:18, 1:16, 1:1430, and 1:1250, respectively. We observed that the greater the imbalance scale, the lower the DR results of the baseline methods. Our model achieves better results on different imbalance scales. The results indicate that the baseline methods might alleviate the imbalanced problem when facing small-scale data, but the large scale of data will cause their models to collapse when dealing with the imbalanced problem.

On the other hand, the multiclass classification results of DR on NSL-KDD is shown in Fig. 8. Our model achieves the best results in R2L and Probe, while the other baseline methods achieve relatively poor performances, due to the large-scale, imbalanced, and out-of-distribution issues in NSL-KDD. The results prove that our proposed method can effectively alleviate the out-of-distribution problem in a large-scale imbalanced dataset.

Furthermore, we employ the improved adversarial method JSMA [30], and generate adversarial samples based on NSL-KDD, to evaluate the robustness of our model. We use DNN as the black-box attack model, and set the penalty strength of JSMA as 0.1. As shown in Table IV, adversarial samples cause an approximately 5% loss of the accuracy of the trained DNN model, which are approximately 7%, 6%, 10%, and 5% of that



(a) (b) (c)

Fig. 6. Feature representation visualization based on PCA. (a) DNN. (b) LSTM. (c) OICS-VSFL.

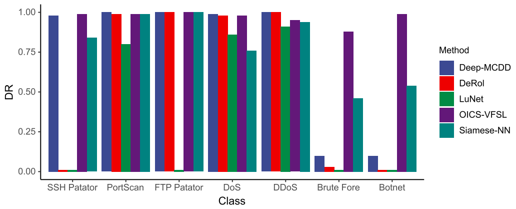


Fig. 7. Detection performance on imbalanced data.

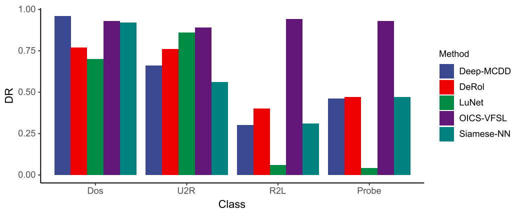


Fig. 8. Detection performance on the out-of-distribution issue.

of LuNet, DeRol, Siamese-NN, and Deep-MCDD, respectively. In comparison, our model only results in less than 1% loss of accuracy. This result demonstrates the robustness of our proposed model, which can effectively resist adversarial attacks to a certain extent.

## V. Conclusion

In this article, to cope with the out-of-distribution issue in imbalanced data, we proposed an OICS-VFSL model for microservice-oriented intrusion detection, which could be applied on resource-constrained computing nodes across distributed IoT systems.

We first presented an integrated few-shot learning framework with variational feature representation, which mainly included the following two basic functions: 1) intra-class distance optimization based on variational feature representation and 2) inter-class distance optimization based on feature concatenation. The intra/inter-class optimization scheme was then introduced based on the reconstructed feature embeddings. Specifically, the KL divergence was utilized to optimize the intra-class distance based on the approximation during a variation Bayesian process, while the similarities between different classes were maximized to optimize the inter-class distance during a feature concatenation process. An intelligent detection algorithm was finally developed for multiclass classification with imbalanced data. Experiments were designed and conducted using the following two public datasets: 1) NSL-KDD and 2) CIC-IDS 2017. Comparing with four baseline methods, the evaluation result demonstrated the outstanding performance of our proposed model especially in identifying new type of attacks with extremely imbalanced data for intrusion detection in distributed IoT.

In future studies, we will explore more deep learning techniques to refine our model. More evaluation experiments will be conducted to improve our algorithms with better performance for intrusion detection in more complex distributed IoT environments.

## References

[1] J. Delsing, "Local cloud Internet of Things automation: Technology and business model features of distributed Internet of Things automation solutions," *IEEE Ind. Electron. Mag.*, vol. 11, no. 4, pp. 8–21, Dec. 2017.

[2] Y. Wu, Y. Ma, H.-N. Dai, and H. Wang, "Deep learning for privacy preservation in autonomous moving platforms enhanced 5G heterogeneous networks," *Comput. Netw.*, vol. 185, 2021, Art. no. 107743.

[3] N. Dragoni *et al.*, "Microservices: Yesterday, today, and tomorrow," in *Present and Ulterior Software Engineering*. Berlin, Germany: Springer, 2017, pp. 195–216.

[4] Z. Cai and X. Zheng, "A private and efficient mechanism for data uploading in smart cyber-physical systems," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 2, pp. 766–775, Apr.–Jun. 2018.

[5] Y. Zuo, Y. Wu, G. Min, C. Huang, and K. Pei, "An intelligent anomaly detection scheme for micro-services architectures with temporal and spatial data analysis," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 2, pp. 548–561, Jun. 2020.

[6] J. Ren, Y. Pan, A. Goscinski, and R. A. Beyah, "Edge computing for the Internet of Things," *IEEE Netw.*, vol. 32, no. 1, pp. 6–7, Jan./Feb. 2018.

[7] Z. Cai and Z. He, "Trading private range counting over big IoT data," in *Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst.*, 2019, pp. 144–153.

[8] X. Liu *et al.*, "Attention-based bidirectional GRU networks for efficient HTTPS traffic classification," *Inf. Sci.*, vol. 541, pp. 297–315, 2020.

[9] Y. Wu, H.-N. Dai, H. Wang, and K.-K. R. Choo, "Blockchain-based privacy preservation for 5G-enabled drone communications," *IEEE Netw.*, vol. 35, no. 1, pp. 50–56, Jan./Feb. 2021.

[10] Y. Wang, Q. Yao, J. T. Kwok, and L. M. Ni, "Generalizing from a few examples: A survey on few-shot learning," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 1–34, 2020.

[11] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Prog. Artif. Intell.*, vol. 5, no. 4, pp. 221–232, 2016.

[12] Y. Bengio *et al.*, "Deep learners benefit more from out-of-distribution examples," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 164–172.

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. 3rd Int. Conf. Learn. Representations*, 2015, pp. 1–11.

[14] Y. Wu, H.-N. Dai, and H. Wang, "Convergence of blockchain and edge computing for secure and scalable IIoT critical infrastructures in industry 4.0," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 2300–2317, Feb. 2021.

[15] Z. Cai, Z. Xiong, H. Xu, P. Wang, W. Li, and Y. Pan, "Generative adversarial networks: A survey towards private and secure applications," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–37, Jul. 2021.

[16] A. Santoro, S. Bartunov, M. Botvinick, D. Wierstra, and T. Lillicrap, "Meta-learning with memory-augmented neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 1842–1850.

[17] L. Yang, Y. Li, J. Wang, and N. N. Xiong, "FSLM: An intelligent few-shot learning model based on Siamese networks for IoT technology," *IEEE Internet Things J.*, vol. 8, no. 12, pp. 9717–9729, Jun. 2021.

[18] X. Zhou, W. Liang, S. Shimizu, J. Ma, and Q. Jin, “Siamese neural network based few-shot learning for anomaly detection in industrial cyber-physical systems,” *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5790–5798, Aug. 2021.

[19] P. Bedi, N. Gupta, and V. Jindal, “Siam-IDS: Handling class imbalance problem in intrusion detection systems using Siamese neural network,” *Procedia Comput. Sci.*, vol. 171, pp. 780–789, 2020.

[20] S. Basodi, C. Ji, H. Zhang, and Y. Pan, “Gradient amplification: An efficient way to train deep neural networks,” *Big Data Mining Analytics*, vol. 3, no. 3, pp. 196–207, Sep. 2020.

[21] X. Zhou, Y. Hu, W. Liang, J. Ma, and Q. Jin, “Variational LSTM enhanced anomaly detection for industrial big data,” *IEEE Trans. Ind. Informat.*, vol. 17, no. 5, pp. 3469–3477, May 2021.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, 2002.

[23] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, 2008, pp. 1322–1328.

[24] B. Santoso, H. Wijayanto, K. Notodiputro, and B. Sartono, “Synthetic over sampling methods for handling class imbalanced problems: A review,” *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 58, no. 1, 2017, Art. no. 012031.

[25] T. Zhou, W. Liu, C. Zhou, and L. Chen, “GAN-based semi-supervised for imbalanced data classification,” in *Proc. 4th Int. Conf. Inf. Manage.*, 2018, pp. 17–21.

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2980–2988.

[27] P. Wu and H. Guo, “LuNET: A deep neural network for network intrusion detection,” in *Proc. IEEE Symp. Ser. Comput. Intell.*, 2019, pp. 617–624.

[28] A. Puzanov, S. Zhang, and K. Cohen, “Deep reinforcement one-shot learning for artificially intelligent classification in expert aided systems,” *Eng. Appl. Artif. Intell.*, vol. 91, 2020, Art. no. 103589.

[29] D. Lee, S. Yu, and H. Yu, “Multi-class data description for out-of-distribution detection,” in *Proc. 26th ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2020, pp. 1362–1370.

[30] A. S. Ross and F. Doshi-Velez, “Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients,” in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018, pp. 1–10.

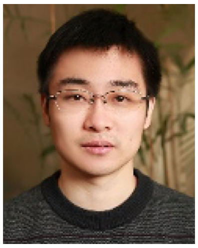
**Wei Liang** (Member, IEEE) received the M.S. and Ph.D. degrees in computer science from Central South University, Changsha, China, in 2005 and 2016, respectively.

From 2005 to 2012, he was with Microsoft, China, for soft engineering. From 2014 to 2015, he was an Exchange Researcher with the Department of Human Informatics and Cognitive Sciences, Faculty of Human Sciences, Waseda University, Tokyo, Japan. He is currently with the Base of International Science and Technology Innovation and Cooperation on Big Data Technology and Management, Hunan University of Technology and Business, Changsha. He has authored or coauthored more than 20 papers at various conferences and journals, including *Future Generation Computer Systems*, *Journal of Computer and System Sciences*, and *Personal and Ubiquitous Computing*. His research interests include information retrieval, data mining, and artificial intelligence.

Dr. Liang is a Member of the IEEE Computer Society.

**Yiyong Hu** (Member, IEEE) received the bachelor’s degree in measurement and control technology and instrumentation from the Harbin University of Science and Technology, Harbin, China, in 2017. He is currently working toward the M.S. degree in computer science with the Hunan University of Technology and Business, Changsha, China.

He was a Software Development Engineer, from 2017 to 2018. His research interests include cybersecurity, artificial intelligence, and natural language processing.



**Xiaokang Zhou** (Member, IEEE) received the Ph.D. degree in human sciences from Waseda University, Tokyo, Japan, in 2014.

From 2012 to 2015, he was a Research Associate with the Faculty of Human Sciences, Waseda University. Since 2017, he has been a Visiting Researcher with the RIKEN Center for Advanced Intelligence Project (AIP), RIKEN, Tokyo. He is currently an Associate Professor with the Faculty of Data Science, Shiga University, Hikone, Japan. He has been engaged in interdisciplinary research works in the fields of computer science and engineering, information systems, and social and human informatics. His research interests include ubiquitous computing, big data, machine learning, behavior and cognitive informatics, cyber-physical-social systems, cyber intelligence and security.

Dr. Zhou is a Member of the IEEE Computer Society; ACM, USA; IPSJ; JSAI, Japan; and CCF, China.

**Yi Pan** (Senior Member, IEEE) received the B.Eng. and M.Eng. degrees in computer engineering from Tsinghua University, Beijing, China, in 1982 and 1984, respectively, and the Ph.D. degree in computer science from the University of Pittsburgh, Pittsburgh, PA, USA, in 1991.

In 2000, he joined Georgia State University, Atlanta, GA, USA, was promoted to a Full Professor in 2004, named a Distinguished University Professor in 2013, and designated a Regents’ Professor (the highest recognition given to a faculty member by the University System of Georgia) in 2015. He is currently a Regents’ Professor and the Chair of the Department of Computer Science with Georgia State University. He was an Associate Dean and the Chair of the Department of Biology during 2013–2017. His profile has been featured as a distinguished alumnus in both *Tsinghua Alumni Newsletter* and the *University of Pittsburgh CS Alumni Newsletter*. He has authored or coauthored more than 400 papers, including more than 230 SCI journal papers and 90 IEEE transactions papers. He has also edited/authored 43 books. His work has been cited more than 11 600 times based on Google Scholar and his current H-index is 55. His research interests include parallel and cloud computing, big data, and bioinformatics.

Dr. Pan was the recipient of many awards, including one IEEE Transactions Best Paper Award, five IEEE and other international conference or journal Best Paper Awards, four IBM Faculty Awards, two JSPS Senior Invitation Fellowships, IEEE BIBE Outstanding Achievement Award, IEEE Outstanding Leadership Award, the NSF Research Opportunity Award, and AFOSR Summer Faculty Research Fellowship. He has organized numerous international conferences and delivered keynote speeches at more than 60 international conferences around the world. He was the Editor-in-Chief or an Editorial Board Member for 20 journals, including seven IEEE transactions.

**Kevin I-Kai Wang** (Member, IEEE) received the B.E. (Hons.) degree in computer systems engineering and the Ph.D. degree in electrical and electronics engineering from the Department of Electrical and Computer Engineering, The University of Auckland, Auckland, New Zealand, in 2004 and 2009 respectively.

He is currently a Senior Lecturer with the Department of Electrical and Computer Engineering, The University of Auckland. He was also a Research Engineer designing commercial home automation systems and traffic sensing systems from 2009 to 2011. His research interests include wireless sensor network-based ambient intelligence, pervasive healthcare systems, human activity recognition, behavior data analytics, and biocybernetic systems.