Computers, Materials & Continua
DOI:10.32604/cmc.2021.017665
Article

Tech Science Press

Enhanced Deep Autoencoder Based Feature Representation Learning for Intelligent Intrusion Detection System

Thavavel Vaiyapuri* **and Adel Binbusayyis**

College of Computer Engineering and Sciences, Prince Sattam Bin Abdulaziz University, AlKharj, Saudi Arabia
*Corresponding Author: Thavavel Vaiyapuri. Email: t.thangam@psau.edu.sa
Received: 02 February 2021; Accepted: 06 March 2021

Abstract: In the era of Big data, learning discriminant feature representation from network traffic is identified has as an invariably essential task for improving the detection ability of an intrusion detection system (IDS). Owing to the lack of accurately labeled network traffic data, many unsupervised feature representation learning models have been proposed with state-of-the-art performance. Yet, these models fail to consider the classification error while learning the feature representation. Intuitively, the learnt feature representation may degrade the performance of the classification task. For the first time in the field of intrusion detection, this paper proposes an unsupervised IDS model leveraging the benefits of deep autoencoder (DAE) for learning the robust feature representation and one-class support vector machine (OCSVM) for finding the more compact decision hyperplane for intrusion detection. Specially, the proposed model defines a new unified objective function to minimize the reconstruction and classification error simultaneously. This unique contribution not only enables the model to support joint learning for feature representation and classifier training but also guides to learn the robust feature representation which can improve the discrimination ability of the classifier for intrusion detection. Three set of evaluation experiments are conducted to demonstrate the potential of the proposed model. First, the ablation evaluation on benchmark dataset, NSL-KDD validates the design decision of the proposed model. Next, the performance evaluation on recent intrusion dataset, UNSW-NB15 signifies the stable performance of the proposed model. Finally, the comparative evaluation verifies the efficacy of the proposed model against recently published state-of-the-art methods.

Keywords: Cybersecurity; network intrusion detection; deep learning; autoencoder; stacked autoencoder; feature representational learning; joint learning; one-class classifier; OCSVM

1 Introduction

Following a decade of rapid advances, networking technologies has driven dramatically the global connectivity and online businesses worldwide. Apparently, this cyber dependency has opened opportunities for cybercriminals to increase the odds of cyberattacks inflicting catastrophic

damage against business continuity and eventually impacts the political economy [1]. As such, the dire need for cybersecurity solutions takes a new crucial stand in safeguarding the cyberspace from the fast-evolving threat landscape. Recently, IDS has become an indispensable part of cybersecurity as it not only detects the anomalous activities from attackers but also monitors the network for any attempts to break the cybersecurity and alerts with timely information to secure the system [2]. For these reasons, IDS has gained momentous interest across the industries and research communities.

The wide adoption of internet technologies is closely followed by massive increase of network traffic volume which contains redundant and irrelevant network data. These undesired data make the classification process increasingly complex in the existing IDS and limits the detection accuracy for intrusion [3,4]. Consequently, efficient and effective approach for learning the most robust feature representation from massive network data is of paramount importance to boost detection accuracy of IDS.

In recent years, several research works have explored the potential of feature representation learning (FRL) to handle massive amount of data and solve the classification problems in various fields with state-of-the-art performance. For example, to handle the arrival of information explosion effectively in field of multimedia technology, Wang et al. [5] presented a number of FRL approaches based on deep autoencoders and compared their efficiency for text categorization process. Similarly, Tang et al. [6] embedded unsupervised FRL to make the feature selection process robust to noises and improve the performance of data mining tasks on massive datasets. In [7], the authors applied FRL within a multi-scale framework to learn more compact discriminative descriptors for effective person reidentification in a video surveillance system. Apart from the above applications, several researchers have reported the prime role of FRL in handling the large volume of biological data to identify therapeutic peptides and serve as future benchmark in designing promising tools for disease screening [8,9]. More potentially in recent studies, many authors have claimed that the application of FRL can boost the performance of real-time analytic tasks on massive IoT data [10–13]. Taking inspiration from these literatures, the impetus of this work is to apply FRL and develop an efficient IDS which can be in pace with current trends to handle the large volume of network traffic in a big data environment and display higher detection accuracy for intrusion.

Deep learning (DL), a new version of machine learning has shown a series of breakthroughs in recent years in wide range of applications [14,15]. More essentially, the recent research trends are recognizing DL as a most promising approach for FRL deeming that the hierarchical nonlinear mappings via multiple activation layers in DL will facilitate to learn the robust feature representation from the given raw data through successive transformations [16]. But as stated in literature [17], the success of DL merely depends on the quality of the employed training set. The increasing network size indeed may complicate the labeling process and may lead to error prone training set. Owing to these reasons, unsupervised deep learning approaches has gained revival of interest in the field of intrusion detection.

Amongst different unsupervised DL approaches, the autoencoder (AE) architecture has shown immense potential for impressive feature representation and is under intensive research. For example, the authors in [18] developed an online light weight AE model utilizing random forest to select the effective features for representation learning and displayed improved accuracy for intrusion detection. correspondingly, an AE model is designed in [19] combining the advantages of data analytics and statistical techniques to extract strongly correlated features. The model gained better intrusion detection performance for modern attacks. Musafer et al. [20] proposed a mathematical

model to optimize the hyperparameters of sparse AE and enhanced the model capability for feature representation learning. Besides these efforts on improving the performance of AE for intrusion detection, few variants of AE are also put forward for intrusion detection. For instance, in 2017, Yu et al. [21] introduced dilated convolutional AE combining the strength of AE and CNN. The model proved its potential for learning robust feature representation from large volume of raw network traffic and meeting high accuracy demand of modern network IDS. Following this variant of AE, in 2018, [22] a non-symmetric DAE is proposed for unsupervised feature representation learning, to increase the detection capability of random forest classifier for intrusion detection. Yan et al. [23] designed a stacked sparse AE model based on unsupervised learning strategy to learn useful feature representation of intrusive behavior and compared its performance on three shallow learning classifiers. Al-Qatf et al. [24] presented an AE model based on self-taught learning framework with SVM for classification to gain improved detection accuracy with regard to attacks. In 2019, the authors in [25] developed a two-stage semi-supervised stacked AE model to learn useful feature representation from large volume of network traffic data. Then, used the learnt feature representation with softmax classifier to achieve increased detection rate for unseen attacks. Recently in 2020, a convolutional AE [26] is proposed for multi-channel feature representation learning and demonstrated that the class-specific features of network traffic can improve the model accuracy significantly for intrusion detection. The model potentially accelerated the intrusion detection process with good accuracy for attacks.

It is worth noting in these literatures, that the application of AE for FRL has contributed to achieve better detection accuracy yet are confronted with two main issues while considering their practical applications. First, due to system uncertainty, the abnormal network traffics are not collected in large size. Intuitively, the available intrusion datasets are inherently imbalanced with more normal traffic samples. The existing models trained under this scenario are biased towards normal traffic behavior degrading the detection accuracy for intrusion. Second, the existing models learn feature representation minimizing reconstruction loss on training sets. Instinctively, there is no guarantee that the learnt feature representation is optimal for intrusion detection task.

Taking into account the aforementioned factors, this paper proposes a novel unsupervised IDS model integrating DAE and one-class classifier within a joint framework for intrusion detection. The joint framework guides the DAE to learn optimal feature representation and enhance the discriminative ability of the classifier for intrusion detection. Further, to address the class imbalance problem, both feature representation and one-classifier learning are trained only with normal samples. In short, the major contributions of this work are highlighted below,

a. To the best of authors' knowledge, this is the first study to propose a joint optimization framework that simultaneously optimizes DAE for feature representation learning and one-class classifier for intrusion detection.
b. Different from existing works, a unified objective function is defined combining the reconstruction error and classification error to ensure that the learnt feature representation is robust to minimize the classification error and achieve higher accuracy for intrusion detection.
c. The proposed model is trained only with the given normal samples to address the class imbalance problem and overfitting that may more likely occur due to the lack of insufficient intrusion traffic samples. This ensures and improves the generalization ability of proposed model.
d. Extensive ablation experiments on benchmark intrusion datasets demonstrate the potential of the proposed model to gain improved detection rate for intrusion through robust feature

representation learning. The comparative analysis results manifest the effectiveness of the proposed model against the state-of-the-art methods.

2 Proposed Methodology

The proposed unsupervised IDS model as shown in Fig. 1 includes two essential components namely, DAE for normal traffic feature representational learning and one-class classifier for intrusion detection. The two subsections that follow elaborates the technical details of the two components, respectively. Subsequently, the objective function and training process of the proposed model is presented.

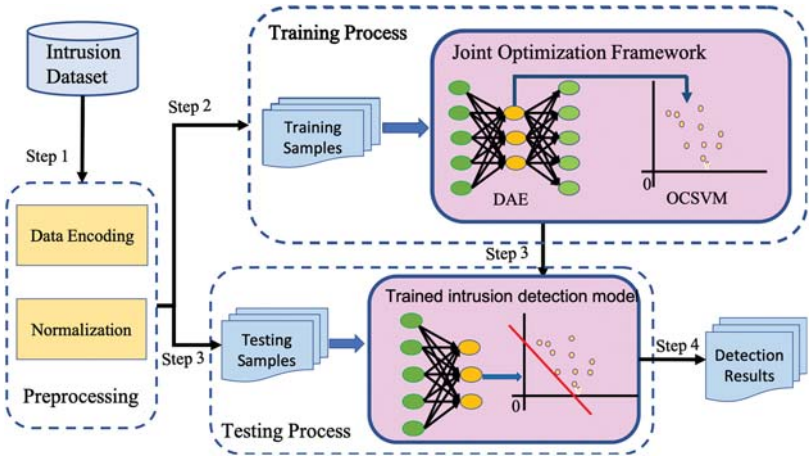

Figure 1: Illustration of proposed IDS architecture

2.1 Deep Autoencoder for Feature Representation Learning

An AE is neural network that learns the intrinsic network traffic features reconstructing the original network traffic at its output layer [27]. The general architecture of an AE consists of two key networks, encoder and decoder connected in serial. As represented by Eq. (1), the encoder network generates the feature representation by mapping the given input network traffic to hidden layer using an activation function f parameterized by W and b.

$$H = f(WX + b) \tag{1}$$

Similarly, the decoder network reconstructs the original input network traffic from the generated feature representation using the activation function g parameterized by W′ and b′ as given below

$$Z = \mathrm{g}(W'H + b') \tag{2}$$

The AE is trained jointly with given training samples to learn the parameter set $\theta = \{\mathrm{W}, \mathrm{W}', \mathrm{b}, \mathrm{b}'\}$ of two networks, encoder and decoder minimizing the reconstruction error which is determined as follows,

$$L_r(X) = \min_{\theta} \frac{1}{2N} \|X - Z\|^2 = \min_{\theta} \frac{1}{2N} \|X - \mathrm{g}(f(X))\|^2 \tag{3}$$

In general, AE demonstrates lower training efficiency and poor generalization ability while dealing diverse and massive network traffic data, due to its simple network structure with single hidden layer [28]. To address this limitation, this work constructs DAE stacking multiple AEs successively such that the output of first AE is fed as input to next AE and so on as shown in Fig. 2. Notably, this hierarchical structure benefits to drive deeper and learn more abstract high-level features that can support better feature representation learning. The output of the kth AE is computed as follows setting $H^0 = X$.

$$H^k = f(W^k H^{k-1} + b^k) \tag{4}$$

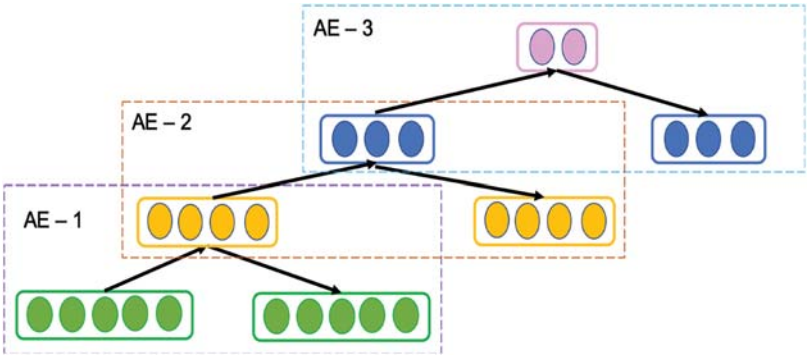

Figure 2: Structure of general AE network

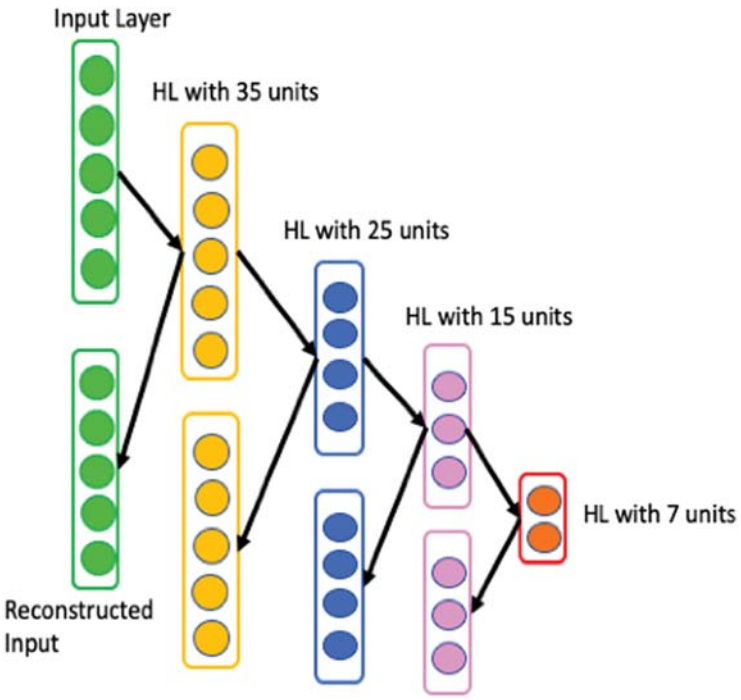

Figure 3: DAE network in the proposed IDS model consists of four AEs with hidden layer dimension 35, 25, 15, 7 respectively

The DAE network of the proposed model is formed by stacking four AEs with the hidden layer dimension of 35, 25, 15 and 7 respectively to learn the most robust feature representation hierarchically from input network traffic data in an unsupervised manner. The structure of DAE of the proposed model is shown in Fig. 3. Here, the number of hidden layer and hidden units is decided conducting practical experiments and utilizing best results without following the rule of thumb defined in previous literature. This is due to the fact that no evidence of any kind exists to confirm the validity of these rules for network generalization. Also, all hidden layers in our DAE use Rectified linear unit (ReLU) activation function. And the number of input units is set in concord to the dimension of input network traffic data. More specifically, the number of trainable parameters is reduced using tied weights where parameter sharing is enforced to obtain decoder weights transposing the encoder weights [29].

2.2 One-Class Classifier for Intrusion Detection

In most real networking environments, acquisition of network traffic with various anomalous behavior is practically impossible. Therefore, this work focuses on unsupervised one-class classification (OCC) method for intrusion detection. Essentially, OCC methods aim to build classifier model using only normal traffic behavior and detect a new incoming traffic as intrusion if its behavior deviates from normal behavior [30]. Thus, OCC methods play significant role in successfully modeling the normal traffic behavior without any a priori knowledge about its underlying distribution. Among different OCC methods, OCSVM method has attracted lot of attention in recent literatures due to its several merits in solving OCC problems [31,32], such as its kernel trick to deal with nonlinearity in input data, its ν trick to deal effectively the outliers in training set and its sparseness of solution to deal effectively with massive input data.

Inspired by these merits, this work employs OCSVM method for intrusion detection. In real scenario, the distribution of normal samples in training set are non-linearly separable. Hence, OCSVM maps the given normal samples to feature space using $\varphi(X)$ to make them linearly separable and finds a decision boundary that separates all mapped normal samples from origin with maximum margin solving the optimization problem [30] given below,

$$L_c(X) = \min_{\omega \in F, \rho, \xi} \frac{1}{2}\|\omega\|^2 - \rho + \frac{1}{\upsilon N}\sum_{i=1}^{N}\xi_i$$
$$s.t. \quad \omega.\varphi(X) - \rho + \xi_i \geq 0, \quad \xi_i \geq 0, \quad \forall i \tag{5}$$

Here, the mapping $\varphi(.)$ is usually implicit and indefinite. Therefore, the inner product of mapped data is generally specified by kernel function $K(x_i, x_j)$ in practice. The most commonly used kernel functions are linear, sigmoid, polynomial and radial basis (RBF). In this work, RBF is chosen to achieve better performance. The RBF is given as follows,

$$K(x_1, x_2) = \exp\left(-\gamma . \|x_1 - x_2\|^2\right) \tag{6}$$

Further, in Eq. (5), F denotes feature space, $\frac{\rho}{\|\omega\|}$ denotes margin size, the term ξ_i models classification $\|w\|$ error with respect to the ith sample and the regularization term $\nu \in (0, 1]$ is outlier score that controls the tradeoff between maximizing the margin from origin and minimizing the

classification error. Usually, the optimization problem in Eq. (5) is solved introducing Lagrangian multiplier α and the final decision hyperplane of OCSVM is obtained as follows,

$$f(x) = sign\left(\sum_{i=1}^{N} \alpha_i K(x_i, x) - \rho\right),$$
$$= \begin{cases} 1, & if\ x\ belongs\ to\ target\ class \\ 0, & otherwise \end{cases} \tag{7}$$

In the above equation, α_i is obtained by solving its dual form as follows,

$$\max_{\alpha} -\frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j K(x_i, x_j)$$
$$s.t. \quad \sum_{i=1}^{N} \alpha_i = 1, \quad 0 \le \alpha_i \le \frac{1}{\upsilon N}, \quad \forall\, i \tag{8}$$

Once the optimal solution for α_i is obtained, the constant ρ is computed by selecting a sample from the training set that satisfies $0 \le \alpha_j \le \frac{1}{\upsilon N}$ and that the sample is a support vector.

$$\rho = \sum_{j=1}^{N} \alpha_j \mathrm{K}(x_i, x_j) \tag{9}$$

Observing the decision function of OCSVM, it is evident that OCSVM can effectively detect malicious activities just with knowledge of normal network traffic samples with an optimal hyperplane. Notwithstanding, the performance of OCSVM is sensitive to the data-dependent parameters (γ, υ) that are difficult to tune. The subsection following will brief how these parameters are fine-tuned during training process in an unsupervised manner through a unified objective function defined in this work.

2.3 Unified Objective Function

In all existing IDS, AE and classifier are trained independently without joint optimization [19,21,24]. In that case, the learnt feature representation does not guarantee strong discriminative ability for intrusion detection task. The work here aims to combat this problem combining the reconstruction loss in Eq. (3) with classification loss in Eq. (5), as given below,

$$L = L_r(X) + L_c(H^k) \tag{10}$$

The above defined unified objective function guides the proposed model to learn robust feature representation for an improved effective intrusion detection by integrating the feature representation and classification process into joint optimization framework. In doing so, the proposed model reduces the reconstruction loss and at the same time ensures that the classification hyperplane margin is maximized for improving detection accuracy of the proposed model.

Algorithm 1: Algorithm of the proposed IDS model

Input: X-Training Set

Initialization:

1. DAE parameter $\theta = (W, b, \mathrm{W}', b')$ using Xavier algorithm
2. OCSVM parameter $\{\nu, \gamma\}$ using grid search algorithm

Procedure:

Stage-I : Greedy Layer-wise Pretraining Process

for each pretraining epoch **do**

for each mini batch **do**

1. $H \leftarrow$ Feature representation using Eq. (1)
2. $Z \leftarrow$ Reconstructed Input using Eq. (2)
3. $Lr \leftarrow$ Reconstruction Loss using Eq. (3)

end for

end for

Stage II : Training for Fine tuning Process

repeat

Feature representation Learning:

1. $H \leftarrow$ Feature representation using Eq. (1)
2. $Z \leftarrow$ Reconstructed Input using Eq. (2)
3. $Lr \leftarrow$ Reconstruction Loss using Eq. (3)

Classifier Learning:

1. Obtain optimal kernel parameter γ using grid search
2. Transform H to *kernel space* applying Eq. (6)
3. Find the hyperplane computing αi as in Eq. (8) and ρ as in Eq. (9)

Optimization:

1. Compute the gradient minimizing the objective function in Eq. (10)
2. Update model parameter θ

until $\langle$Convergence of $\theta\rangle$

2.4 Training Process

As advised in previous literature [33,34], the training of DAE comprises two stages, pretraining and fine tuning to avoid local minima and to ensure fast convergence. During pretraining, each AE in the proposed DAE component is trained individually for its parameters minimizing the reconstruction loss and then its encoding is used as input to the next AE.

After the weight and bias vector of all AEs are initialized through pretraining, first the hyperparameter ν of OCSVM is optimized using grid search algorithm on the given training set. Later, the two key components, DAE and OCSVM in the proposed model are trained jointly in an unsupervised manner to fine tune their hyperparameter optimizing the unified objective function defined in Eq. (10). During each iteration of fine-tuning process, the parameter γ of OCSVM is optimized using grid search on the feature representation learnt during that iteration. This sequence of training ensures for robust feature representation that not only demonstrates ability for reconstruction of input network traffic but also in enhancing the discriminative ability of OCSVM for intrusion detection.

Furthermore, Xavier algorithm is used to initialize the trainable parameters of DAE to keep the gradient values and activation values within a reasonable range [35]. Also, adaptive moment

estimation (Adam) method [36] which is regarded as better gradient optimization method for deep learning networks is chosen in this work to compute the gradient values of θ as it presents the benefits of both AdaGrad and RMSProp algorithms. Notably, to obtain stabilized results, the training process is terminated when the number of epochs exceed 15 and loss value of the model falls below the threshold value of 0.005. The Algorithm-1 summarizes the training procedure adopted for the proposed model.

3 Experimental Setup

This section first describes the experimental datasets. Then details the methods used for preprocessing the datasets. Subsequently, the implementation details and the metrics used for experimental evaluation are presented.

3.1 Datasets

A number of datasets are available publicly for IDS research evaluation. Nonetheless, these datasets suffer from absences of traffic diversity and lack of sufficient number of sophisticated attack styles. Therefore, in order to conduct a fair and effective evaluation of the proposed model, an old benchmark NSL-KDD dataset and a new contemporary UNSW-NB15 dataset are considered in this work. A brief description of these two intrusion datasets is given below.

A. NSL-KDD Dataset

The NSL-KDD dataset is an improved version of KDD'99 dataset, presented by Tavallaee et in 2009 resolving the redundancy in KDD '99 dataset [37]. This dataset contains an optimal ratio of 125,973 training samples to 22,543 testing samples. Thus NSL-KDD is regarded as one of the most valuable benchmark resource in the field cybersecurity research for IDS evaluation. Each sample in NSL-KDD contains 41 features and 1 class label to characterize whether the network traffic is normal or belongs to attack category. The distribution of normal traffic samples in the training and testing sets with regard to attacks are given in Tab. 1.

Table 1: Data distribution in NSL-KDD

Class	Training set	Testing set
Normal	67,343	9,710
Attack	58,630	12,833
Total	125,973	22,543

B. UNSW-NB15 Dataset

The UNSW-NB15 is a modernized dataset recently developed by ACCS with hybrid of real normal and synthesized contemporary attack behavior from network traffic flow [38]. This dataset includes 9 families of attacks namely DoS, Analysis, Generic, Fuzzers, Backdoors, Exploits, Shellcode, Reconnaissance, and Worms. The dataset consists of 175,341 training samples and 82,332 testing samples, each characterized with 42 features and a class label to discriminate the network traffic as normal or malicious activities. The distribution of samples against normal and attack class is shown in Tab. 2.

Table 2: Data distribution in UNSW-NB15

Class	Training set	Testing set
Normal	56,000	37,000
Attack	119,341	45,332
Total	175,341	82,332

3.2 Data Preprocessing

Data preprocessing is essentially crucial for providing quality input for model training and boost the detection ability of the IDS. It includes two main operations namely, data encoding and normalization.

a) ***Data Encoding:*** In this work, label encoding method is used to map all non-numeric or nominal features to numeric values. This method maps a nominal feature with C different values to an integer in the range of 0 to C − 1. For example, the NSL-KDD dataset includes three nominal features namely, protocol_type, service_type, TCP status flag with 3, 70 and 11 distinct nominal values respectively. After label encoding, the feature protocol_type with three values is mapped as follows, tcp:0, udp:1 and icmp:2.

b) ***Normalization:*** Generally, the machine learning algorithms are biased by input features with large numeric value. To combat this effect, min-max normalization is applied to adjust the value range of all input features within the range [0, 1].

3.3 Implementation Details

All the experiments are conducted on a personal computer with the specifications as follows, Intel Core i7-8565H@1.8 GHz, 128 GB RAM and Windows 10 operating system. The proposed model is implemented in Jupyter development environment using Python 3 as programming language. More specifically, the python libraries, Keras and Tensorflow are used to implement various deep learning tasks [39]. Also, python Scikit-learn library is used to implement various evaluation measures and data preprocessing tasks.

3.4 Evaluation Metrics

The effectiveness of the proposed IDS model is measured by analyzing four evaluation metrics that are most commonly used in the field of intrusion detection. The relevant definition of these four metrics are as follows,

a) ***Accuracy (ACC):*** measures the proportion of network traffic flows that are correctly classified and is computed as follows,

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

b) ***Detection rate (DR):*** Also called Recall or Sensitivity, measures the proportion of intrusion traffic flow that are correctly classified as given below,

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

c) ***F1-measure (F1):*** Also termed as F1-Score, is considered as more effective measure than accuracy to evaluate the performance of intrusion detection model especially for imbalanced datasets. It is an hormonic average of detection rate and precision as follows

$$F1 = \frac{2 \times (DR \times Precision)}{DR \times Precision} \tag{13}$$

d) *False alarm rate (FAR):* Also termed as false positive rate, measures the proportion of normal network traffic flows that are incorrectly classified. It is computed as follows,

$$FAR = \frac{FP}{FP + TN} \tag{14}$$

4 Experimental Results and Discussion

The potentiality of the proposed model is demonstrated designing three set of evaluation experiments. Specially, these experiments are designed with the following objectives

a) Validate the design decision of the proposed model on benchmark dataset NSL-KDD
b) Verify the stable performance of the proposed model on recent intrusion dataset
c) Compare the potential of the proposed model against recently published state-of-the-art methods.

4.1 Ablation Evaluation

This study involves two sets of analyses, first the design decision of the proposed model is validated, next the structural configuration of the DAE network is analyzed with regard to intrusion detection performance. Both these analyze are conducted on the standard benchmark intrusion dataset, NSL-KDD in terms of ACC, F1, DR and FAR. The subsections below describe these two analyses in detail.

4.1.1 Ablation Analysis I

The aim of this analysis is to investigate the significance and contribution of different components of the proposed model to the overall detection performance. For this purpose, the following three variants of proposed model are developed to conduct the ablation analysis,

(a) **OC:** This variant is developed removing DAE components to demonstrate the significance of feature representation learning in the proposed model
(b) **DAE + Softmax:** This variant is developed replacing the OCSVM component with Softmax layer as shown in Fig. 4. to demonstrate the significance of one-class unsupervised classification in the proposed model
(c) **DAE+OC:** This version indeed is developed to demonstrate the significance of the training jointly DAE and OCSVM through the defined unified objective function. To this purpose, the DAE is first pretrained and fine-tuned to learn essential feature representation. Then, the learnt feature representation is used to train OCSVM for intrusion detection.

In order to conduct a reasonable comparison, the above variants are developed under the same environmental setup using the same parameter as the proposed model and the results are reported in Tab. 3. The results clearly reveal the independent effects and relevance of all components of the proposed model for the obtained performance improvement with regard to intrusion detection. In particular, the results of the variant OC with high FAR value on training

set and low DR value indicates that the DAE plays a very crucial role in learning the most robust feature representation from network traffic and improves the discrimination ability of the proposed model for intrusion detection. This in accordance to the claim of recent literature [24] and provides a new insight for improving the intrusion detection accuracy.

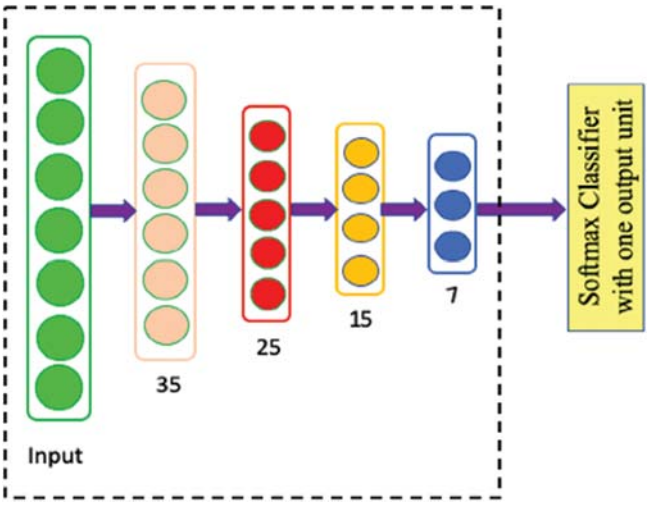

Figure 4: Structure of the DAE + softmax variant

Table 3: Ablation analysis results on NSL-KDD dataset for different variants of proposed model

Variants	Training set				Testing set			
	DR	FAR	ACC	F1	DAR	FAR	ACC	F1
OC	95.80	6.28	95.75	95.45	87.2	12.2	87.11	88.51
DAE + Softmax	94.69	6.25	94.85	94.48	81.42	14.75	85.65	86.59
DAE + OC	97.88	5.00	95.97	95.52	93.20	12.89	87.99	89.83
Proposed model	96.55	5.02	95.71	95.44	94.53	12.03	91.25	92.93

Similarly, the significant drop in the evaluation results of the variant DAE + Softmax suggests that the kernel trick of OCSVM classifier is more effective in contributing the compact representation of normal samples and improves the detection accuracy significantly without any knowledge about the intrusion behavior.

Finally, the results of the variant DAE + OCSVM reveals the potential of the proposed model over the three variants. The improved performance of the proposed model confirms the significance of the unified objective function for joint training of DAE and OCSVM to learn the robust feature representation that can enhance the discriminative ability of classifier for intrusion detection. Importantly, the outcome of this analysis suggests that the proposed model will serve as a new promising way for improving the intrusion detection accuracy in future studies.

4.1.2 Ablation Analysis II

This ablation analysis aims to observe the influence of number of hidden layers (number of stacking AEs) on intrusion detection performance exploring different structural configuration for DAE. As claimed in previous literature [40], it is quite intuitive that adding more hidden layers in DAE can improve the model performance. In tandem, a series of ablation experiments were conducted to examine the increase in number of hidden layers on intrusion detection performance. For this purpose, the number of hidden layers was varied from two to five and the corresponding results are summarized in Tab. 4. It can be observed from the results that the proposed model improved its performance with increase in number of hidden layers. But surprisingly, the performance with five hidden layers was not significantly better compared to that displayed by four hidden layers. In particular, it induced lower DR value indicating that the increase in number of hidden layers above four leads to overfitting with the trained dataset.

Table 4: Ablation analysis results on NSL-KDD dataset for different number of hidden layers in DAE network of proposed model

Number of filters	Training set				Testing set			
	DR	FAR	ACC	F1	DAR	FAR	ACC	F1
{35, 28, 21, 14, 7}	97.91	4.23	96.82	96.57	93.22	10.68	91.67	91.93
{35, 25, 15, 7}	98.29	4.41	96.62	95.75	97.11	2.43	91.58	92.87
{30, 15, 7}	97.53	5.01	95.65	95.39	91.76	11.98	90.29	90.53
{20, 7}	96.55	5.02	95.71	95.44	87.11	13.74	85.60	87.06

Added to, Fig. 5 illustrates the ability of DAE to learn the robust feature representation that can reconstruct the original input with small variation, when the number of hidden layer is 4 with dimension of 35, 25, 15 and 7 respectively. A similar observation is also reported in previous literature that increasing the number of hidden layers deeper may weaken the classifier performance [41]. Taking into account this observation, the number of hidden layers in the subsequent experiments is set to 4 to achieve remarkable detection performance.

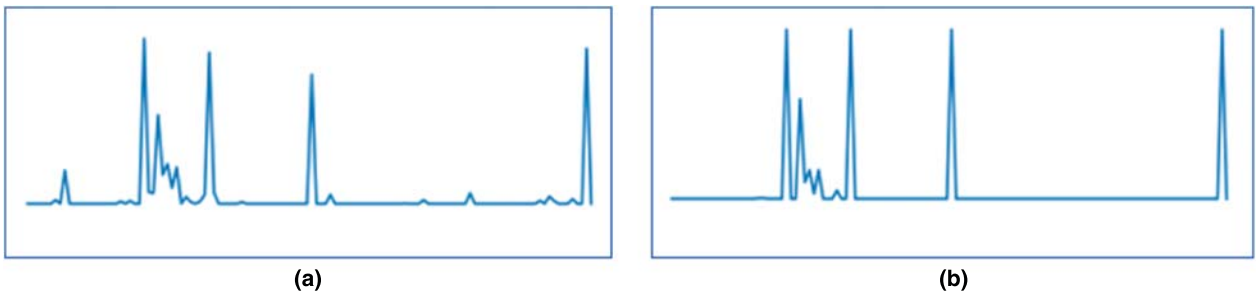

Figure 5: Illustration of reconstructed network traffic sample (b) by the proposed DAE network from the original 1D network traffic (a)

4.2 Performance Evaluation

In literature, it is stated that the change of datasets considerably affects and varies the performance of detection process [42]. Accordingly, to investigate the stable performance of the proposed model on different datasets, this experiment is conducted choosing a most recent benchmark dataset, UNSW-NB15 that includes many new modern attack styles.

The confusion matrix delivered by the proposed model on UNSW-NB15 training and testing datasets are shown in Fig. 6. The evaluation metrics computed using these confusion matrices are presented in. Fig. 7. These figures demonstrate that the proposed model is very effective in achieving a DR of 97.16, FAR of 6.6, ACC of 95.93 and F1 of 97.27 on training dataset. Comparably a DR of 96.63, FAR of 2.61, ACC of 96.97 and F1 of 97.33 on testing dataset clearly reveals the efficacy of the proposed model to generalize even on a complex dataset such as UNSW-NB15 and at the same time confirms that the proposed model is very competitive for modern attack detection.

Detection Results →	Normal	Attack
Normal	36034	966
Attack	1524	43808

(a)

Detection Results →	Normal	Attack
Normal	52259	3741
Attack	3386	115955

(b)

Figure 6: Confusion matrix of proposed model on UNSW-NB15. (a) Training set (b) testing set

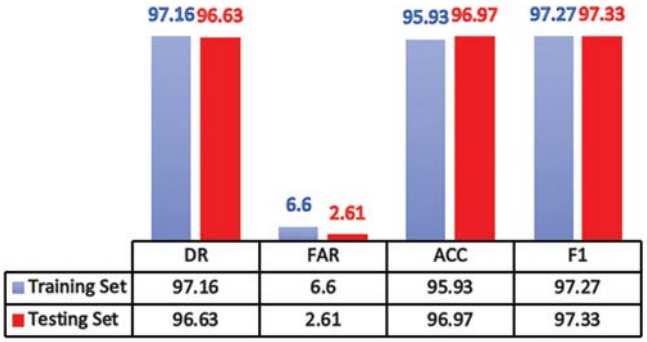

Figure 7: Performance analysis of proposed IDS model on UNSW-NB15

It is can be noted that, similar to the results on NSL-KDD, the performance improvement of the proposed model on UNSW-NB15 dataset, also remains at a promising level. This consistent performance of the proposed model is evidently attributed to the joint optimization of feature representation and classification learning for intrusion detection task.

4.3 Comparative Analysis with Related Works

The effectiveness of the proposed model is further highlighted comparing with recent and relevant IDS models based on deep learning approaches. Since its impractical to compare with all latest approaches, only those approaches that have used both NSL-KDD datasets are considered to have a rational comparison. Also, the results provided in their published papers are used to

maintain fair comparison and the results of this comparison are presented in Tab. 5. Here, for clarity purpose, the highest score is highlighted in bold for each metrics.

Table 5: Comparative evaluation of proposed model against recent IDS models

Recent IDS models	NSL-KDD testing set			
	DAR	FAR	ACC	F1
Non symmetric DAE [22]	85.42	14.58	85.42	87.37
SSAE + SVM [24]	76.56	NA	84.96	85.28
Statistical analysis + AE [19]	80.37	NA	87	81.98
ICVAE [43]	77.43	**2.74**	85.97	86.27
One-class ContAE [44]	89.23	13.66	87.98	88.41
SAVAER [45]	84.86	4.70	89.36	90.08
Proposed DAE + OCSVM	**94.53**	12.03	**91.25**	**92.93**

Now observing the results, it can be realized that the proposed model outperforms all the recent IDS approaches for all metrics except for the model introduced in [44] with improved conditional variational autoencoder (ICVAE) displays the very low FAR of 2.74. Though, ICVAE model shows lower probability for FAR than the proposed model, its performance in terms of DR, ACC and F1 metrics are very worst. This indicates that the proposed model is competitively effective in displaying better performance than all other recent approaches. The reason is possibly might be due to the introduced joint optimization framework that enables DAE to generate feature representation with potential ability not only for reconstruction but also for enhancing the classifier discriminative ability for intrusion detection.

In summary, it can be concluded that the superior performance of the proposed model demonstrates that it has great potential to be a used as promising tool for intrusion detection.

5 Conclusion

This paper has proposed an unsupervised learning model for building an effective IDS. The proposed model resolves the labelled data scarcity challenge associated with supervised learning by integrating DAE and OCSVM within a joint optimization framework. Specifically, it has defined a unified objective function for joint learning of feature representation and classification task. This mechanism has guided the DAE to learn the most robust feature representation and at the same time has ensured to improve the discriminative ability of OCSVM for any unknown attacks. The outcome of ablation experiments has not only validated the design decision of the proposed model but has also signified the crucial contribution of each component in the proposed model in gaining improved detection rate for intrusions. Also, extensive comparative evaluation has manifested the efficacy of the proposed model against recently published state-of-the-art baselines. The proposed model will serve as a new sight for the research communities in the field intrusion detection to explore on joint unsupervised learning and achieve excellent intrusion detection rate for new modern attacks.

Funding Statement: This work was supported by the Research Deanship of Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia (Grant No. 2020/01/17215). Also, the author thanks

Deanship of college of computer engineering and sciences for technical support provided to complete the project successfully.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] X. Liu, J. Yu, W. Lv, D. Yu, Y. Wang *et al.,* "Network security situation: From awareness to awareness-control," *Journal of Network and Computer Applications*, vol. 139, no. 9, pp. 15–30, 2019.

[2] A. Binbusayyis and T. Vaiyapuri, "Identifying and benchmarking key features for cyber intrusion detection: An ensemble approach," *IEEE Access*, vol. 7, pp. 106495–106513, 2019.

[3] A. Binbusayyis and T. Vaiyapuri, "Comprehensive analysis and recommendation of feature evaluation measures for intrusion detection," *Heliyon*, vol. 6, no. 7, pp. e04262, 2020.

[4] K. Selvakumar, M. Karuppiah, L. Sai Ramesh, S. H. Islam, M. Hassan *et al.,* "Intelligent temporal classification and fuzzy rough set-based feature selection algorithm for intrusion detection system in WSNs," *Information Sciences*, vol. 497, no. 11, pp. 77–90, 2019.

[5] S. Wang, J. Cai, Q. Lin and W. Guo, "An overview of unsupervised deep feature representation for text categorization," *IEEE Transactions on Computational Social Systems*, vol. 6, no. 3, pp. 504–517, 2019.

[6] C. Tang, M. Bian, X. Liu, M. Li, H. Zhou *et al.,* "Unsupervised feature selection via latent representation learning and manifold regularization," *Neural Networks*, vol. 117, no. 9, pp. 163–178, 2019.

[7] Y. Wu, K. Zhang, D. Wu, C. Wang, C. A. Yuan *et al.,* "Person re-identification by multi-scale feature representation learning with random batch feature mask," *IEEE Transactions on Cognitive and Developmental Systems*, Early Access, pp. 1–10, 2020.

[8] Y. P. Zhang and Q. Zou, "PPTPP: A novel therapeutic peptide prediction method using physicochemical property encoding and adaptive feature representation learning," *Bioinformatics*, vol. 36, no. 13, pp. 3982–3987, 2020.

[9] X. Qiang, C. Zhou, X. Ye, P. Du, R. Su *et al.,* "CPPred-FL: A sequence-based predictor for large-scale identification of cell-penetrating peptides by feature representation learning," *Briefings in Bioinformatics*, vol. 21, pp. 11–23, 2020.

[10] Z. Huang, X. Xu, J. Ni, H. Zhu and C. Wang, "Multimodal representation learning for recommendation in internet of things," *IEEE Internet of Things Journal*, vol. 6, no. 6, pp. 10675–10685, 2019.

[11] O. Aydogdu and M. Ekinci, "A new approach for data stream classification: Unsupervised feature representational online sequential extreme learning machine," *Multimedia Tools and Applications*, vol. 79, no. 37–38, pp. 1–23, 2020.

[12] N. Wang, W. Zhou, Y. Song, C. Ma, W. Liu *et al.,* "Unsupervised deep representation learning for real-time tracking," arXiv preprint arXiv: 2007.11984, 2020. [Online]. Available: https://arxiv.org/abs/2007.11984.

[13] Z. Wu, Y. Guo, W. Lin, S. Yu and Y. Ji, "A weighted deep representation learning model for imbalanced fault diagnosis in cyber-physical systems," *Sensors*, vol. 18, pp. 1096, 2018.

[14] A. Aldweesh, A. Derhab and A. Z. Emam, "Deep learning approaches for anomaly-based intrusion detection systems: A survey, taxonomy, and open issues," *Knowledge-Based Systems*, vol. 189, no. 5, pp. 105124, 2020.

[15] M. M. Hassan, A. Gumaei, A. Alsanad, M. Alrubaian and G. Fortino, "A hybrid deep learning model for efficient intrusion detection in big data environment," *Information Sciences*, vol. 513, no. 3, pp. 386–396, 2020.

[16] G. Zhong, L. N. Wang, X. Ling and J. Dong, "An overview on data representation learning: From traditional feature learning to recent deep learning," *Journal of Finance and Data Science*, vol. 2, no. 4, pp. 265–278, 2016.

[17] T. A. Tang, L. Mhamdi, D. McLernon, S. A. Zaidi, M. Ghogho *et al.,* "DeepIDS: Deep learning approach for intrusion detection in software defined networking," *Electronics*, vol. 9, pp. 1533, 2020.

[18] X. Li, W. Chen, Q. Zhang and L. Wu, "Building autoencoder intrusion detection system based on random forest feature selection," *Computers & Security*, vol. 95, no. 1, pp. 101851, 2020.
[19] C. Ieracitano, A. Adeel, F. C. Morabito and A. Hussain, "A novel statistical analysis and autoencoder driven intelligent intrusion detection approach," *Neurocomputing*, vol. 387, no. 1, pp. 51–62, 2020.
[20] H. Musafer, A. Abuzneid, M. Faezipour and A. Mahmood, "An enhanced design of sparse autoencoder for latent features extraction based on trigonometric simplexes for network intrusion detection systems," *Electronics*, vol. 9, no. 2, pp. 259, 2020.
[21] Y. Yu, J. Long and Z. Cai, "Network intrusion detection through stacking dilated convolutional autoencoders," *Security and Communication Networks*, vol. 2017, pp. 1–10, 2017.
[22] N. Shone, T. N. Ngoc, V. D. Phai and Q. Shi, "A deep learning approach to network intrusion detection," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 2, no. 1, pp. 41–50, 2018.
[23] B. Yan and G. Han, "Effective feature extraction via stacked sparse autoencoder to improve intrusion detection system," *IEEE Access*, vol. 6, pp. 41238–41248, 2018.
[24] M. Al-Qatf, Y. Lasheng, M. Al-Habib and K. Al-Sabahi, "Deep learning approach combining sparse autoencoder with SVM for network intrusion detection," *IEEE Access*, vol. 6, pp. 52843–52856, 2018.
[25] F. A. Khan, A. Gumaei, A. Derhab and A. Hussain, "A novel two-stage deep learning model for efficient network intrusion detection," *IEEE Access*, vol. 7, pp. 30373–30385, 2019.
[26] G. Andresini, A. Appice, N. Di Mauro, C. Loglisci and D. Malerba, "Multi-channel deep feature learning for intrusion detection," *IEEE Access*, vol. 8, pp. 53346–53359, 2020.
[27] D. E. Rumelhart, G. E. Hinton and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.
[28] H. Shao, H. Jiang, H. Zhao and F. Wang, "A novel deep autoencoder feature learning method for rotating machinery fault diagnosis," *Mechanical Systems and Signal Processing*, vol. 95, no. 8, pp. 187–204, 2017.
[29] P. Li and P. M. Nguyen, "On random deep weight-tied autoencoders: Exact asymptotic analysis, phase transitions, and implications to training," in *Proc. ICLR*, New Orleans, U.S. of Louisiana, 2018.
[30] S. S. Khan and M. G. Madden, "One-class classification: Taxonomy of study and review of techniques," *Knowledge Engineering Review*, vol. 29, no. 3, pp. 345–374, 2014.
[31] S. M. Erfani, S. Rajasegarar, S. Karunasekera and C. Leckie, "High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning," *Pattern Recognition*, vol. 58, no. 7, pp. 121–134, 2016.
[32] J. Parras and S. Zazo, "Using one class SVM to counter intelligent attacks against an SPRT defense mechanism," *Ad Hoc Networks*, vol. 94, no. 694–706, pp. 101946, 2019.
[33] H. Song, Z. Jiang, A. Men and B. Yang, "A hybrid semi-supervised anomaly detection model for high-dimensional data," *Computational Intelligence and Neuroscience*, vol. 2017, no. 1, pp. 1–9, 2017.
[34] W. Wang, M. Zhao and J. Wang, "Effective android malware detection with a hybrid model based on deep autoencoder and convolutional neural network," *Journal of Ambient Intelligence and Humanized Computing*, vol. 10, no. 8, pp. 3035–3043, 2019.
[35] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Proc. JMLR ICAIS*, vol. 9, pp. 249–256, 2010.
[36] K. Da, "A method for stochastic optimization," arXiv preprint, arXiv: 1412.6980, 2014. [Online]. Available: https://arxiv.org/abs/1412.6980.
[37] M. Tavallaee, E. Bagheri, W. Lu and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set. 2009," in *Proc. IEEE CISDA*, Ottawa, Canada, pp. 1–6, 2009.
[38] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in *Proc. IEEE MilCIS*, Canberra, Australia, pp. 1–6, 2015.
[39] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, California: O'Reilly Media, Inc., 2019.

[40] Q. Xu, C. Zhang, L. Zhang and Y. Song, "The learning effect of different hidden layers stacked autoencoder," *Proc. IEEE IHMSC*, vol. 2, pp. 148–151, 2016.

[41] P. Raut and A. Dani, "Correlation between number of hidden layers and accuracy of artificial neural network," in *Proc. Springer ICACTA*, Mumbai, India, pp. 513–521, 2020.

[42] A. Fu, C. Dong and L. Wang, "An experimental study on stability and generalization of extreme learning machines," *International Journal of Machine Learning and Cybernetics*, vol. 6, no. 1, pp. 129–135, 2015.

[43] Y. Yang, K. Zheng, C. Wu and Y. Yang, "Improving the classification effectiveness of intrusion detection by using improved conditional variational autoencoder and deep neural network," *Sensors*, vol. 19, pp. 2528, 2019.

[44] T. Vaiyapuri and A. Binbusayyis, "Application of deep autoencoder as an one-class classifier for unsupervised network intrusion detection: A comparative evaluation," *PeerJ Computer Science*, vol. 6, no. 1, pp. e327, 2020.

[45] Y. Yang, K. Zheng, B. Wu, Y. Yang and X. Wang, "Network intrusion detection based on supervised adversarial variational auto-encoder with regularization," *IEEE Access*, vol. 8, pp. 42169–42184, 2020.