PSYCHOLOGY AND EDUCATION (2019) 56(1): 52-72
ISSN: 1553-3969

Designing an Efficient Intrusion Detection System Based on Feature Selection and Ensemble Classifier

L K Suresh Kumar
University College of Engineering, Osmania University, India

Abstract

Intrusion detection system (IDS) is one of extensively used techniques in a network topology to safe guard the integrity and availability of sensitive assets in the protected systems. Although many supervised and unsupervised learning approaches from the field of machine learning have been used to increase the efficacy of IDSs, it is still a problem for existing intrusion detection algorithms to achieve good performance. First, lots of redundant and irrelevant data in high-dimensional datasets interfere with the classification process of an IDS. Second, an individual classifier may not perform well in the detection of each type of attacks. Third, many models are built for stale datasets, making them less adaptable for novel attacks. Thus, we propose a new intrusion detection framework in this paper, and this framework is based on the feature selection and ensemble learning techniques. In the first step, a heuristic algorithm called PSO-SVM is proposed for dimensionality reduction, which selects the optimal subset based on the correlation between features. Then, we introduce an ensemble approach that combines CART, C4.5. Finally, voting technique is used to combine the probability distributions of the base learners for attack recognition.

Keywords

Ensemble Classifier ,Feature Selection ,Intrusion Detection System ,Machine Learning ,Particle Swarm optimisation

1.Introduction:

In recent years ,the internet is applied in every aspect of society such as teaching ,entertainment ,communication ,IoT ,Digital Banking etc . As technology is emerging ,the data has to be more secured and integrated .With regard to this ,cyber security has become more prone to attacks . Breaching of data threaten confidentiality ,integrity of the data .However ,many security applications exist such as firewall ,data encryption ,authentication etc but these applications could not prevent the data from being breached . To overcome this problem IDS (Intrusion Detection System) is introduced[1]. IDS is a system to detect malicious activities.

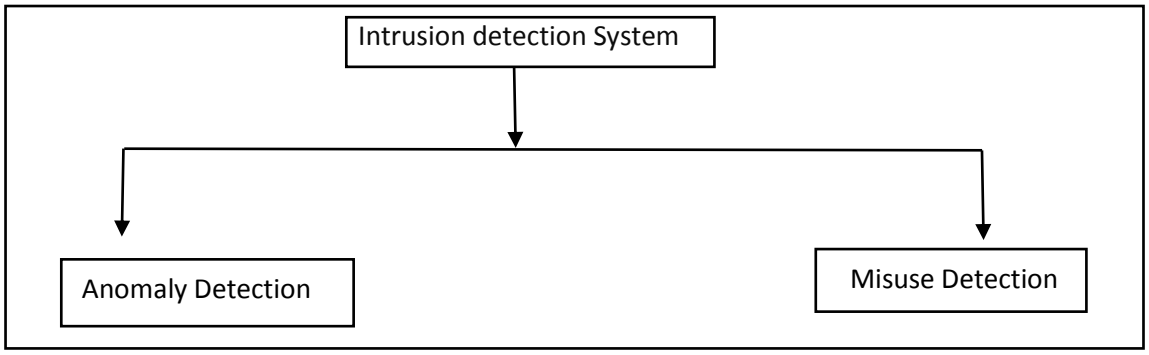

Fig. 1 IDS classification

IDS is used to detect the attacks . The fig.1 interprets that the IDS is classified into two types based on detection mechanism namely Anomaly and Misuse . Anomaly detection system is fabricated in such way to detect suspicious data from the remaining data .This detects any type of attack but does not define the attack . The misuse detection system is designed to detect only known attacks .It is also called as signature detection .This kind of system are well built for only known attacks . It can not identify undefined attacks or deviations of known attacks [1].

As the attackers become more advanced ,new attacks and vulnerabilities emerged drastically . In order to detect such attacks an advanced model should be introduced . Keeping this in mind ,many researches proposed IDS using ML and DL[3,4,5] which can be applied for both anomaly and misuse detection . The IDS not only differentiates between benign and suspicious attacks but also figure outs the specific class of an attack occurring in the protected system[1]. The detection and response mechanism of an IDS is shown in fig.2 .

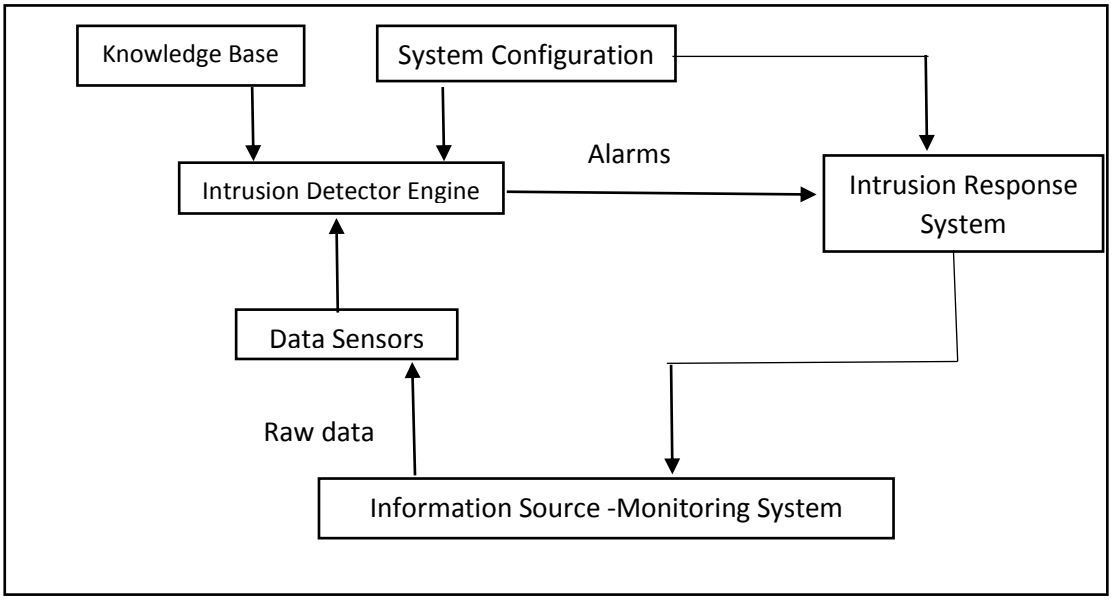

Fig. 2 Intrusion detection System and response System[36]

The aim of building an efficient IDS is it should spot attacks with high Attack Detection Rate(ADR) and low False Alarm Rate(FAR). Initially ,there was problem of applying ML in IDSs , that means a single classifier may not be strong enough to build an efficient IDS[1].There by ,the idea of ensemble classifier came into picture[6,7] .As ensemble classifiers make better segregation of data about the object submitted at the input [2]. An ensemble would average the output of multiple classifiers and therefore become a better option[1].

In this paper we propose a Intrusion Detection System to detect various types of attacks with high accuracy and efficiency .First , as a regular means of dimension reduction Feature Selection is done .Second ,the imbalance between normal and malicious traffic has a

undesirable impact on accuracy and efficiency .To deal with this ,our solution uses ensemble classifier to reduce the bias among different training data .

Due to rise in technology ,breaching of data as become quite simple for attackers . As a result of this the confidentiality and integrity of the data is lost . To detect and prevent these attacks from happening and saving important private information from being exploited by the intruders ,an efficient IDS that suits modern requirements and with high Attack Detection Rate (ADR) and less False Alarm Rate (FAR) is the main motivation behind this work.

• The proposed novel methodology that combines the benefits of feature selection and ensemble classifier with the aim of providing efficient and accurate intrusion detection.

• In the context of feature selection, we provide PSO based approach, which finds the optimal values in a specified class and beneficial for optimizing the efficiency of the training and testing phase.

• To increase the multi-class classification performance on unbalanced datasets, we introduce an ensemble approach by combining decisions from multiple classifiers (ID3 and CART) into one by utilizing a vote classifier based on the average of probabilities (AOP) combination rule.

• The proposal is compared with three datasets, namely: NSL-KDD, AWID, and CIC-IDS2017. Experimental results show that the proposed solution surpasses equivalent methods in terms of Accuracy (Acc), FMeasure, and ADR classification metrics, while keeping FAR at acceptable levels

The rest of the paper is organised as follows

Chapter 2. describes the related work that done on IDS using different approaches .Researches proposed models on feature selection , ensemble classifiers and hybrid approaches too and their respective observations and results are discussed.

Chapter 3. deals with the proposed methodology namely feature selection and ensemble classifier and the datasets that are used to evaluate the proposed model .

Chapter 4 .states the results obtained by the model on different datasets and performance comparison on different approaches .

Chapter 5.concludes the performance of our approach and discusses the future work

2 . Related work

As a significant tool in computer based systems for ensuring cyber security, IDS constantly attracts the research community's attention. Although plenty of solutions have been proposed to improve the performance of IDS, in the context of this section, we only consider related work that falls under the ML based IDS, utilizes feature selection or ensemble classifier, and especially focuses on hybrid approaches.

Hota and Shrivas [8] proposed a model that used different feature selection techniques to remove irrelevant features. The results indicate that C4.5 with information gain can achieve the highest accuracy with only 17 features for the NSL-KDD dataset.Abdullah et al. [31] also proposed a framework of IDS with selection of features within the NSL-KDD

dataset that are based on dividing the input dataset into different subsets, and combining them using Information Gain (IG) filter.Gaikwad and Thool [9] proposed a bagging ensemble method using REPTree as its base classifier, which takes less time to build the model and provide highest classification accuracy with lowest false positives on the NSL-KDD dataset.

Jabbar et al. [10] proposed a cluster-based ensemble classifier for IDS, which is built with Alternating Decision Tree (ADTree) and k-Nearest Neighbor algorithm (kNN). The experimental results show that the proposed ensemble classifier outperforms other existing techniques in terms of accuracy and detection rateMalik et al. [11] proposed a combination approach of Particle Swarm Optimization (PSO) and Random Forest (RF). More appropriate features for each class help the proposed model produce a higher accuracy along with low false positive rate compared with other algorithms.

Pham et al. [32] built a hybrid model, which utilizes gain ratio technique as feature selection and bagging to combine tree-based base classifiers. Experimental results show that the best performance was produced by the bagging model that used J48 as the base classifier and worked on 35-feature subset of the NSL-KDD dataset.

Abdullah et al. [30] also built an IDS using IG based feature selection and ensemble learning algorithms. The experiment on NSL-KDD dataset indicates that the highest accuracy obtained when using RF and PART as base classifiers under the product probability rule. In addition, Salo et al. [33] proposed a hybrid IDS which combines the feature selection approaches of IG and Principal Component Analysis (PCA) with an ensemble classifier based on Support Vector Machine (SVM), Instance-Based learning algorithms (IBK), and Multi-Layer Perceptron (MLP). A comparative analysis performed on several IDS datasets has proven that IG-PCA Ensemble method exhibits better performance than the majority of existing approaches.

Due to large-scale data produced from a massive network infrastructure, Khan et al. [34] proposed a scalable and hybrid IDS, which is based on Spark ML and Convolutional-LSTM (Conv-LSTM) network to employ the anomaly and misuse detection separately. Zhong et al. [35] also proposed a new anomaly detection model called HELAD, which is based on the Damped Incremental Statistics algorithm for feature selection and organic integration of multiple deep learning techniques for classification.

In 2008, Zhou, Jianguo, et al. Proposed system a Culture Particle Swarm Optimization algorithm (CPSO) used to optimize the parameters of SVM. By using the colony aptitude of particle swarm and the ability of conserving the evolving knowledge of the culture algorithm, this CPSO algorithm constructed the population space based on particle swarm and the knowledge space. The proposed CPSO-SVM model that can choose optimal values of SVM parameters was test on the prediction of financial distress of listed companies in China .

In 2011, Kolias, Constantinos, Georgios Kambourakis, and M. Maragoudakis et al. suggested that the RBF has certain parameter that affects the accuracy. PSO is used along with RBF artificial neural network it will improve the accuracy. If it is used in IDS it will improves the accuracy of classification.

In 2011, Homg, Shi-Jinn, et al. proposed an SVM based intrusion detection

system, which used a hierarchical clustering algorithm, leave one out, and the SVM technique. The hierarchical clustering algorithm provided the SVM with fewer, abstracted, and higher-qualified training instances that are derived from the KDD Cup 1999 training set. It was able to greatly minimize the training time, and improve the performance of SVM. The simple feature selection procedure (leave one out) was applied to eliminate unimportant features from the training set so the obtained SVM model could classify the network traffic data more accurately.

In 2012, Gaspar, Paulo, Jaime Carbonell, and Jose Luis Oliveira et al. gave the review on strategies that are used to improve the classification performance in term of accuracy of SVMs and perform some experimentation to study the influence of features and hyper-parameters in the optimization process, using kernels function. Huang et al provide a study on the joint optimization of C and g parameters (using the RBF kernel), and feature selection using Grid search and genetic algorithms .

In 2014, Ahmad, Iftikhar, et al. proposed a genetic algorithm to search the genetic principal components that offers a subset of features with optimal sensitivity and the highest discriminatory power. The support vector machine (SVM) is used for classification. The results show that proposed method enhances SVM performance in intrusion detection.

3.Developed Methodology

In an effort to increase the detection ability of IDS and prevent the service providers from attacks , we propose an efficient ML-based IDS using a metaheuristic optimization algorithm based feature selection approach, and a vote classifier which is an ensemble of classifiers method . During the experiments, 10-fold cross-validation (CV) approach is used to validate the performance of the model and classify benign traffic and various types of attacks.

Fig. 3 demonstrates the detection framework of the proposed ML-based IDS, which consists of the following four main phases:

• Datasets pre-processing: The first phase is to remodel raw data into a format suitable for analysis by applying pre-processing to the original datasets.

• Dimensionality reduction: In order to overcome the problem of high-dimensional datasets, the feature selection approach based on PSO is used with SVM to reduce the dimensionality of the datasets and select the most fitting features for each type of attacks.

• Classifiers training: For purpose of improving the accuracy of the IDS, we train three individual classifiers as base learners using ID3 and CART, and build an ensemble classifier based on them.

• Attack recognition: The detection model is tested using a 10-fold cross-validation approach, and voting technique is used to combine the probability distributions of the base learners with the AOP combination rule to make classification decision

Finally ,Based on the results of the ensemble classifier, benign traffic and various intrusive events can be detected and classified with high classification accuracy. Detailed information about the framework is provided in Sections .

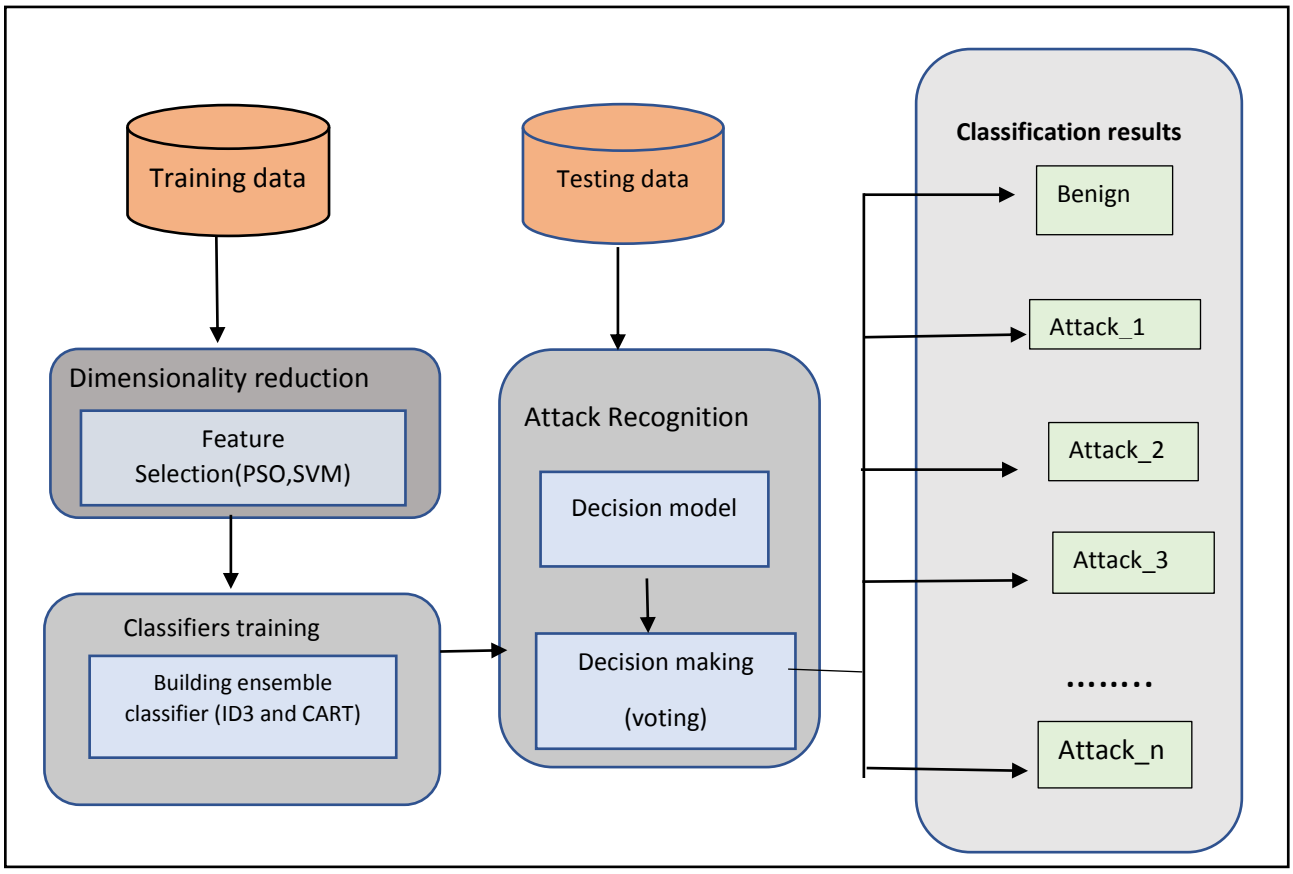

Fig. 3 The structure of the proposed Feature selection -Ensemble method

3.1Feature Selection

Feature Selection helps in reducing the dimensionality of the data set . Feature Selection extracts apt features from original data set effectively but are not suitable for all learning problems .Feature Selection as an alternative to feature extraction is often used as a pre-processing step in Machine Learning . Feature Selection approaches can be mainly categorised into wrapper ,filter and embedded approaches[12].Filter method pre-processes the data .These consider the relationship between features to calculate and predict the target features . Wrapper methods evaluate subset of features by their predictive by statistical reasoning or cross validation .Wrapper method is dependent on classification algorithm[13] .It has two parts named as search and evaluation .The search process deals with parameter initialisation that are used for evaluation of feature using evaluation function .It consist of both forward and backward selection.

The forward selection method initialises an empty set of features and iteratively evaluates features one by one .For every step , the feature that gets the maximum value of the evaluation function compared to the available set is included[15].The process terminates when there is no improvement in the evaluation function is found .While the backward elimination method initiates the selection process with the entire data set and removes the features one by one in each iteration , if the elimination of that particular feature improves the performance . The search process stops if the elimination of the feature decreases the value [15].

The embedded feature selection is combination of wrapper and filter based approach . It implicitly or explicitly uses FS technique to improve the performance of the classifier .

3.1.1 Feature Selection Process

In order to select the features in FS process ,initially entire feature set is considered for classification .The features are then selected by applying the FS methods. The basic steps in the FS process are given below fig.4 .

→Generating the subset of features

→Evaluation of the generated feature set

→Termination criterion

→Validating the results obtained for the given subset of feature

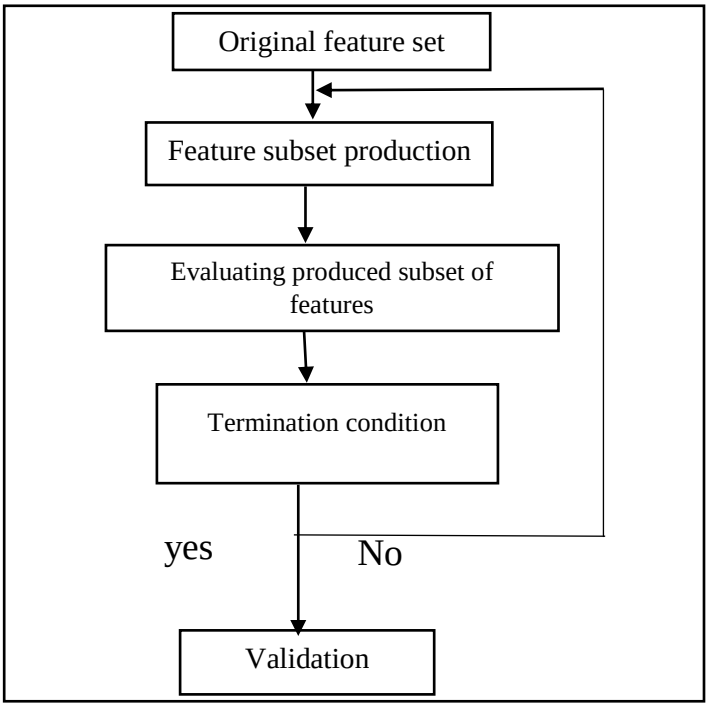

Fig .4 Steps of feature selection process [14]

3.1.2 Particle Swarm Optimisation approach for feature selection

In this section, we propose Particle Swarm Optimisation (PSO) with SVM based feature selection approach .Particle Swarm Optimisation is a parallel evolutionary computation technique developed by Mishra and Senguta[16]. The PSO algorithm's performance is greatly influenced by the included tuning parameters ,often referred to as the exploration-exploitation trade off : whereby exploration describes the ability to assess various regions in the problem space to an attempt to pinpoint a good optimum ,preferably the global one and exploitation describes the ability to focus the search within near vicinity of a promising candidate solution , to effectively and quickly locate the optimum [17].

The objective function of PSO algorithm used to evaluate its solutions , and operate upon the resultant fitness values . Each particle saves its position , composed of the candidate solution and its evaluates fitness , and its velocity [18]. PSO algorithm has been used in many applications to solve many problems[16,19-22]

The velocity and position are updated using below equations

$$\vec{v}_i^{d+1} = w\vec{v}_i^{d} + C1\gamma_1(\overrightarrow{P_i}d - \vec{x}_i^{d}) + C2\gamma_2(\overrightarrow{G_d} - \overrightarrow{x_{id}}) \quad (1)$$

$$\vec{x}_i^{d+1} = \vec{x}_i^{d} + \vec{v}_i^{d+1} \quad (2)$$

The velocity and position of each particle are represented as the vectors $\vec{v}_{\mathrm{i}}^{d} = [\vec{v}_1^{d}\ \vec{v}_2^{d}\ \vec{v}_3^{d}, \ldots\ldots\ldots\ldots]$ and $\overrightarrow{x_{id}} = [x_1^d, x_2^d, x_3^d, \ldots\ldots\ldots.]$ respectively .In equation 1 $\overrightarrow{P_i}d$ represents the local best and $\overrightarrow{G_d}$ represents global best positions . $C1$ and $C2$ are called acceleration factors known as cognitive and social parameters . γ_1 and γ_2 are random number between 0 and 1 .i is the iteration index . w is the inertia weight parameter.

Local best and global best are found using below formulae

$$\vec{P}\mathrm{i}^{d+1} = \max(CF(\mathrm{i}+1) + \vec{P}_i d) \quad (3)$$

$$\vec{G}_i^{d} = \max(\vec{P}_i^{d}) \quad (4)$$

Where CF is current fitness of the particular particle and $\vec{P}_i d$ is the local best of the particular particle .

The Particle Swarm Optimisation Algorithm is presented in Algorithm 1.The key parts of the PSO can be summarised as follows :

- Initialisation : The parameters of the algorithm and initialisation of population is done here.
- New solution generation : Here, the particles are moved in the search space according to the updating rules .
- Fitness Evaluation : The particles' positions are substituted in fitness function.
- Local best : The particles position for which maximum fitness is obtained is considered as local best .
- Global best :Minimum of all particles local best is considered as global best

3.2 Support Vector Machine (SVM)

SVM has been applied to variety of applications such as text categorization ,image processing ,attack classification are few applications [24].It is used for both classification and regression problems .It supports vectors .SVM performs better with linearly separable and also handles non -linear data by transforming data using kernel function to high dimensional feature space[13].Based on this learning ,the data set can be separated into two parts as working set and set of free variables .In the beginning ,SVM was used to address binary classification problems but it can also used for multi-class classification problems by decomposing the multi-class problems into several two class problems that can be addressed further by several SVMs[25].

Algorithm 1 Particle Swarm Optimisation approach for Feature Selection

Input: S_{tr} , S_{test}, C_l , C_u ,g_1 ,g_p

Output : C ,g

/* S_{tr} , S_{test} are the scaled training and testing dataset . C_l, C_u is the lower and upper limit of parameter C . g_1 ,g_p is the lower and upper limit of parameter g. */

Step 1: particle ={pos ,fitness ,velocity , bestpos ,bestfitness }

Step 2: initialise population of parameter [max_size] ,GlobalBestpos ,GlobalBest ,GlobalBestfit

For each swarm i from 1 to 10

Initialise pos in range [C_1 ,C_p] and [g_1 ,g_p]

//particle consist of two dimension C and g

$P_{fitness}$ = SVM(S_{tr} , S_{test} ,pos);

//calculating of fitness value based on mean square error(MSE) using SVM

Initialise velocity in range[C_l , C_u] and [g_1 ,g_p]

Swarm[i]← {pos, $P_{fitness}$, velocity,pos,$P_{fitness}$}

if(swarm[i].fitness < GlobalBestfit) then

GlobalBestfit =swarm[i].fitness;

GlobalBestpos=swarm[i].pos;

End if

End for

Step 3: choose particle with best fitness value

While(i <max_iteration)

Do for j from 1 to 10

Particle currPos =swarm[i**]**

Newvelocity = w* velocity[j] +(c1 * r1 *(curPos.bestPos - curPos.pos)) +(c2 *r2 *(GlobalBestpos -curPos.pos));

// w is inertia c1 ,c2 cognitive local and global weight

Newpos =pos+Newvelocity;

Newfit =SVM(Newpos);

if(Newfit < curPos.bestfit) then

GlobalBestpos=Newpos;

GlobalBestfit=Newfit;

End if

End for

End while

Feature Optimisation

Step 1: take l as the binary string of size 50

// as l = 010101010101010101…….

Step 2: particle ={pos ,fitness ,bestpos ,bestfitness }

Step 3: Particle [ ] swarm =new Particle [max _size], GlobalBestpos, GlobalBestfit

Step 4: do for each particle in swarm i from 1 to 10

pos=random_string(l);

writeRandomFeatures(pos,S_{str});

//In this function , featureSelection.txt produced from binary string S_{str} is scaled training dataset

Fitness =SVMF(featureSelection.txt ,$S_{tetx,}$C,g);

//In this function ,S_{test} is scaled test dataset , C and g is parameter obtain from parameter optimisation

swarm[i]← particle{pos ,fitness ,pos,fitness}

if(swarm[i].fitness <GlobalBestfit) then

GlobalBestfit=swarm[i].fitness;

GlobalBestpos=swarm[i].pos;

End if

End for

Step 5: do while i from 1 to max_iteration

Newpos;Newfit ;

Do for j from 1 to 10

Particle P=swarm[i];

Newpos=random_string(pos);

writeRandomFeatures(Newpos,S_{str});

Newfit=SVMF(featureSelection.txt,S_{tet} ,C,g);

if(Newfit <P.bestfitness)

P.bestpos=Newpos;

P.bestfitness=Newfit;

End if

if(Newfit<GlobalBestpos)

GlobalBestpos =Newpos;

GlobalBestfit =Newfit;

End if

End for

End while

3.3 Ensemble Classifier

For ensemble learning ,the classification methods usually combine multiple base classifiers in some way to produce better accuracy[26].These classifiers are powerful to solve the same problem and collectively achieve a forecasting result with higher stability and accuracy by creating multiple independent models and combining them [27]. The need of employing ensemble classifiers to improve the effectiveness are representational issue ,statistical reason , and computational reason . First, sometimes a single classifier is not qualified to obtain the best representation in the hypothesis space, therefore, it is necessary to combine independent classifiers to improve the predictive performance. Second, if the input dataset is not sufficient to train the learning algorithm, the result may lead to a weak or false hypothesis. In the last case, in order to produce a suitable hypothesis, an individual classifier could spend a significant amount of computing time, in which the procedure will be more likely to cause problems[1].

There are two popular algorithms in ensemble learning namely ,Bagging[28] and Boosting[29] , usually produce good results and widely chosen to build many ensemble models .The other popular ensemble learning methods for improving the performance of classification are Voting ,Bayesian parameter averaging and Stacking .

Among all decision tree algorithms ,using ID3 algorithm decision trees are built iteratively by finding out the maximum Information Gain(IG) among all featured data columns to be represented as the node of the tree . This algorithm builds a short tree relatively in less time .Mean while CART is easily used in conjugation with different algorithms and unwrapping complex interdependence .Other key reasons include the ability to do data-set and cross validation and not affected by the outliners. Therefore ,ID3 and CART algorithms are selected to construct the ensemble for multi-class intrusion detection in this paper.

3.3.1 ID3

In decision tree learning, **ID3** (**Iterative Dichotomiser 3**) is an algorithm invented by Ross Quinlan[30] used to generate a decision tree from a dataset. ID3 is the precursor to the C4.5 algorithm, and is typically used in the machine learning and natural language processing **.** It is a classification algorithm that follows a greedy approach of building a decision tree by selecting a best attribute that yields maximum Information Gain (IG) or minimum Entropy(H) [1].

$$\text{Gain Ratio(A)} = \frac{Gain(A)}{SplitInfo(A)} \qquad (5)$$

In the process , an attribute with the highest information gain is chosen as splitting attribute for the node N .Information gain represents how much uncertainty in the set D is

reduced after it is partitioned on attribute A ,where uncertainty can be calculated by entropy as

$$\text{Entropy(D)} = -\sum_{x \in X} p(x) \log_2 p(x) \qquad (6)$$

Where X is the set of classes in D and p(x) is the proportion of the number of elements in class x to the number of elements in set D .

Likewise ,SplitInfo is the term which describes how equally the attributes splits the data and can be calculated as :

$$\text{SplitInfo(A)} = \sum_{j=1}^{n} \frac{|Dj|}{|D|} log(\frac{|Dj|}{|D|}) \qquad (7)$$

where $\frac{|Dj|}{|D|}$ represents the weight of the jth partition in the set D.

3.3.2 CART

The CART algorithm is a type of classification algorithm that is required to build a decision tree on the basis of Gini's impurity index. It is a basic machine learning algorithm and provides a wide variety of use cases. A statistician named Leo Breiman coined the phrase to describe Decision Tree algorithms that may be used for classification or regression predictive modelling issues.

CART is an umbrella word that refers to the following types of decision trees:

- **Classification Trees:** When the target variable is continuous, the tree is used to find the "class" into which the target variable is most likely to fall.
- **Regression trees:** These are used to forecast the value of a continuous variable.

In CART ,we create three initial splits based on each feature .Then we evaluate how well the split either minimised the error or improved the prediction .This process of branch splitting is iterated for further subdivisions until we reach the leaf node . This uses Gini index for splitting .

3.3.3 Vote

Vote is a meta algorithm which performs the decision process by applying several classifiers . It uses the power of several individual classifiers and applies a combination rule for the decision. For example, minimum probability, maximum probability, majority voting, product of probabilities, and average of probabilities are different algorithms for combination rules. In order to deal with the multi-class classification, majority voting could not be chosen because the number of classes is more than that of base classifiers. In this paper, average of probabilities approach is used to make decision, where the class label is determined based on the maximum value of the average of predicted probabilities.

4.Evaluation and results

As stated before, this paper aims to develop an efficient intrusion detection system with high accuracy and low false alarms. For this purpose, a hybrid method, combined PSO and SVM named PSO-SVM , is performed to determine a subset of the

original features in order to eliminate the irrelevant features, and improve the classification efficiency. In the classification step, an ensemble classifier combined two different algorithms, ID3 and CART based on AOP combination rule, is trained and tested based on three datasets . Even in the cases where the data is allowed to be released or shared for public use, it will be heavily anonymized or severely altered. This will cause a lot of the essential data components that are considered critical to the researchers to be lost or no longer reliable.

4.1 Description of datasets

During the evaluation of IDS, one of the challenges faced by researchers is finding a suitable dataset. Acquiring a real world dataset that represents the traffic flowing through the network without any sort of anonymization or modification is a problem that has been continuously encountered by the cybersecurity research community[37].

For this reason, many researchers have decided to use simulated datasets such as the most well-known KDDCup'99 dataset , or one of its contemporaries the NSL-KDD dataset. Recently there has been a significant effort to try and develop data sets that are reflective of real world data. In 2015, Kolias et al. [38] published Aegean WiFi Intrusion Dataset (AWID) dataset, which includes real traces of both normal and intrusive 802.11 traffic. In addition, in 2017, the Canadian Institute for Cybersecurity (CIC) published an intrusion detection dataset named CIC-IDS2017 [78], which resembles the true real-world data packet capture (PCAPs). Therefore, in this paper, experiments are conducted based on the NSL-KDD, AWID, and CIC-IDS2017 datasets.

4.1.1 NSL-KDD dataset

The NSL-KDD dataset was proposed in 2009 as a new revised version of the original dataset KDDCup'99 . On the one hand, NSL-KDD retained the advantageous and challenging characteristics of KDDCup'99. On the other hand, it addressed some drawbacks inherited from the original dataset by eliminating redundant records, rationalizing the number of instances, and maintaining the diversity of selected samples. It is worth noting that the NSL-KDD dataset is compiled to maximize the difficulty of prediction, which constitutes its outstanding characteristics. In order to group the records into five difficulty levels, the initial dataset was evaluated using several benchmark classifiers, and each instance was annotated with the number of its successful predictions . For each difficult level group, the amount of selected records is inversely proportional to the record percentages from the original KDDCup'99 dataset.

In this study, KDDTrain+, KDDTest+, and KDDTest21 sets of the NSL-KDD dataset are used. The KDDTrain+ set contains total 125,973 instances comprising of 58,630 instances of attack traffic and 67,343 instances of normal traffic. Whereas, the KDDTest+ set contains total 22,544 instances, and as a subset of the KDDTest+ set, the KDDTest21 set includes total 11,850 instances. Cross-validation is done on the the KDDTrain+ set in our experiments, and to extend this benchmark, we also consider a validation test using simple hold-out (train-test) approach applied on KDDTest+ and KDDTest-21 sets. A detailed overview of the instances is shown in Table 1.

Table 1. Statistics of the three sets of NSL-KDD data set

Class	KDD Train+	KDD Test+	KDDTest-21
Normal	67343	9711	2152
DoS	45927	7458	4342
PRB	11656	2421	2402
R2L	995	2754	2754
U2R	52	200	200
Attacks	58630	12833	9698
Total	125973	22544	11850

4.1.2. Aegean WiFi Intrusion Dataset (AWID) dataset

AWID was publicly available in 2015 as a collection of sets of WiFi network data, which contain real traces of both normal and intrusive data collected from real network environments [48]. Each record in the dataset is represented as a vector of 155 attributes, and each attribute has numeric or nominal values. Based on the number of target classes, the dataset can be classified into AWID-CLS dataset and AWID-ATK dataset. AWID-CLS dataset groups the instances into 4 main classes including normal, flooding, impersonation, and injection, while AWID-ATK dataset has 17 target classes that belong to the 4 main classes. On the other hand, based on the number of instances, all the datasets have two different versions: Full Set and Reduced Set. It is important to mention that these two versions are not related. The reduced set was collected independently from the full set at different times, with different tools, and in different environments. For this research we have conducted experiments on the the reduced four class dataset (AWID-CLS-R-Tst) by using cross-validation method for classification purposes. In general, AWID-CLS-R-Tst set includes total 575,643 instances, and more detailed information about the numbers of specific attacks can be seen in Table 2.

4.1.3. CIC-IDS2017 dataset

The CIC-IDS2017 dataset was published by Canadian Institute for Cybersecurity (CIC) in 2017, it contains benign and the most up-to-date common attacks [78]. It also includes the results of the network traffic analysis using CICFlowMeter with labeled flows based on the time stamp, source and destination IPs, source and destination ports, protocols, and attacks (CSV files). This is one of the newest intrusion detection datasets, which covers necessary criteria with updated attacks such as DDoS, Brute Force, XSS, SQL Injection, Infiltration, Port Scan, and Botnet. In detail, this dataset contains 2,830,743 records devised on 8 files and each record includes 78 different features with its label. In order to maintain the same order of magnitude of each dataset while taking into account the requirements of multi-classification, the Wednesday-working Hours set has been chosen for experiments through cross-validation method. This set includes total 691,406 instances belonging to 6 categories, and the static information of the set is given in Table 2.

Table 2. Statistics of the AWID and CIC-IDS2017 datasets

Class	KDD Train+	Class	KDDTest-21
Normal	530785	Normal	439683

Injection	16682	DoS slowloris	5796
Flooding	8097	DoS Slowhttptest	5499
Impersonation	20079	DoS Hulk	230124
		DoS GoldenEye	10293
		Heartbleed	11
Attacks	44858	Attacks	251723
Total	575643	Total	691406

4.2 Results and discussions

The performance of IDS is evaluated based on its capability of classifying network traffic into a correct type. In order to avoid the effect of data sampling when assessing the IDS, therefore, we conducted experiments by using repeated k-fold (kf) cross-validation method, and the value of k is considered as 10. In this paper, all the performance results reported are the average value of outputs from 10 iterations of 10f validation approach, and each experiment is repeated with different seed for avoiding biased results. More specifically, for each dataset, we provide the confusion matrix derived from the testing process of PSO-SVM -Ensemble, and compare the performance of the proposed algorithm with no feature selection and some state-of-the-art methods in terms of several detection metrics, including Accuracy (Acc), precision, Detection Rate (DR), F-Measure, Attack Detection Rate (ADR), and False Alarm Rate (FAR). The mathematical calculations of the utilized evaluation metrics are explained in[39].

Table 3.
The performance results based on the original features based on NSL- KDD with 10f validation. (41 features)

classifer	Acc	precision	DR	Fmeasure	ADR	FAR	MBT
ID3	0.942	0.945	0.940	0.944	0.913	0.034	16.89
CART	0.948	0.945	0.949	0.946	0.903	0.019	15.02
Ensemble	0.954	0.952	0.954	0.951	0.920	0.016	51.45

Table 3 summarises the performance based on the NSL-KDD dataset ,which includes the results of the base and ensemble classifier is not good enough in some metrics without implementing feature selection.

Table 4.
The performance results based on the selected features using PSO-SVM (10 features)

classifer	Acc	precision	DR	Fmeasure	ADR	FAR	MBT
ID3	0.988	0.987	0.987	0.986	0.987	0.012	2.92
CART	0.992	0.987	0.988	0.988	0.987	0.009	8.63
Ensemble	0.998	0.997	0.998	0.997	0.997	0.001	34.45

Table 4 says that the proposed PSO-SVM Ensemble method performs best on all the two sets. In detail, our model exhibits the highest accuracy of 0.998, FMeasure of 0.997, ADR of 0.997 and the lowest FAR of 0.001 based on the NSL-KDD dataset.

Table 5.

The performance results based on the original features based on AWID with 10f validation. (84 features)

classifier	Acc	precision	DR	Fmeasure	ADR	FAR	MBT
ID3	0.956	0.952	0.998	0.976	0.788	0.034	93.95
CART	0.977	0.985	0.995	0.988	0.784	0.005	14198
Ensemble	0.985	0.979	0.999	0.990	0.783	0.002	488.43

Table 6.

The performance results based on the selected features using PSO-SVM (10 features)

classifer	Acc	precision	DR	Fmeasue	ADR	FAR	MBT
ID3	0.985	0.984	0.985	0.985	0.914	0.011	9.97
CART	0.992	0.990	0.991	0.991	0.945	0.003	26.51
Ensemble	0.995	0.995	0.996	0.997	0.957	0.001	34.45

As seen in Table 5 and Table 6, the proposed PSO-SVM ,Ensemble approach still achieves the best performance results in most respects on the AWID dataset, such as the highest accuracy of 0.995, the highest ADR of 0.957, and the lowest FAR of 0.002. Each base classifier using the selected feature exhibits higher accuracy and ADR than the ensemble classifier with the original features, which strongly proves the effectiveness of the proposed feature selection.

Table 7.

The performance results based on the original features on CIC-IDS2017 with 10f validation. (78 features)

classifier	Acc	precision	DR	Fmeasure	ADR	FAR	MBT
ID3	0.960	0.962	0.983	0.974	0.917	0.016	212.95
CART	0.948	0.944	0.948	0.946	0.904	0.021	244.98
Ensemble	0.952	0.950	0.952	0.951	0.918	0.016	976.43

Table 8.

The performance results based on the selected features using PSO-SVM on CIC-IDS2017 with 10f validation (13 features)

classifer	Acc	precision	DR	Fmeasure	ADR	FAR	MBT
ID3	0.985	0.995	0.988	0.992	0.974	0.011	212.49
CART	0.994	0.994	0.988	0.985	0.987	0.009	244.85
Ensemble	0.998	0.997	0.999	0.999	0.997	0.001	97.95

Similarly, the result of the comparison on the CIC-IDS2017 dataset is shown in Table 7 and Table 8, the observation is that the performance of the proposed feature selection approach outperforms that of all features in every respect, and the CFS-BA-Ensemble approach achieves the highest accuracy rate of 0.998, DR of 0.999, and ADR of 0.997 with only 13 features, which also outperforms all other individual classifiers.

5.Conclusion

Although many machine learning approaches have been proposed to increase the efficacy of IDSs, it is still a problem for existing intrusion detection algorithms to achieve good performance. In this paper, to deal with the high-dimensional and unbalanced network traffic,

we propose a novel intrusion detection framework, which is based on the feature selection and ensemble learning techniques. First, we proposed PSO-SVM algorithm with the aim of selecting the optimal subset features. Then, the ensemble classifier based on CART, ID3 with the AOP rule is introduced to construct the classification model. Finally, the proposed IDS is evaluated by 10f cross-validation over three intrusion detection datasets.

The experimental results are promising with an accuracy of classification equal to 99.82%, 99.8% DR and 0.07% FAR with a subset of 10 features for the NSL-KDD dataset, and the obtained results for the AWID provide accuracy of 99.49% and 0.15% FAR with a subset composed of only 8 features. Remarkably, our model achieves the highest accuracy of 99.89% and DR of 99.9% on the subset of 13 features for the CIC-IDS2017 dataset.

Although the proposed PSO-SVM Ensemble method has indicated superior performance, in the future work, its capability could be further improved to deal with rare attacks from the massive network traffic

References

[1] Yuyang et.al Building an Efficient Intrusion Detection System Based on Feature Selection and Ensemble Classifier arXiv:1904.01352v4 [cs.CR] 2 Apr 2020

[2]Pham, N.T., Foo, E., Suriadi, S., Jeffrey, H., Lahza, H.F.M., 2018. Improving performance of intrusion detection system using ensemble methods and feature selection, in: Proceedings of the Australasian Computer Science Week Multiconference, ACM. p. 2. doi:10.1145/ 3167918.3167951.

[3] Du, M., Wang, K., Chen, Y., Wang, X., Sun, Y., 2018a. Big data privacy preserving in multi-access edge computing for heterogeneous internet of things. IEEE Communications Magazine 56, 62–67. doi:10.1109/MCOM.2018.1701148.

[4] Du, M., Wang, K., Xia, Z., Zhang, Y., 2018b. Differential privacy preserving of training model in wireless big data with edge computing. IEEE Transactions on Big Data doi:10.1109/TBDATA.2018.2829886.

[5] Mishra, P., Varadharajan, V., Tupakula, U., Pilli, E.S., 2018. A detailed investigation and analysis of using machine learning techniques for intrusion detection. IEEE Communications Surveys & Tutorials doi:10.1109/COMST.2018.2847722.

[6] Feng, X., Xiao, Z., Zhong, B., Qiu, J., Dong, Y., 2018. Dynamic ensemble classification for credit scoring using soft probability. Applied Soft Computing 65, 139–151. doi:10.1016/j.asoc.2018.01.021.

[7] Salo, F., Nassif, A.B., Essex, A., 2019. Dimensionality reduction with ig-pca and ensemble classifier for network intrusion detection. Computer Networks 148, 164–175. doi:10.1016/j.comnet.2018.11. 010.

[8] Hota, H., Shrivas, A.K., 2014. Decision tree techniques applied on nsl-kdd data and its comparison with various feature selection techniques, in: Advanced Computing, Networking and InformaticsVolume 1. Springer, pp. 205–211. doi:10.1007/978-3-319-07353-8_24.

[9] Gaikwad, D., Thool, R.C., 2015. Intrusion detection system using bagging ensemble method of machine learning, in: 2015 International Conference on Computing Communication Control and Automation, IEEE. pp. 291–295. doi:10.1109/ICCUBEA.2015.61.

[10] Jabbar, M., Aluvalu, R., Reddy, S.S.S., 2017. Cluster based ensemble classification for intrusion detection system, in: Proceedings of the 9th International Conference on Machine Learning and Computing, pp. 253–257. doi:10.1145/3055635.3056595.

[11] Malik, A.J., Shahzad, W., Khan, F.A., 2015. Network intrusion detection using hybrid binary pso and random forests algorithm. Security and Communication Networks 8, 2646–2660. doi:10.1002/sec.508.

[12] Hajisalem, V., Babaie, S., 2018. A hybrid intrusion detection system based on abc-afs algorithm for misuse and anomaly detection. Computer Networks 136, 37–50. doi:10.1016/j.comnet.2018.02.028

[13]Ankit Thakkar ,Ritika Lohiya . Journal of Ambient intelligence and Humanized Computing (2021) 12:1249 -1266 doi.org/10.1007/s12652-020-02167-9

[14] Kumari B,Swarnkar T(2011) filter versus wrapper feature subset selection in large dimensionality micro array : a review .Int J Comput Sci Inf Technol 2(3) :10488-1053

[15] Balasaraswathi VR,Sugumaran M ,Hamid Y (2017) Feature selection techniques for intrusion detection using non-bio-inspired and bio-inspired optimisation algorithms. J Commun Inform Netw 2(4):107-119

[16] V. K. Mishra and A.Sengunta,"MO-PSE:Adaptive multi-objective particle swarm optimisation based design space exploration in architectural synthesis for applications specific processor design,"Adv. Eng.softw., vol.67 ,pp. 111-124 ,Jan. 2014

[17] I. C.Trelea ,"The particle swarm optimisation algorithm :Convergence analysis and parameter selection ,"Inf.Process. Lett., vol.85 ,no. 6, pp. 317-325,2003.

[18] J.Blondin. Particle swarm optimisation: A Tutorial .[Online]. Available: //cs.armstrong.edu/saad/csci8100/pso_tutorial.pdf

[19]A.Sengupta, S. Bhadauria, and S. P.Mohanty, "TL-HLS : Methodology for low cost hardware trojan security aware scheduling with optimal loop unrolling factor during high level synthesis," IEEE Trans. Comput.-Aided Des. Integr . Circuits Syst., vol . 36, no. pp. 660-673, Apr .2017

[20] V. K. Mishra and A.Sengunta ,"swarm -inspired exploration of architecture and unrolling factors for nested-loop-based application in architectural synthesis," Electron .Lett.,vol. 51, no. 2, pp. 157-159,2015

[21]A. Sengupta and S.Bhadauria, "User power -delay budget driven PSO based design space exploration of optimal k-cycle transient fault secured datapath during high level synthesis ," in proc.-Int. Symp. Qual.Electron . Des. (ISQED), Apr. 2015 ,pp. 289-292

[22] A.Sengupta and V.K. Mishra,"Expert systems with applications automated exploration of datapath and unrolling factor during power-Performance tradeoff in architectural synthesis using multi-dimensional PSO algorithm ," Expert Syst. Appl., vol. 41 ,no. 10, pp. 4691-4703 ,2014.

[23] Vitthal Manekar, Kalyani Waghmare International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISS(online): 2277-7970)Volume-4 Number-3 Issue-16 September-2014

[24] Sutharaharan S (2016) support vector machine .In :Machine Learning models and algorithms for big data classification ,vol 36.Springer ,pp 207-235

[25] Meyer D. Wein FT (2015) support vector machines. Interf Libsvm pack e1071:28

[26] Feng, X., Xiao, Z., Zhong, B., Qiu, J., Dong, Y., 2018. Dynamic ensemble classification for credit scoring using soft probability. Applied Soft Computing 65, 139–151. doi:10.1016/j.asoc.2018.01.021.

[27] Li, H., Sun, J., 2013. Predicting business failure using an rsf-based case-based reasoning ensemble forecasting method. Journal of Forecasting 32, 180–192. doi:10.1002/for.1265.

[28] Breiman, L., 1996. Bagging predictors. Machine learning 24, 123– 140. doi:10.1007/BF00058655.

[29] Freund, Y., Schapire, R.E., et al., 1996. Experiments with a new boosting algorithm, in: icml, Citeseer. pp. 148–156

[30] Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81–106

[31] Abdullah, M., Balamash, A., Alshannaq, A., Almabdy, S., 2018. Enhanced intrusion detection system using feature selection method and ensemble learning algorithms. International Journal of Computer Science and Information Security (IJCSIS) 16

[32] Pham, N.T., Foo, E., Suriadi, S., Jeffrey, H., Lahza, H.F.M., 2018. Improving performance of intrusion detection system using ensemble methods and feature selection, in: Proceedings of the Australasian Computer Science Week Multiconference, ACM. p. 2. doi:10.1145/ 3167918.3167951.

[33] Salo, F., Nassif, A.B., Essex, A., 2019. Dimensionality reduction with ig-pca and ensemble classifier for network intrusion detection. Computer Networks 148, 164–175. doi:10.1016/j.comnet.2018.11. 010

[34]Khan, M.A., Karim, M., Kim, Y., et al., 2019. A scalable and hybrid intrusion detection system based on the convolutional-lstm network. Symmetry 11, 583. doi:10.3390/sym11040583

[35] Zhong, Y., Chen, W., Wang, Z., Chen, Y., Wang, K., Li, Y., Yin, X., Shi, X., Yang, J., Li, K., 2020. Helad: A novel network anomaly detection model based on heterogeneous ensemble learning. Computer Networks 169, 107049. doi:10.1016/j.comnet.2019.107049.

[36] Denning DE (1987) An intrusion -Detection Model .IEEE Trans Softw Eng 2:222-232

[37] Aldwairi, T., Perera, D., Novotny, M.A., 2018. An evaluation of the performance of restricted boltzmann machines as a model for anomaly network intrusion detection. Computer Networks 144, 111– 119. doi:10.1016/j.comnet.2018.07.025.

[38] Kolias, C., Kambourakis, G., Stavrou, A., Gritzalis, S., 2015. Intrusion detection in 802.11 networks: empirical evaluation of threats and a public dataset. IEEE Communications Surveys & Tutorials 18, 184–208. doi:10.1109/COMST.2015.2402161.

[39] Elhag, S., Fernández, A., Altalhi, A., Alshomrani, S., Herrera, F., 2019. A multi-objective evolutionary fuzzy system to obtain a broad and accurate set of solutions in intrusion detection systems. Soft Computing 23, 1321–1336. doi:10.1007/s00500-017-2856-4.

[40].R Gangadhar Reddy, M. Srinivasa Reddy, P R Anisha, Kishor Kumar Reddy C, "Identification OF EARTHQUAKES USING WAVELET TRANSFORM AND CLUSTERING METHODOLOGIES", International Journal of Civil Engineering and Technology, 2017. (http://www.iaeme.com/MasterAdmin/UploadFolder/IJCIET_08_08_067/IJCIET_08_08_067.pdf)

[41].B Subbarayudu, Srija Harshika D, E Amareswar, R Gangadhar Reddy, Kishor Kumar Reddy C, "Review and Comparison on Software Process Models", International Journal of Mechanical Engineering and Technology, 2017.

(http://www.iaeme.com/MasterAdmin/UploadFolder/IJMET_08_08_105/IJMET_08_08_105.pdf)

[42].Kishor Kumar Reddy C and Vijaya Babu B, "ISLGAS: Improved Supervised Learning in Quest using Gain Ratio as Attribute Selection Measure", International Journal of Control Theory and Applications, 2016.

[43].Kishor Kumar Reddy C and Vijaya Babu B "A Survey on Issues of Decision Tree and Non-Decision Tree Algorithms", International Journal of Artificial Intelligence and Applications for Smart Devices, 2016.

[44].Kishor Kumar Reddy C, Rupa C H and Vijaya Babu B, "ISLIQ: Improved Supervised Learning in Quest to Nowcast the presence of Snow/No-Snow", WSEAS Transactions on Computers, 2016.

(http://www.wseas.org/multimedia/journals/computers/2016/a055705-872.pdf)

[45].Kishor Kumar Reddy C and Vijaya Babu B, "ISPM: Improved Snow Prediction Model to Nowcast the Presence of Snow/No-Snow", International Review on Computers and Software, 2015.

(http://www.praiseworthyprize.org/jsm/index.php?journal=irecos&page=article&op=view&path%5B%5D=17055)

[46].Kishor Kumar Reddy C, Rupa C H and Vijaya Babu B, "SLGAS: Supervised Learning using Gain Ratio as Attribute Selection Measure to Nowcast Snow/No-Snow", International Review on Computers and Software, 2015.

(http://www.praiseworthyprize.org/jsm/index.php?journal=irecos&page=article&op=view&path%5B%5D=16706)

[47].Kishor Kumar Reddy C, Vijaya Babu B, Rupa C H, "SLEAS: Supervised Learning using Entropy as Attribute Selection Measure", International Journal of Engineering and Technology, 2014.

(http://www.enggjournals.com/ijet/docs/IJET14-06-05-210.pdf)

[48].Kishor Kumar Reddy C, Rupa C H and Vijaya Babu B, "A Pragmatic Methodology to Predict the Presence of Snow/No-Snow using Supervised Learning Methodologies", International Journal of Applied Engineering Research, 2014.

(http://www.ripublication.com/Volume/ijaerv9n21.htm)

[49].Kishor Kumar Reddy C, Rupa C H and Vijaya Babu, "SPM: A Fast and Scalable Model for Predicting Snow/No-Snow", World Applied Sciences Journal, 2014.

(http://www.idosi.org/wasj/wasj32(8)14/14.pdf)

[50].Kishor Kumar Reddy C, Anisha P R, Narasimha Prasad L V and Dr. B Vijaya Babu, "Comparison of HAAR, DB, SYM and COIF Wavelet Transforms in the Detection of Earthquakes Using Seismic Signals", International Journal of Applied Engineering Research, 2014, pp. 5439-5452.