

We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,500

Open access books available

176,000

International authors and editors

190M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



Chapter

Efficient Machine Learning Classifier for Fault Detection in Wireless Sensor Networks

Poornima G. Miathali

Abstract

The deployment of wireless sensor networks in unpredictable and dangerous conditions makes them prone to software, hardware, and communication errors. Sensors are physical devices that are deployed in inaccessible environment which makes them malicious. The Fault occurs in the sensed data and its detection should be precise and rapid to limit the loss. The status of sensed data should be explicitly determined to guarantee the normal function of the Sensor Networks. For the purpose of fault detection machine learning classifiers are employed because they are effective and used to classify sensed data into faulty and non-faulty data. The faults due to Dos, Probe, R2L, and U2R are considered for implementation. KDD CUP 99 dataset is chosen for training and test purpose, and the dataset contains 41 features which are categorized as content, basic, TCP features. The required feature for each fault category is selected through recursive feature elimination technique. The performance of the classifier is measured and evaluated in terms of Accuracy, precision, recall, and F-measures. From experimental results, it is observed that Random Forest classifier is best suited for Wireless Sensor Networks fault detection. The simulation result shows that Multi-layer perceptron outperforms the other classifier with 92% of accuracy.

Keywords: attacks, classifiers, sensor networks, machine learning, random forest, support vector machine, multilayer perceptron, stochastic gradient descent, faults

1. Introduction

Wireless sensor networks are widely used for a variety of purposes, including systems that must function safely. More mission-critical subsystems like cars, drones, and others are joining the area of WSNs, although historically, geographically close systems would link wirelessly over time. As a result, it has become imperative to create WSNs that are fault tolerant. Data security plays a crucial part in successful communication. In earlier days for the security purpose Encryption, Firewall, Virtual Private networks (VPN) were used to provide data security. But these methods are not enough to secure the data. Therefore, machine learning approach gives an effective way to deal with the problem. Many researches have performed studies and arrived at various conclusions on data safety. With hardware implementation, these

methods seem to be complex. So, machine learning gives the best solution for the problem. This is the easiest way and does not consume large amount of time compared to other methods and at the same time it is a cost-effective method.

From Education to Entertainment industry data is the backbone. Therefore data security and safety is significant. Hackers may duplicate the data packets or even IP address itself therefore it is difficult to identify the malicious data in the network. Machine learning techniques give the efficient solution.

A Hardware model is implemented [1], using sensors. This method seems to be complicated. Attack detection has achieved through block chain technology [2], But this method suffers from computational delay, block chain overheads, cost of implementations. Other machine learning classifiers are used to find the attacks. The major drawback from the research is more computational time and more false positive values [3, 4]. In the present study false positive values are comparatively less and it is shown in the confusion matrix and it is discussed in the result section.

The Data set considered for the present study is KDD Cup 99 dataset which contains large number of data sets and is publically available. The major attacks that are considered in this attack are DoS (Denial of Service) attack in which an unauthorized user getting access to the network. Probe attack. R2L (Remote to user) attack in which an unauthorized user can send data packets to the system where he or she cannot have the access as a local user. U2R (User to root) attack in which the unauthorized can get into the root.

To analyze the data as attacked or normal data four classifiers are considered, RF (Random Forest), Support Vector machine (SVM), Multilayer Perceptron (MLP), Stochastic Gradient Descent (SGD) Classifiers are used. Raw data cannot be used to test and train the machine learning model. So, Data preprocessing steps such as Feature Selection, Encoding. The Preprocessed data are applied to the different classifiers. Efficiency parameters such as accuracy, precision, recall, F-measure, selectivity, specificity, G-mean are found out. By comparing all these parameters the final result can be achieved. Different percentage of attacks can be introduced in the data set. So that an efficient classifier can be found out for different percentage of attacks. The Efficiency parameter can be obtained from Confusion matrix. Confusion matrix contains True positive (TP), True Negative (TN), False Positive (FP), False Negative (FN) value.

In the present study, a brief description on available data set in the internet is presented. Further, pre-processing of data is discussed in detail. When data sets are applied to 4 different types of classifiers, efficient classifier is derived with respect to confusion matrix parameter.

The paper is organized as follows, Section 2 the motivation for the present study is discussed, Section 3 reviews the related works carried out in the field of intrusion detection system and various data faults that occur and the type of classifiers used is presented. Section 4 introduces the proposed Method, In Section 5 discusses the performance measures and analyses. The Paper finally concludes with Section 6 with future research directions.

2. Motivation

Wireless sensor networks are widely used for a variety of purposes, including systems that must function safely. More mission-critical subsystems like cars, drones, and others are joining the area of WSNs, although historically, geographically close systems would link wirelessly over time. As a result, it has become imperative to

create WSNs that are fault tolerant. The nature of the defects that are likely, as described in the introduction section, makes it appropriate to cutting-edge technology, such as machine learning, to find such faults.

3. Literature survey

The focus of the research work presented in this paper is on the detection of faults due to attacks and the methods used to detect and classify the data.

Zainib Noshad, Nadeem Javaid, Tanzila Saba [1] The use of wireless sensor networks (WSNs) in a variety of environments makes them susceptible to errors. Unstable and dangerous conditions. This makes WSN vulnerable to errors in software, faults in hardware and communication. Fault detection in WSNs has become a challenging task because of the sensor's constrained resources and varied deployment environments. The classification of gain, offset, spike, data loss, out of bounds, and stuck-at faults at the sensor level is done using Support Vector Machine (SVM), Convolutional Neural Network (CNN), Stochastic Gradient Descent (SGD), Multi-layer Perceptron (MLP), Random Forest (RF), and Probabilistic Neural Network (PNN) classifiers. Two of the six faults—the spike and data loss faults—are brought on by the datasets. The Detection Accuracy (DA), True Positive Rate (TPR), Matthews Correlation Coefficients (MCC), and F1-score are used to compare the results. Simulations demonstrate that the RF method achieves a higher rate of defect detection than the other classifiers.

Salah Zidi, Tarek Moulahi, and Bechir Alaya [2] one of the easiest ways to find failure in WSNs appears to be to use machine learning. SVM is employed in our context to define a decision function, which is based on statistical learning theory. This technique has a lot of potential for multidimensional data learning in addition to having demonstrated performance in a number of fields. This method, which makes use of kernel functions, has a significant capacity for adaptability for nonlinear classification scenarios, such as our case of fault detection. This has the potential to be very helpful in fault prevention. The goals of this research are to use a dynamic classification approach to track sensor activity through its data in order to predict errors as quickly as feasible in the same context of prevention.

Terry Windeatt [3] Multilayer Perceptron (MLP) classifier settings can be difficult to adjust, as is widely known. In this study, a metric that can forecast how many classifier training iterations will take to get the best results from an ensemble of MLP classifiers is described. The measure, which is based on a spectral representation of a Boolean function, is computed between pairs of patterns on the training data. With this representation, accuracy and diversity can be combined into a single statistic that describes the mapping from classifier decisions to the target label.

Luofan Dong, Huaqiang Du, Fangjie Mao [4] Convolutional neural networks (CNNs) recently demonstrated outstanding performance in a variety of applications, including computer vision and remote sensing semantic segmentation. Much interest is focused on the capacity to learn CNN's high-representational properties. On the other hand, the random forest (RF) technique is frequently used for variable selection, classification, and regression. This article tested a technique based on the fusion of an RF classifier and the CNN for a very high-resolution remote sensing (VHRRS) based forests mapping. This method was based on the previous fusion models that fused CNN with the other models, such as conditional random fields (CRFs), support vector machine (SVM), and RF. Huwaida T. Elshoush, Esraa A. Dinar [5] Spam prevalence is

rising daily as electronic emails are used more frequently. As a result, spam emails have grown to be a serious issue that reduces the use of electronic emails for communication. Several machine learning approaches, including Naive Bayes (NB), K-Nearest Neighbor (KNN), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Decision Tree, provide email spam filtering solutions (DT). This study examines various machine learning methods, namely Adaboost and Stochastic Gradient Descent, to filter spam emails (SGD).

Mrutyunjaya Panda, Ajith Abraham [6] Security of network traffic is growing to be a significant issue for computer networks as the internet expands. The frequency of attacks on the network is rising over time. Such network attacks are nothing more than intrusions. The network and data have been protected against threats by using intrusion detection systems to identify intrusions. Large amounts of network data are monitored, analyzed, and classified into abnormal and regular data using data mining algorithms. Poornima G, K Suresh Babu, K B Raja, K R Venugopal, and L M Patnaik [7] proposed to find the probability of correctly identifying a faulty node for three different types of faults based on normal bias. The nodes fault status is declared based on its confidence score that depends on the threshold value. Uma R. Salunkhe, Suresh N. Mali [8] used an intrusion detection system (IDS) to detect hostile activity has been an efficient technique to increase security. An intrusion detection system is anomaly detection. Due to its inability to accurately detect all sorts of attacks, current anomaly detection is frequently characterized by high false alarm rates and only modest accuracy and detection rates. Using the KDD-99 Cup and NSL-KDD datasets, a test is run to assess how well the various machine learning methods perform.

Miao X, Liu Y, Zhao H, Li C [9] system which detects the attacks in the wireless sensor network the KDD cup 99 data set is used in the present paper and the to classify the attacks in the WSN's the KNN classifier is used, But the detection rate achieved with this classifier is very poor and the highest detection rate is 75% and that is for $k = 5$. Gharghan S.K, Nordin R, Ismail M, Ali J. A [10] a hardware model for intrusion detection system is suggested this model has failed to give the accurate result, due to some hardware vulnerabilities and it is complex to design and human intervention is required.

In [11] the authors have discussed Intrusion detection system and used Decision tree, SVM, MLP algorithm. The result shows that MLP outperforms the other classifier with accuracy of 91%. In [12] the authors elaborate on layer wise DoS attack and its defense mechanisms and classification. In [13] the authors detect faults in WSN using hidden Markov model, KDD cup 99 data set is used, the accuracy they have achieved for test data is 77.11%. In paper [14] fault detection using deep learning algorithms is done. KDD cup 99 data set is used and MLP, SVM algorithms are used and the accuracy is 91%. In [15] the fault detection in WSN using Internet of things based on improved BP Neural network Leven berg- Marquard algorithm is applied with a accuracy result of 91%.

From the papers surveyed, for selecting feature subset Recursive feature elimination method is implemented. All the independent variables in supervised learning is known as features of the data. Elimination in this context means eliminating the features. Doing a process repetitively to eliminate the features of the data is known as Recursive feature elimination. KDD is a type of data set and an online repository that contains data from all different sorts of intrusion attempts. It mainly includes DOS, R2L, U2R, and PROBE. RF, SVM, SGD, MLP classifiers will be assessed on the KDD dataset in this research.

4. Methodology

The methodology used in this work is shown in the **Figure 1** and it involves preprocessing the KDD dataset initially, using the prepared dataset in a fair environment with equal access to resources, and then comparing classifier performance across all analyzed attacks (DOS, R2L, U2R, and PROBE) and their faults. Machine learning model needs large number of data set to avoid the problem of over fitting. The Proposed optimal feature subset selection algorithm includes feature normalization, feature scoring, feature subset selection and feature subset elimination.

Data Preprocessing is the most time-consuming task but plays significant role in machine learning model. Raw data cannot be used for training or testing the machine learning model. Hence data preprocessing is required in machine learning. Encoding is the process of converting the categorical data into numerical value. Categorical values are the string values that are stored as components of the input features. Features/ attributes that have strings or categories as their values are termed as categorical attributes. These Categorical values can be represented in two forms, namely Nominal and Ordinal. When there is no ordering between the attributes those are referred to as Nominal attributes. When there is an ordering between the attributes those are referred to as Ordinal attributes. KDD cup 99 data set contains 125,973 train data and 22,544 test data. So, it helps to build and test an efficient machine learning model. It also contains different type of attacks called Neptune, pod, smurf, etc. which can be further categorized as DoS, Probe, U2R and R2L attacks [16] as shown in **Table 1**.

In NSL-KDD cup 99 data set with 41 input features are present. In that protocol feature contains tcp, udp, icmp etc., Service feature contains http, ftp, telnet etc., Flag feature contains SF, REL, ROIT etc. These three columns contain symbolic and continuous data which cannot be used. Because the classifier that are considered, accepts only numerical values. Hence One Hot Encoding technique is used. One hot encoding columns are exactly equal to the number of the values a particular feature is having. While encoding, there should be only one value present only once in the encoded values.

Feature Scaling is the process of transforming the data value into 0 to 1. For example if we consider two weights whose values are 80 and 40 respectively, by

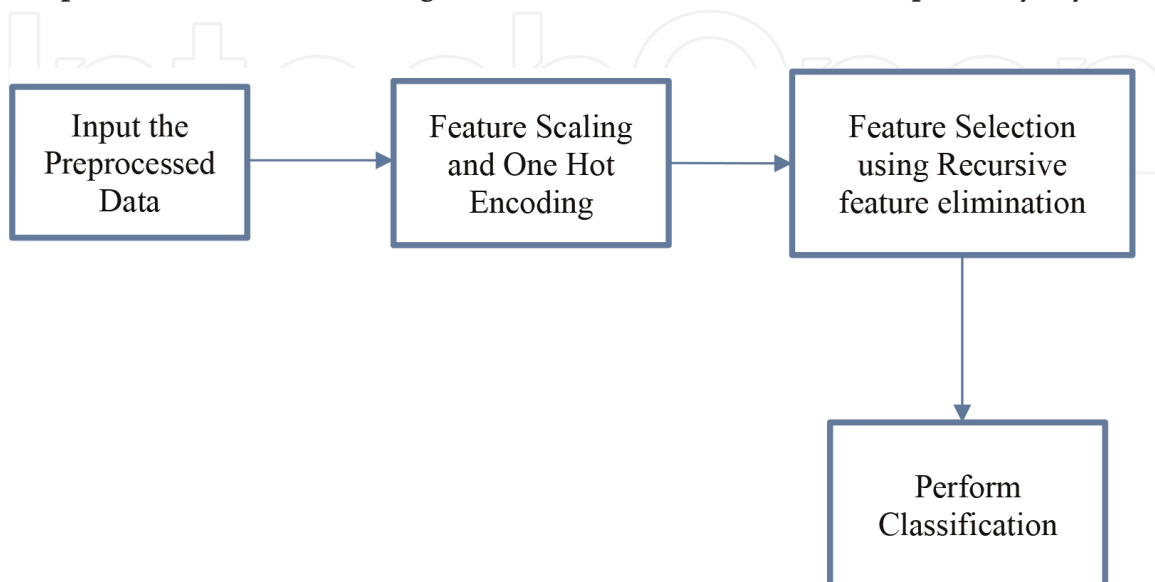


Figure 1.
Block diagram of the proposed method.

Attack types	DoS attack	Probe attack	U2R attack	R2L attack
Known attack	Neptune, back, land, pod, smurf teardrop	Ipsweep, nmap, portsweep, satan	ftp_write, guess_password, immap, multihop, phf, spy, wazerclient, wazermaster	Buffer_overflow, landmod, perl, rootkit
New attack	Mailbomb, apache2, processtable, worm	mscan, saint	Sendmail, snmpattack, snmpguess, httptunnel	Ps, sqlattack, xterm

Table 1.
Types of attacks and faults in the data.

feature scaling these can be represented as 1 and 0 where 0 is the lowest possible score/weight and 1 is the highest possible score/weight.

Feature Selection: Machine learning model needs to be trained by huge number of data set for the accurate result. But some data does not contain useful information, without considering that feature also classification can be done. This process is known as Feature selection. Feature Selection basically a process where in only few features that contains the useful information can be selected and machine learning model will get rid of the noise data. Recursive feature elimination method is used for feature selection. All the independent variables in supervised learning is known as features of the data. Elimination in this context means eliminating the features. Doing a process repetitively to eliminate the features of the data is known as Recursive feature elimination (RFE). RFE works by searching for a subset of features by starting with all features in the training dataset and successfully removing features until the desired number remains. KDD cup 99 data set has 41 input features. Among 41 features, 23 features are selected for the model by using the recursive feature elimination technique. (Table 2).

Classifiers that are used to classify the malicious and the normal data are:

1. Random Forest Algorithm [17]: Decision Trees are very sensitive by nature, necessitating the use of the Random Forest algorithm. The decision tree's entire structure may change if the training data set has a slight difference. Because of this, it is extremely sensitive, and the outcome is highly variable. Decision trees are the binary trees that recursively splits the data set until we are left with pure

Selected features
src_bytes; Dst_bytes; Wrong_fragment; Num_compromised; Count; Srv_count;
Same_srv_rate; Diff_srv_rate; Dst_host_same_srv_rate; Dst_host_error_rate;
Dst_host_srv_error_rate; Protocol_type; Dst_host_diff_srv_rate; Dst_host_same_src_port_rate
Dst_host_srv_diff_host_rate; Service_eco_i; Hot; Logged_in
Is_guest_login; Dst_host_count; Dst_host_srv_count

Table 2.
Selected features.

leaf nodes. Bootstrapping is the process of building a new data set from an existing one. We must use the bootstrapped data sets to train the decision trees. This is how the data is aggregated using Random Forest Classifier.

2. Support Vector machine: SVM classifier comes under the supervised learning. When the data is of 2- dimensional space then a line which separates the two classes needs to be created. But the data set is in 3 dimensional then a hyper plane that will separate the 3-dimensional data sets needs to be created. a line can be used to demarcate two-dimensional linearly separable data. $Y = Ax+B$ is the definition of the line's function. The equation becomes $x_2 = ax_1 + b$ if we replace x in this case with x_1 and y with x_2 . The new form of this equation is $ax_1 - x_2 + b = 0$. We will obtain $wx + b = 0$ if we define $x = (x_1, x_2)$ and $w = (a,-1)$. Thus, we shall obtain the line's equation. The same line is used to divide the two data classes in logistic regression as well, however the problem with logistic regression is that it does not care if the cases are actually close to the line or not.
3. Multilayer Perceptron: A fully connected class of feed forward artificial neural network is called a multilayer perceptron (MLP) (ANN). The term "MLP" is used ambiguously; sometimes it is used broadly to refer to any feed forward ANN, and other times it is used specifically to describe networks made up of several layers of perceptron's (with threshold activation). Especially when they comprise a single hidden layer, multilayer perceptron's are commonly referred to as "vanilla" neural networks in common parlance. In MLP input layer, output layer and number of hidden layers are used depending on weights and activation function the classification is done in output layer. Each node, with the exception of the input nodes, is a neuron that employs a nonlinear activation function. Back propagation is a supervised learning method that is used by MLP during training. MLP differs from a linear perceptron [18] due to its numerous layers and non-linear activation. It can identify non- linearly separable data.
4. Stochastic gradient descent: It is an iterative process to optimize the objective function. Gradient simply refers to a surface's slope or tilt. Gradient descent is an iterative process that descends a function's slope in steps from a random point until it reaches the function's lowest point.

Performance Evaluation Measures: In this section, we provide a detailed evaluation of the machine learning techniques with various performance measures to detect network faults caused due to intrusions.

4.1 Confusion matrix

Confusion matrices are a widely used measurement when attempting to solve classification issues. Both binary classification and multi class classification issues can be solved with it. In Confusion matrix there are values which are called True Positive, True Negative, False Positive and False Negative. True Positive Constitutes the data features that are correctly identified by the Algorithm. True Negatives are also the values that are correctly identified by the algorithm. False Positive and the False Negative are the data features that are wrongly identified by the Algorithm. The Confusion matrix in machine learning is used to find the Precision and Accuracy of the Classifier which we can obtain those from True and False Values. After the

classifiers are trained the performance of all 4 classifiers are measured in terms of these metrics using test data set. Based on the Confusion Matrix, Accuracy, Precision, Recall, F-measure, Specificity, Selectivity, G-mean are calculated as mentioned below,

1. **Accuracy:** Accuracy of a classifier can be calculated as ratio total true values with all the values present in the confusion matrix.

$$\text{Accuracy} = (\text{TP} + \text{TN}) / (\text{TP} + \text{TN} + \text{FP} + \text{FN}) \quad (1)$$

2. **Precision:** Precision is determined by dividing the total number of optimistic predictions by the actual number of optimistic predictions.

$$\text{Precision} = (\text{TP}) / (\text{TP} + \text{FP}) \quad (2)$$

3. **Recall:** It is obtained by dividing the sum of all valid samples by the total number of valid positive predictions.

$$\text{Recall} = (\text{TP}) / (\text{TP} + \text{FN}) \quad (3)$$

4. **F-measure:** The F1 score is defined as the harmonic mean of precision and recall.

$$F - \text{measure} = 2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall}) \quad (4)$$

5. **G-mean:** Geometric mean is the square root of true positive rate and true negative rate.

$$\text{Gmean} = \sqrt{\text{TPR} * \text{TNR}} \quad (5)$$

6. **Selectivity:** Sensitivity is the ratio between the total number of positive samples to the number of samples tested as positive in the test

$$\text{Sensitivity} = (\text{TP}) / (\text{TP} + \text{FN}) \quad (6)$$

7. **Specificity:** Specificity is the ratio between total numbers of negative samples to the number of samples tested as negative in the test.

$$\text{Specificity} = (\text{TN}) / (\text{FP} + \text{TN}). \quad (7)$$

5. Result and discussion

In this section, the experimental results of our machine learning techniques with four class classification methodology using NSL-KDD intrusion detection dataset are provided in order to detect network intrusions and then comparison with the existing approaches is done to evaluate the efficacy of our network intrusion detection model. Confusion matrix is drawn for each type of attack and their faults of all four classifiers. So we obtain 16 set of confusion matrix and which are presented in this study. The performance of the classifiers is measured in terms of Confusion matrix,

Accuracy, Precision, Recall, F-measure, Specificity, Selectivity, G-mean. After the classifiers are trained the performance of all 4 classifiers are measured in terms of these metrics using test data set.

All the experiments are conducted using NSL-KDD dataset that has 125,973 training instances, 22,544 instances for testing with 41 attributes and 4 attack types for four classifiers to build an efficient network fault detection system. We have evaluated all algorithms with various evaluation measures, as discussed in the above section.

Confusion matrix for Random Forest:

Confusion matrix for major types of faults is shown in the **Figures 2–5**. For U2R attack the True negative is zero since the fault data is very low compared to the normal data, In R2L also number of fault data is very low, therefore the true negative value is low. In DoS and Probe attack also number of False positive data is more therefore the accuracy will be less. In the similar way the Confusion matrix for other classifiers are also constructed.

Tables 3–6 show the performance of the all 4 classifiers and from the result obtained we see the MLP classifier performs better than the other Classifier. False positive rate is less for MLP Classifier, True Positive and True Negative values are more. Therefore, MLP classifier is efficient classifier for fault detection Wireless Sensor Networks. The comparative plot of Accuracy for all 4 types of classifier algorithm is shown in the **Figure 6**, and it's evident that MLP on an average has an accuracy of 89.725%.

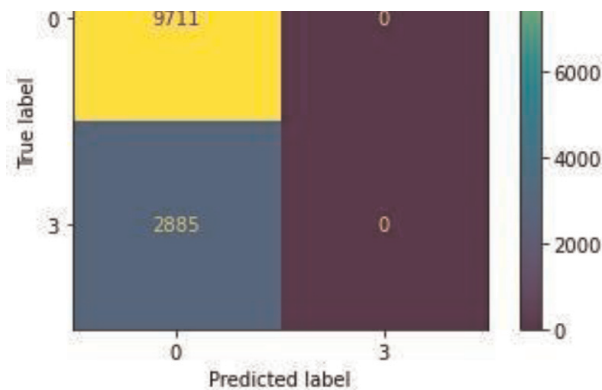


Figure 2.
 U2R attack for RF.

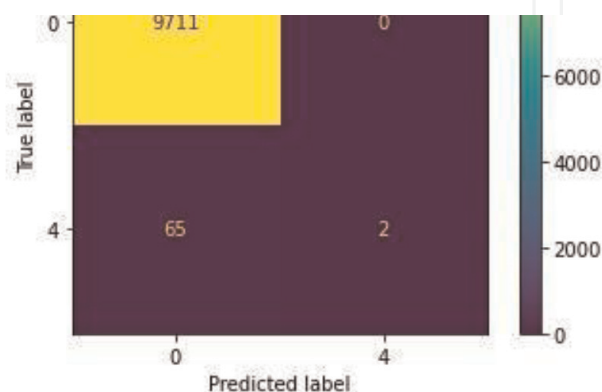


Figure 3.
 R2L attack for RF.

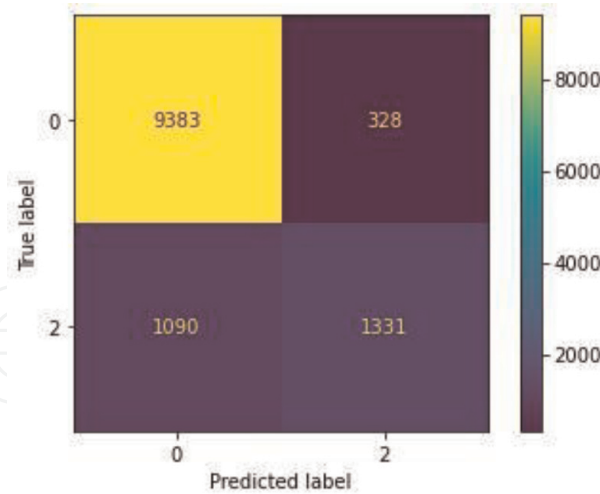


Figure 4.
Probe attack for RF.

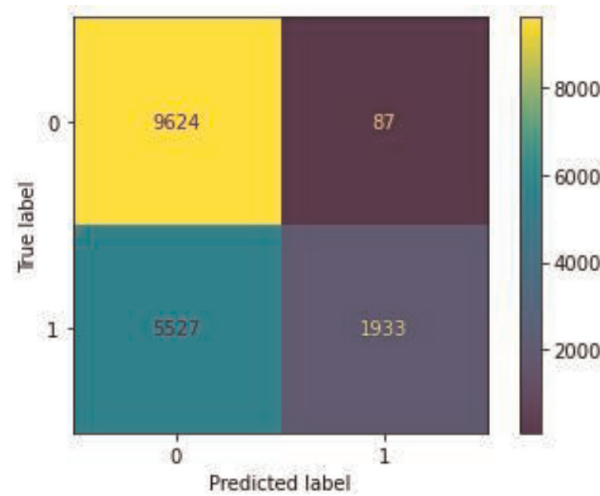


Figure 5.
DoS attack for RF.

The performance of the classifiers is also studied by introducing different Fault Probability Rates (FPR) and the results of the same are shown in **Tables 7–10**. The major goal of the fault percentage variation is that how accurately a classifier classifies the attack or normal data irrespective of the percentage of the fault present in particular test data. In the present study a classifier classifies the data with good amount of accuracy even if the percentage of fault is high.

Attack Types	Accuracy	Precision	Recall	F-measure	Specificity	Selectivity	G-mean
DoS	0.880	0.828	0.995	0.904	0.003	1.00	0.05
Probe	0.867	0.885	0.957	0.920	0.508	0.975	0.70
R2L	0.770	0.770	1.000	0.870	0.000	1.000	0.00
U2R	0.993	0.993	1.000	0.996	0.000	1.000	0.00

Table 3.
Accuracy for random Forest classifier.

Attack Types	Accuracy	Precision	Recall	F-measure	Specificity	Selectivity	G-mean
DoS	0.868	0.820	0.983	0.894	0.720	0.983	0.84
Probe	0.876	0.900	0.950	0.925	0.579	0.950	0.74
R2L	0.993	0.993	1.000	0.996	0.000	1.000	1.00
U2R	0.771	0.771	0.999	0.870	0.000	0.998	0.02

Table 4.
Accuracy for support vector machine classifier.

Attack Types	Accuracy	Precision	Recall	F-measure	Specificity	Selectivity	G-mean
DoS	0.886	0.842	0.982	0.907	0.782	0.981	0.87
Probe	0.920	0.935	0.935	0.935	0.591	0.969	0.75
R2L	0.993	0.994	0.998	0.996	0.000	0.999	0.00
U2R	0.790	0.790	0.998	0.880	0.265	0.999	0.51

Table 5.
Accuracy for MLP classifier.

Attack Types	Accuracy	Precision	Recall	F-measure	Specificity	Selectivity	G-mean
DoS	0.885	0.895	0.903	0.899	0.851	0.900	0.87
Probe	0.810	0.812	0.992	0.893	0.663	0.992	0.25
R2L	0.993	0.993	1.000	0.996	0.000	1.000	0.00
U2R	0.771	0.771	0.999	0.870	0.001	0.001	0.04

Table 6.
Accuracy for SGD classifier.

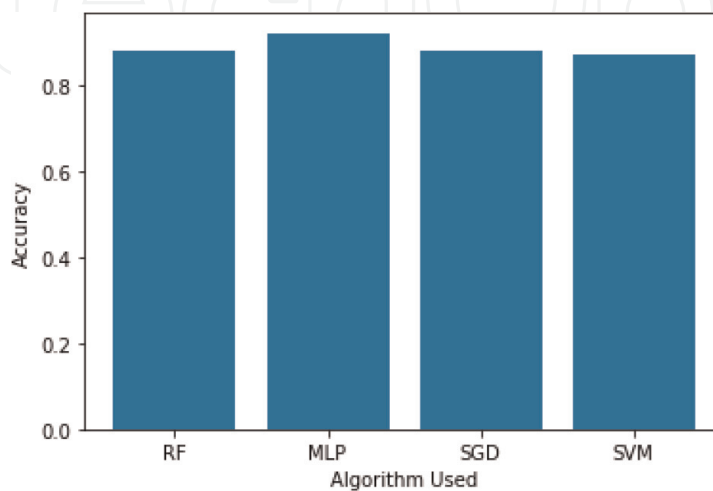


Figure 6.
Comparison of accuracy for all four classifiers.

Efficiency Parameter	FPR = 0.5	FPR = 0.4	FPR = 0.3	FPR = 0.2	FPR = 0.1	FPR = 0.05
Accuracy	0.880	0.750	0.834	0.932	0.873	0.978
Precision	0.828	0.823	0.830	0.964	0.994	0.991
Recall	0.995	0.969	0.996	0.962	0.876	0.998
F-measure	0.904	0.892	0.875	0.961	0.931	0.988
Specificity	0.993	0.968	0.962	0.997	0.930	0.993
Selectivity	0.011	0.523	0.684	0.652	0.781	0.669
G-mean	0.167	0.711	0.811	0.800	0.852	0.815

Table 7.
Performance of RF for varying fault rate.

Efficiency Parameter	FPR = 0.5	FPR = 0.4	FPR = 0.3	FPR = 0.2	FPR = 0.1	FPR = 0.05
Accuracy	0.882	0.926	0.940	0.936	0.926	0.914
Precision	0.906	0.976	0.992	0.979	0.997	0.999
Recall	0.953	0.946	0.942	0.935	0.926	0.937
F-measure	0.929	0.961	0.966	0.965	0.960	0.974
Specificity	0.999	0.999	0.979	0.949	0.824	0.819
Selectivity	0.999	0.993	0.973	0.953	0.884	0.821
G-mean	0.999	0.993	0.973	0.933	0.829	0.799

Table 8.
Performance of SVM for varying fault rate.

Efficiency Parameter	FPR = 0.5	FPR = 0.4	FPR = 0.3	FPR = 0.2	FPR = 0.1	FPR = 0.05
Accuracy	0.894	0.922	0.944	0.961	0.944	0.956
Precision	0.905	0.947	0.963	0.981	0.998	0.982
Recall	0.964	0.966	0.976	0.978	0.945	0.972
F-measure	0.936	0.956	0.969	0.979	0.970	0.977
Specificity	0.782	0.750	0.794	0.779	0.787	0.790
Selectivity	0.981	0.975	0.980	0.986	0.979	0.985
G-mean	0.876	0.855	0.882	0.877	0.878	0.882

Table 9.
Performance of MLP for varying fault rate.

6. Conclusions

The proposed system uses different Machine learning classifiers to recognize and categorize faults in Wireless Sensor networks. The dataset has four major classes, they are DoS, Probe, R2L, U2R which are further categorized. In this paper for the purpose of fault detection Random Forest (RF), Support Vector Machine (SVM), Stochastic

Efficiency Parameter	FPR = 0.5	FPR = 0.4	FPR = 0.3	FPR = 0.2	FPR = 0.1	FPR = 0.05
Accuracy	0.880	0.758	0.647	0.734	0.944	0.939
Precision	0.887	0.932	0.924	0.986	0.994	0.998
Recall	0.900	0.717	0.609	0.720	0.984	0.940
F-measure	.893	0.810	0.734	0.833	0.999	0.968
Specificity	0.851	0.864	0.634	0.661	0.757	0.358
Selectivity	0.900	0.717	0.963	0.950	0.945	0.940
G-mean	0.875	0.787	0.782	0.793	0.846	0.886

Table 10.
Performance of SGD for varying fault rate.


Gradient Descent (SGD), Multi-layer Perceptron (MLP) a classifiers are used to classify sensed data into faulty and non-faulty data Fault detection is a challenging task since wireless networks are placed in confined spaces. Machine learning classifiers are employed in this project because they are effective. The ML algorithms are trained using preprocessed data sets. One Hot Encoding is the method that is used to pre-process the data. Since in the data set few columns does not contain Numeric values. Recursive feature elimination is used to select the features that are applicable and which helps to find the specific attack. The system is put to the test on data set that were not seen during the training phase, some new attacks are introduced in the test data and the result show that the system is effective in identifying faults in the WSN. Since fault detection in the WSN can be challenging, due to harsh environment where the WSN are deployed makes them vulnerable to faults. Therefore machine learning is essential for fault detection since it is less time consuming, faster and also gives the good accuracy.

Author details

Poornima G. Miathali
Department of Electronics and Communication Engineering, BMS College of
Engineering, India

*Address all correspondence to: gpoornima.ece@bmsce.ac.in

IntechOpen

© 2023 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

References

- [1] Noshad Z, Javaid N, Saba T, Wadud Z, Saleem MQ, Alzahrani ME, et al. Fault detection in wireless sensor networks through the random Forest classifier. *Sensors*. 2019;**19**:1568. DOI: 10.3390/s19071568
- [2] Zidi S, Moulahi T, Alaya B. Fault detection in wireless sensor networks through SVM classifier. *IEEE Sensors Journal*. 2018;**18**(1):340-347. DOI: 10.1109/JSEN.2017.2771226
- [3] Windeatt T. Accuracy/diversity and ensemble MLP classifier design. *IEEE Transactions on Neural Networks and Learning Systems*. 2006;**17**(5):1194-1211. DOI: 10.1109/TNN.2006.875979
- [4] Elshoush HT, Dinar EA, Using Adaboost and Stochastic gradient descent (SGD) Algorithms with R and Orange Software for Filtering E-mail Spam. In: 2019 11th Computer Science and Electronic Engineering (CEEC). Colchester, UK: IEEE; 2019. pp. 41-46. DOI: 10.1109/CEEC47804.2019.8974319
- [5] Dong L et al. Very high-resolution remote sensing imagery classification using a fusion of random Forest and deep learning technique—Subtropical area for example. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*. 2020; **13**:113-128. DOI: 10.1109/JSTARS.2019.2953234
- [6] Jazzar M, Yousef RF, Eleyan D. Evaluation of machine learning techniques for email spam classification. *International Journal of Education and Management Engineering (IJEME)*. 2021;**11**(4):35-42. DOI: 10.5815/ijeme.2021.04.04
- [7] Panda M, Abraham A. Hybrid evolutionary algorithms for classification data mining. *Neural Computing and Applications*. 2015;**26**:507-523. DOI: 10.1007/s00521-014-1673-2
- [8] Poornima G, Suresh Babu K, Raja KB, Venugopal KR, Patnaik LM. Fault diagnosis approach for WSN using Normal bias technique. *ACEEE International Journal on Communication*. 2013;**4**(2):29-36
- [9] Zidi S, Moulahi T, Alaya B. Fault detection in wireless sensor networks through SVM classifier. *IEEE Sensors Journal*. 2017;**18**:340-347
- [10] Muhammed T, Shaikh RA. An analysis of fault detection strategies in wireless sensor networks. *Journal of Network and Computer Applications*. 2017;**78**:267-287
- [11] Miao X, Liu Y, Zhao H, Li C. Distributed online one-class support vector machine for anomaly detection over networks. *IEEE Transactions on Cybernetics*. 2018;**99**:1-14
- [12] Gharghan SK, Nordin R, Ismail M, Ali JA. Accurate wireless sensor localization technique based on hybrid PSO-ANN algorithm for indoor and outdoor track cycling. *IEEE Sensors Journal*. 2016;**16**:529-541
- [13] Swain RR, Khilar PM, Dash T. Neural network based automated detection of link failures in wireless sensor networks and extension to a study on the detection of disjoint nodes. *Journal of Ambient Intelligence and Humanized Computing*. 2018;**10**: 593-610
- [14] Yuan Y, Li S, Zhang X, Sun J. A Comparative Analysis of SVM, Naive Bayes and GBDT for Data Faults Detection in WSNs. In: 2018 IEEE

International Conference on Software
Quality, Reliability and Security
Companion (QRS-C). Lisbon, Portugal:
IEEE; 2018. pp. 394-399. DOI: 10.1109/
QRS-C.2018.00075

[15] Cheng Y, Liu Q, Wang J, Wan S,
Umer T. Distributed fault detection for
wireless sensor networks based on
support vector regression. *Wireless
Communications and Mobile
Computing*. 2018;**2018**:8. DOI: 10.1155/
2018/4349795

[16] Abdullah MA, Alsolami BM,
Alyahya HM, Alotibi MH. Intrusion
detection of DoS attacks in WSNs using
classification techniques. *Journal of
Fundamental and Applied Sciences*.
2018;**10**:298-303

[17] Zhang D, Qian L, Mao B, Huang C,
Huang B, Si Y. A data-driven Design for
Fault Detection of wind turbines using
random forests and XGboost. *IEEE
Access*. 2018;**6**:21020-21031

[18] Zhang X, Zou J, He K, Sun J.
Accelerating very deep convolutional
networks for classification and detection.
*IEEE Transactions on Pattern Analysis
and Machine Intelligence*. 2016;**38**:
1943-1955