Avrupa Bilim ve Teknoloji Dergisi
Sayı 31 (Ek Sayı 1), S. 104-113, Aralık 2021
© Telif hakkı EJOSAT'a aittir
Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 31 (Supp. 1), pp. 104-113, December 2021
Copyright © 2021 EJOSAT
Research Article

A Tree Based Machine Learning and Deep Learning Classification for Network Intrusion Detection

Şeyma Cihan[1*], Murat Aydos[2], Nihat Yılmaz Şimşek[3]

[1*] TUBITAK-Defense Industries Research and Development Institute-06261, Ankara, Turkey, (ORCID: 0000-0001-6267-2441), seyma.cihan@tubitak.gov.tr
[2] Department of Computer Engineering, Hacettepe University, Ankara, Turkey, (ORCID: 0000-0002-7570-9204), maydos@hacettepe.edu.tr
[3] Department of Computer Engineering, Hacettepe University, Ankara, Turkey, (ORCID: 0000-0003-0577-2766), nihatysimsek@gmail.com

(First received 2 March 2021 and in final form 16 December 2021)

(**DOI:** 10.31590/ejosat.889994)

ATIF/REFERENCE: Cihan, S., Aydos, M. & Simsek, N, Y. (2021). A Tree Based Machine Learning and Deep Learning Classification for Network Intrusion Detection. *European Journal of Science and Technology*, (31), 104-113.

Abstract

Parallel to the developments in network technology, the number of attacks on the network has increased significantly. The need for powerful intrusion detection systems to maintain network security and stability is increasing on a daily basis. This study proposes an intrusion detection system using traditional machine learning and deep learning algorithms. In this study, the NSL-KDD dataset has been classified using Random Forest, Decision Tree and Deep Neural Network algorithms. In addition, variable subsets were determined by using the Gini index and CFS (Corelation Based Feature Selection) to decrease dimension of the dataset. As a result of the study, the highest accuracy rate was 99.972%, and it was obtained from Random Forest algorithm applied on the dataset that was reduced to 11 variables by CFS method. In addition, 99.64% accuracy rate was obtained from Deep Neural Network without feature engineering.

Keywords: Intrusion detection system, Machine learning, Decision tree.

Ağ Saldırı Tespiti için Ağaç Temelli Makine Öğrenimi ve Derin Öğrenme Sınıflandırması

Öz

Ağ teknolojisindeki gelişmelere paralel olarak ağa yönelik saldırıların sayısı önemli ölçüde artmıştır. Ağ güvenliğini ve istikrarını korumak için güçlü izinsiz giriş tespit sistemlerine olan ihtiyaç her geçen gün artmaktadır. Bu çalışma, geleneksel makine öğrenimi ve derin öğrenme algoritmalarını kullanan bir saldırı tespit sistemi önermektedir. Bu çalışmada, NSL-KDD veri seti Random Forest, Decision Tree ve Deep Neural Network algoritmaları kullanılarak sınıflandırılmıştır. Ayrıca, veri kümesinin boyutunu azaltmak için Gini indeksi ve CFS (Korelasyona Dayalı Özellik Seçimi) kullanılarak değişken alt kümeleri belirlenmiştir. Çalışma sonucunda en yüksek doğruluk oranı %99.97 olarak CFS yöntemi ile 11 değişkene indirgenen veri kümesi üzerinde uygulanan Random Forest algoritması ile elde edilmiştir. Ayrıca özellik mühendisliği olmadan Deep Neural Network'ten %99,64 doğruluk oranı elde edilmiştir.

Anahtar Kelimeler: Saldırı tespit sistemi, Makine öğrenimi, Karar ağacı.

[*] Corresponding Author: seyma.cihan@tubitak.gov.tr

1. Introduction

In today's dizzying data flow traffic, protecting network technology against cyber attacks has become a vital necessity. NIDS (Network Intrusion Detection System) is the main security solutions that supports today's military applications, social systems, social networks, cloud services and other critical applications. NIDS automatically monitors network traffic to detect malicious activity and policy violations. Today, NIDS often uses signature-based detection techniques. However, the fact that signature databases need to be updated frequently in signature-based systems and can identify only known attacks motivates researchers to use approaches based on anomaly detection [1]. Systems based on anomaly detection primarily determine the normal behavior profile and can identify potential attacks in case of a significant deviation from normal [2,3]. Furthermore, NIDS based on anomaly detection is generally perceived as a stronger method in academic researches due to its theoretical potential to evaluate new attacks [4].

In systems based on anomaly detection, data mining and machine learning algorithms are often used to detect and prevent attacks [5,6]. In addition, effective NIDS are developed by using data mining methods as they are required to be able to identify and generalize attacks [7]. The developed software using machine learning algorithms can easily interpret complex and large amount of data related to network traffic to provide real-time detection, analysis and classification of attacks.

In classical machine learning methods, defining the features by experts with domain knowledge can reduce the complexity of patterns and make them more understandable and visible for algorithm. This process, requiring technical expert, is important and time-consuming. It also brings some problems like workload and high error rate because it is made by human [8]. Feature selection has vital importance for the performance of algorithm. In comparison of deep learning to other machine learning methods, there is no need feature selection process. Because the most relevant features are determined by deep learning model during the classification of dataset.In addition to feature selection problem mentioned above, classical machine learning approaches are not sufficient to solve a massive intrusion data classification problem which comes out a real network application environment [9]. In another study, it is presented that deep learning algorihm outperforms the other classical machine learning methods when they applied to high-dimensional data. Hence, it is shown importance of deep learning technology in high-dimensional data classification problems [10].

In this study, a new network intrusion detection technique based on Deep Learning is presented and then new method is compared to classicial machine learning algorithms including Random Forest and Decisin Tree. The two main contribution of the study proposed in this paper is:
i) comparison of classical machine learning algorithms and deep learning method on NSL-KDD dataset for a network intrusion detection system. The deep learning model was built based on extensive experiments for a large number of deep learning models to create most suitable and efficient model for NIDS.
ii) This study shows the effect of feature selection on classification of NSL-KDD dataset. Correlation based feature selection and Gini Index were applied to dataset for eliminating the irrelevant features and then these new dataset with the less features given as input data to created classification models.

1.1. Literature

In recent years, many of studies have been conducted in the literature to identify intrusion detection systems by using machine learning and deep learning approaches. Tavallaee et al. [4], in their study, developed a new set of training and test data, named NSL-KDD, to solve the problems of the KDD dataset that affect the classification performances of the algorithms. Researchers firstly extracted repetitive and invalid records in the training and test dataset. Then, the KDD training set was randomly divided into 3 subsets and each subset was trained three times with the J48 decision tree algorithm: Naive Bayes, NBTree, Random Forest, Random Tree, Multilayer Perceptron and Support Vector Machine (SVM) algorithms. The “successful prediction score” was kept for each record that was correctly labeled by machine learning algorithms and training and test datasets were grouped according to the scores obtained. According to these groups, the test datasets consisting of 125,973 records and 22,544 records were formed by a random selection. As a result, it was determined that machine learning algorithms showed better classification performance on newly created datasets without bias.

Olusola et al. [5], in their study, analyzed each variable in the KDD’99 dataset in terms of its impact on the classification performance and its distinctive power. In order to make these analyses, 10% KDD (kddcup_data_gz file) dataset was used. Two approaches were applied to assess the importance of the variable for the given attack class. Those were: calculating the rough set dependency rating for each class and calculating the dependency ratios of each class. As a result, the most relevant variables for each class were determined.

Horng et al. [2] developed an SVM-based classification system on the KDD-99 dataset. In the study, hierarchical clustering method was used in the preprocessing of the dataset. The accuracy of the developed system was 95.72% and false positive rate was 0.7%. In addition, the classification algorithm showed a better performance in DoS and Probe attacks than in similar studies. The system was also able to successfully identify the types of attacks that were not in the training dataset.

In Lin et al. studies [3], SVM (Support Vector Machine), DT (Decision Tree) and SA (Simulated Annealing) algorithms were applied on KDD dataset in order to develop intrusion detection sytem based on anomaly detection. In this study, to determine the best parameters for the DT and SVM algorithms, SA algorithm was also used. It was also used with SVM to determine the best subset of variables to improve classification performance. In order to evaluate the classification performance and obtain highest accuracy rates and decision rules, the k-fold method was used. Accuracy rate was determined as 99.96% on 23 variables in the developed classification system.

George [11] in his study, has made the anomaly detection using Principal Component Analysis (PCA) and Support Vector Machine (SVM) algorithms on KDD dataset. The SVM algorithm was applied on the original dataset and also the 28 variable dataset generated by the PCA algorithm, and the classification results were compared in terms of execution time, precision and recall. The SVM algorithm used with PCA algorithm, reduced execution time and achieved a higher classification accuracy.

In the study [12], the number of variables on the KDD99 dataset was reduced to 16 variables by PCA algorithm, then applied the Naive Bayes classification algorithm on the WEKA platform. In this study, the Naive Bayes classification algorithm was applied on the reduced dataset and the original dataset, and their performance was compared in terms of classification accuracy and execution times. In the same study, the False Positive ratio of the classification algorithm applied to the reduced dataset was found to be higher than the original dataset, while the classification time and memory requirement were significantly less.

Revathi and Malathi [13] studied on the NSL-KDD dataset in which Random Forest, J48, SVM, CART and Naive Bayes algorithms were used for classification. The researchers compared the results by applying the classification algorithms on the a dataset of 15 variables reduced with CFS method and the original dataset with 41 variables. The highest accuracy rates were obtained by Random Forest classification algorithm. In addition, the accuracy rates obtained from the classification on 15 variables were higher than the classification results for the original dataset for all attack types.

In their work, Siddiqui and Naahid [14] applied k-Means clustering algorithm on the 10% KDD dataset using the Oracle 10g Data Miner (ODM) as a data mining tool, and 1000 clusters were created on the training dataset containing 494.019 records. The Euclidean distance was used as a distance function when applying the clustering algorithm. As a result of clustering algorithm, it was determined that attacks were more in TCP protocol. The highest number of attacks to the TCP protocol were DoS attacks with a rate of 51%. At the same time, the most common attacks in all protocols were found to be DoS and PROBE.

Shrivas and Dewangan [15] applied CART, ANN, Bayes net, ANN and Bayes net, and CART and Bayes net classification algorithms on NSL-KDD and KDD-99 datasets in the first part of their work. Then, they applied the GR (Gain Ratio) feature selection algorithm to the best classifying algorithm. As a result of the classification, the best performance was obtained by the combined application of ANN and Bayes net algorithms. In the classification applied by decreasing the number of the features with GR on the 35 variables, the accuracy rate was 99.42% and in the NSL-KDD datasets, while on the 31 variables, the accuracy rate was 98.07%.

Al-Jarrah et al. [16], proposed two new feature selection methods for the NSL-KDD dataset, namely the RFFSR (RandomForest-Forward Selection Ranking) and the RF-BER (RandomForest-Backward Elimination Ranking). The features chosen by the proposed methods were compared with the three feature subset, which are well known in the literature of IDS and are selected by methods such as information gain, entropy, hybrid methods, and expert opinion. The performances of the generated subsets were compared by applying the RF classification algorithm, and experimental results showed that the features selected by the proposed methods improved detection and false positive rates by 99.8% and 0.001%, respectively.

Hasan et al. [17] applied Random Forest and SVM classification algorithms on the KDD99 dataset. Precision and false negative rate parameters were used in the study to evaluate the classification performances. The precision ratio of the RF classification algorithm was found to be higher by 80%, whereas the false negative rate of SVM algorithm was found to be lower by 31.69%.

In Dhanabal and Shantharajah [18], the J48, SVM, and Naïve Bayes classification algorithms were applied using the WEKA tool for the detection of attacks on the NSL-KDD dataset. Firstly, researchers reduced the number of features to 6 by using the correlation-based feature selection algorithm. The highest accuracy rate obtained from the J48 machine learning algorithm was 99.8%.

Özgür and Erdem [6] identified descriptive statistics by using the KDD99 dataset between the years 2010-2015 as mentioned in their literature review. As a result of the study, the most commonly used algorithms were SVM and decision tree derivatives, the most commonly used software tools were MATLAB and WEKA, and the detection rate was the most commonly used metric.

In Farnaaz and Jabbar [19] studies, the Symmetrical Uncertainty (SU) method was used to reduce the number of variables, and the RF and J48 classification algorithms were applied on the original and reduced NSL-KDD dataset. In the study, accuracy, detection rate, false alarm rate, and mathews correlation coefficient parameters were used for performance comparison. The experimental results showed that the RF classification algorithm had better performance in terms of comparison parameters. Additionaly, using the Symmetrical Uncertainty (SU) method increased the detection rate, decreased the false alarm rate.

Javaid et al. [20] presented a deep learning-based model to create an efficient Network Intrusion Detection System. They developed a deep learning model with self-taught learning algorithm and applied this model on NSL-KDD dataset. In this study, binary and multiclass classification methods were used on test dataset. As a result of performed classification with binary and multiclass classification types have revealed an accuracy rate of 88.39% and 79.10%, respectively.

In another study, Tang et al. [21] developed a deep neural network-based model for intrusion detection system and six major features are selected from the original NSL-KDD dataset for training of the developed model. Different learning rate values were used for optimization of the model and at the end of that study, learning rate 0.001 is found to be most successfully with an accuracy rate of 91.62% in terms of determined all metrics including precision, recall and f-measure.

Yin et al. [9] applied Recurrent Neural Networks algorithm with binary and multiclass classification for intrusion detection and compared results to classical machine learning algorithms including J48, SVM and Random Forests. RNN-IDS model achieved 97.09% accuracy on test dataset compared to applied classification algorithms.

KDD Cup ’99 and NSL-KDD datasets were given as test dataset to developed Non-Symmetric Deep Auto-Encoder model on GPU based Tensorflow by Shone et al. [8] for creating an intrusion detection system. As a result of performed classification, proposed model have achieved an accuracy rate of 97.85% on KDD Cup ’99 dataset and 85.42% on NSL-KDD dataset.

Aljawarneh et al. [7] proposed a hybrid classification model on the NSL-KDDTrain + 20% dataset in their study. First, using the Information Gain (IG) method, they reduced the variable number to 8 variables by selecting variables with an IG score above 0.40. Then, J48, Meta Pagging, Random Tree, REPTree, Ada BoostM1, Decision Stump, and Naive Bayes algorithms were applied to the selected variable subset, and the attack classification was performed. The performance of the developed model was compared with the performance of the J48, SVM and Naive Bayes classification algorithms. As a result, the proposed hybrid model showed the best performance with 96.2% to 99.9% accuracy in determining all attack classes.

In the Biswas’ [22] study, on the NSL-KDD dataset, subsets of the variables were determined by the Correlation Based Feature Selection (CFS), the Principal Component Analysis (PCA), the Information Gain Ratio Feature Selection (IGR) and the Minimum Redundancy Maximum Relevance methods. On the dataset consisting of selected variables, Naive Bayes, Support Vector Machine, Decision Tree, Neural Network, and k-Nearest Neighbor classification algorithms were applied by using WEKA tool. According to the results of the experiment, the highest accuracy rate of 99.07% was reached using the K-NN classification algorithm applied after decreasing dimension of the the dataset with the IGR method.

Özgür and Erdem [23] proposed a model called GA-NS-AB (Genetic Algorithm Based Feature Selection and Weighting). The developed model was implemented on the NSL- KDD dataset. In their study, classifier fusion was made with Adaboost, Decision Tree, Logistic Regression, Naive Bayes, Random Forest, Gradient Boosting, K Nearest Neighbor and Artificial Neural Networks. ANOVA statistical test was used to compare the fusion classifier results. Compared to other studies in the literature, the GA-NS-AB model (4 classifier fusion) was found to have a better performance with an accuracy of 90.75%. According to the results of the experiments, it was determined that weighted fusion classifications using 3 and 4 classifiers were sufficient.

Gurung et al. [24] introduced a sparce auto-encoder deep neural network approach with logistic regression. They used NSL-KDD dataset as input and binary classification as the output of model. Confusion matrix was used as evaluation metric of the classification results. In the presented study, it was found 87.2% accuracy rate. In this study, Random Forest and Decision Tree algorithms, which are widely used in machine learning domain, are selected for the purpose of classification on network traffic data to determine intrusion. Additionally, Gini index and CFS methods is also used to reduce the dimension of the dataset, which has critical importance for high volume traffic data. To the best of our knowledge, this is the first study that uses these two methods together for reducing dimension in NSL-KDD dataset. Also, the classification performances obtained by the machine learning algorithms are presented comparatively in terms of determined parameters.

In Addition to Random Forest and Decision Tree algorithms, our proposed deep learning algorithm implemented to dataset for the intrusion detection system. These algorithms are analyzed for the purpose of comparison for their performances and accuracies and it is expected to show deep learning algorithm also has an important potential for creating intrusion detection system.

2. Material and Method

2.1. Dataset Description

In this study, the NSL-KDD dataset was used. The NSL-KDD dataset was created by deleting repetitive and redundant records and reducing the data size on a KDD 99 dataset. Within the scope of the study, classification algorithm was applied on a 20% of the training dataset instead of the whole of the dataset due to execution cost. The dataset consists of 25,192 records and 41 variables.

There are 21 different types of attack in the NSL-KDD training dataset. These types of attacks are grouped into four different categories. Table 1 shows the types of attacks and the classes that they belong to [18, 25]. The attack classes are classified as Probing, Denial of Service (DoS), Remote to Local (R2L) and User to Root (U2R) attacks [26].

Table 1. Attack Class and Type Matching

Dataset Attack Class	Attack Type
DoS	Back,Land,Neptune,Pod, Smurf, Teardrop, Apache2, Udpstorm, Processtable, Worm, Mailbomb
PROBE	Satan,Ipsweep,Nmap, Portsweep, Mscan, Saint
R2L	Guess_Password, Ftp_write, Imap, Phf, Multihop, Warezmaster, Warezclient, Spy, Xlock, Xsnoop, Snmpguess, Snmpgetattack, Httptunnel, Sendmail, Named
U2R	Buffer_overflow, Loadmodule, Rootkit, Perl, Sqlattack, Xterm, Ps

Probe Attacks: They aim to obtain information about the target network from an external source network. Therefore, the basic connection level properties such as “duration of connection” and “source bytes” are significant when detecting the probes. However, it is not expected to provide information such as “number of files creations” and “number of files accessed”.

DoS Attacks: DoS attacks prevent the services provided by the target through illegal requests. For this reason, “percentage of connections having same destination host and same service”, and packet level features such as “source bytes” and “percentage of packets with errors” are important traffic features.

R2L Attacks: R2L attacks are one of the most difficult attacks to detect. Includes network level and host level features. Therefore, to determine the R2L attacks, network-level properties such as "duration of connection" and "service requested", as well as host-level properties such as "number of failed login attempts" are required.

U2R Attacks: U2R attacks contain semantic details that are very difficult to catch at an early stage. Such attacks are often content-based and target an application. Therefore, features such as "number of file creations" and "number of shell prompts invoked" are relevant; however, features such as "protocol" and "source bytes" are ignored.

2.2. Feature Selection

The variables in the dataset have a key role in the performance of machine learning algorithms. The variables in the NSL-KDD dataset are grouped into four main groups: TCP connection basics, TCP connection content properties, time-based network traffic and host-based network traffic. Some of the features included in the dataset are not only important in training of machine learning algorithms, but also have a role in improving the detection rate [27]. However, during the construction of the machine learning model, the use of all features of the dataset is not effective in terms of processing time and cost. Therefore, it is important to reduce the dataset dimension by identifying the relevant features in establishing robust learning models.

In machine learning applications, determining the importance of a variable that is a result of complex interactions with other variables can be a difficult problem for researchers. In order to decrease the dimension of the dataset, the correlation-based feature selection method CFS (Corelation based feature selection) and Gini index are used.

Important variables according to CFS method are having a high correlation with the target or class variable, but a low correlation with other variables in the dataset [22]. CFS method, which is also one of the filter based feature selection, is implemented by using WEKA tool. In CFS method, in addition to intercorrelation between the variables, it also predicts the correlation between a subset of features and a class variable. CFS can be computed by using Equation 1, where Cs is the correlation between summed variable subsets and the target class variable, Sn is the number of subset variables, Rci is the average correlation between variables and target class variable, and Rii is the average intercorrelation between variables [28].

$$Cs = \frac{SnRci}{\sqrt{Sn + Sn(Sn-1)Rii}} \qquad (1)$$

In addition to the CFS method, significant variables were determined for the classification by calculating the Gini index (Mean Decrease Gini) [29]. The Random Forest algorithm offers two different methods, which can be used for feature selection or ranking. These methods are Mean Decrease Accuracy (MDA) and Mean Decrease Gini (MDG). MDA determines the importance of a variable on out of bag observations that are not used for growing tree by measuring the change in classification accuracy when the variable values are randomly permuted as compared to the initial observed values. MDG is obtained by the sum of all decreases in Gini impurity for a given variable after each split [30, 31]. Although Random Forest algorithm provides two different feature selection method, MDG and the Gini index provide more robust subset results as compared to MDA [31]. Therefore, MDG is used for reducing dataset dimension to construct classification model in this study. The Gini index is calculated for all variables in the dataset. Each tree in the forest is used for the calculation of Gini importance. This value gives the value of Gini for any m variable [32]. In Equation 2:

$$GI(t) = 1 - \sum_{k} p(k|t)2 \qquad (2)$$

where GI(t) shows the Gini index and p(k|t) shows the rate at which the class k can be separated correctly in the t node.
In Equation 3:

$$\Delta GI(t)=P_t\, GI(t)\text{- } P_L GI(t_L)\text{- } P_R\, GI(t_R) \qquad (3)$$

where *$\Delta GI(t)$*: the Gini difference; *$P_L GI(t_L)$:* the Gini index on the left side of the node; *$P_R\ GI(t_R)$:* the Gini index on the right side of the node; *P_t*: the number of instances before the division; *P_L*: the number of samples on the left side after the division; *P_R*: the number of samples on the right side after the division [33].

2.3. Classification Models

There are different machine learning algorithms that have been acknowledged in the literature in order to develop attack detection systems.

In this study, Random Forest algorithm, which is a ensemble learning algorithm, has been selected because of its advantages for establishing the classification model. The Random Forest algorithm performs well in most problem areas, provides good results on both numeric and categorical data as well as for noisy or missing datasets, and can be applied on dataset containing a large number of features. It is powerful against overfitting and does not require pruning in the tree after the model constructing process [32, 34].

In order to compare the performance of the Random Forest algorithm with the results obtained from a single decision tree, the J48 Decision Tree algorithm - one that also allows the researchers to interpret the tree results - is used. The classification model and applications arew developed through the R program on RStudio and WEKA platform.

In this study, Deep Neural Network (DNN) is proposed as another classification method. DNN is in fact an artificial neural network (ANN) with several hidden layers of units across the input and output layers [35]. DNN can also get model of

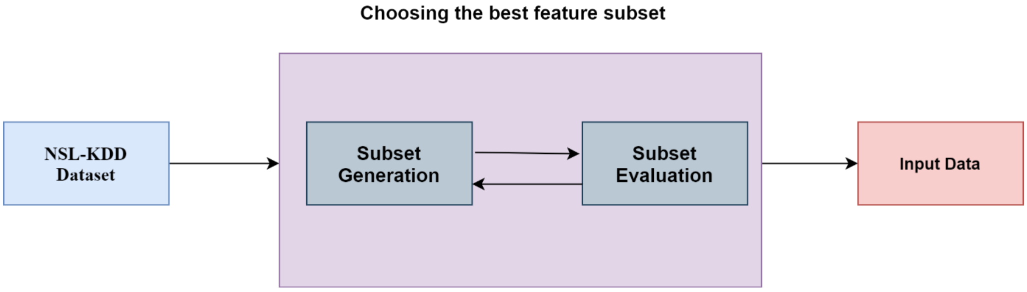

Figure 1. General structure of feature selection

complex non-linear relationships like ANN, but DNN have the extra layers which allows feature combinations from lower layers. Hence, DNN have more capability to create models for complex data with less units than networks designed similarly [36]. DNNs are generally aimed to function as feed-forward networks and they can be discriminatively trained with the standard back-propagation algorithm. Stochastic Gradient Descent is used to update weights with the following equation (4):

$$w_{ij}(t+1) = w_{ij}(t) + \mu \frac{\partial C}{\partial wij} \qquad (4)$$

where, μ denotes the learning rate and C represents the cost function. The selection of the cost function is dependent on parameters like the learning model (supervised, unsupervised etc.) and the activation function. For instance, given that supervised learning is applied on a multiclass classification problem, softmax function can be chosen as the activation function and cross entropy function can be used as cost function. Mathematically, the softmax function can be expressed with the following equation (5):

$$P_j = \frac{exp(x_j)}{\sum_k exp(x_k)} \qquad (5)$$

where, P_j represents the probability of class (output of the unit j) and x_j and x_k represent the total input to units j and k respectively of the same level. Cross entropy (cost function in a supervised learning on multiclass classification problems) is obtained with the equation (6):

$$C_r = \sum_j d_j \, log(P_j) \qquad (6)$$

where d_j represents the target probability for output unit j and P_j is the probability output for j after applying the activation function [37].

In this study, H2O cluster library of R program used for implementation of proposed deep neural network approach on NSL-KDD dataset. H2O cluster library provides an efficient framework for usage of large datasets including network intrusion data in deep learning algorithm.

2.4. Model Evaluation

The Confusion Matrix is an important tool for evaluating the performance of the applied classification models. Table 2 shows the components of the confusion matrix. The parameters used to evaluate the classification performance are obtained from the confusion matrix. The confusion matrix components used in the attack detection classification are defined as follows:

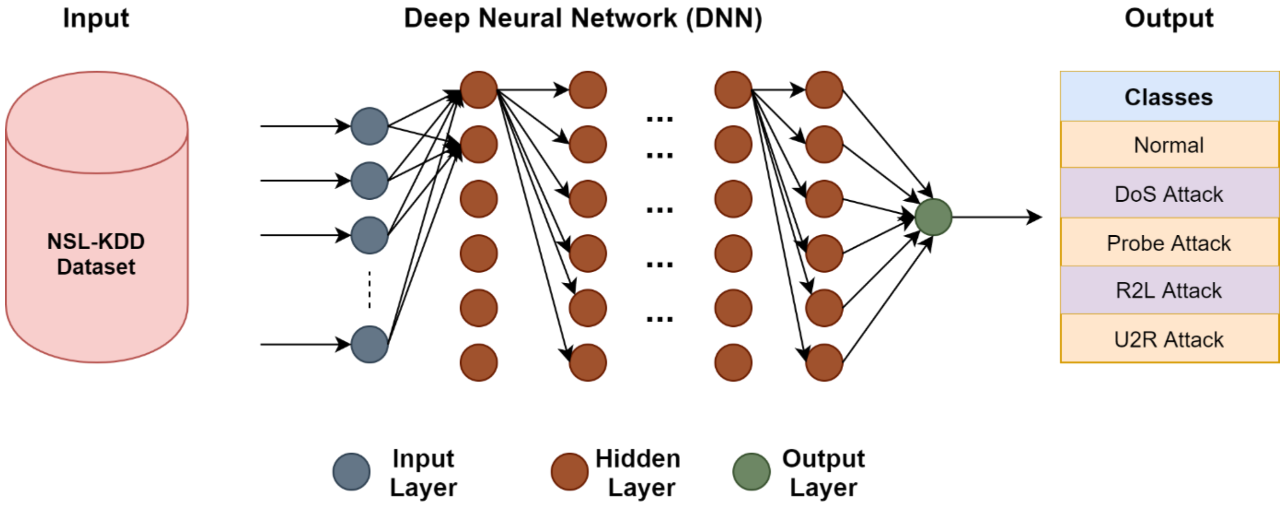

Figure 2. Proposed deep neural network model for NIDS classification

True-Positive (TP): Classification of attacks as attacks
True-Negative (TN): Classification of non-attacks as non-attacks
False-Positive (FP): Classification of normal ones as attacks
FalseNegative (FN): Classification of attacks as normal

Table 2. Confusion Matrix

	Predicted Values		
		No Attack	Attack
Actual values	No Attack	TN	FP
	Attack	FN	TP

To evaluate the classification model performance; Accuracy, TP Rate, FP Rate, Precision, Recall, F-measure, ROC area and Time parameters are used.

3. Results and Discussion

3.1. Experimental Results

In the study, the pre-analysis of the dataset was performed by R program both graphically and statistically before the classification model was established. Figure 3 shows the main classes of the attacks and counts in the NSL-KDD dataset.

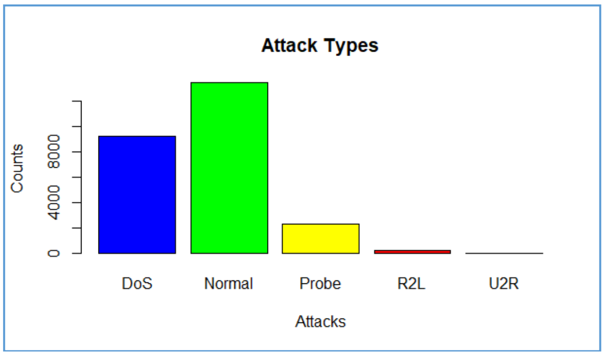

Figure 3. Attack Types in NSL-KDD Dataset

From this, it was seen that the datset has DoS attack types the most. This was followed by Probe attacks. In addition, when the attacks were examined according to the protocol type, it was determined that the maximum number of DoS attacks were in the TCP protocol, whereas the maximum of the Probe attacks were in the ICMP and UDP protocols (Figure 4).

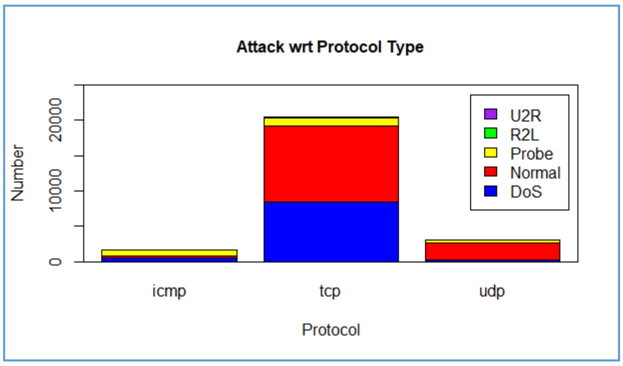

Figure 4. Attacks wrt Protocol Type

To constructing a classification model, J48 Decision Tree and Random Forest algorithms was applied on the original dataset consisting of 41 variables. Then, significant variables were determined by Mean Decrease Gini (Gini Index) and CFS methods. Table 3 shows the selected subset of features.

Table 3. Selected features

Feature Selection Method	Selected Features
Gini Index	service, protocol_type, flag, src_bytes, dst_bytes, count, srv_count, serror_rate, same_srv_rate, diff_srv_rate, dst_host_srv_count, dst_host_srv_serror_rate, dst_host_diff_srv_rate , dst_host_same_src_port_rate, dst_host_serror_rate
CFS	service, flag, src_bytes, dst_bytes, logged_in, root_shell, srv_serror_rate, same_srv_rate, diff_srv_rate, dst_host_srv_diff_host_rate, dst_host_serror_rate

Figure 5 shows the 15 most important variables by Gini index. According to Gini index, the variables src_bytes, same_srv_rate, and flag were determined as the three most important variables for the classification model.

Furthermore, importance rank of features by deep learning model is shown Figure 6. According to the results of the deep learning model, num_compromised, src_bytes and srv_count are the most relevant features in dataset and src_bytes is the common variable in both approaches.

After feature selection process, classification algorithms have been applied on NSL-KDD dataset. Firstly, classical machine learning algorithms have been applied for comparison purpose and then the results were obtained with deep learning model. Used classical machine learning algorithms are J48 DT and RF. Afterwards, deep learning models have been applied on features selected dataset.

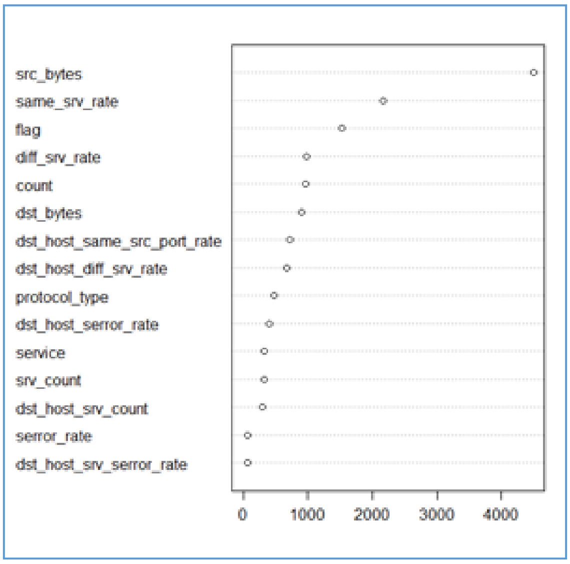

Figure 5. Mean Decrease Gini Result

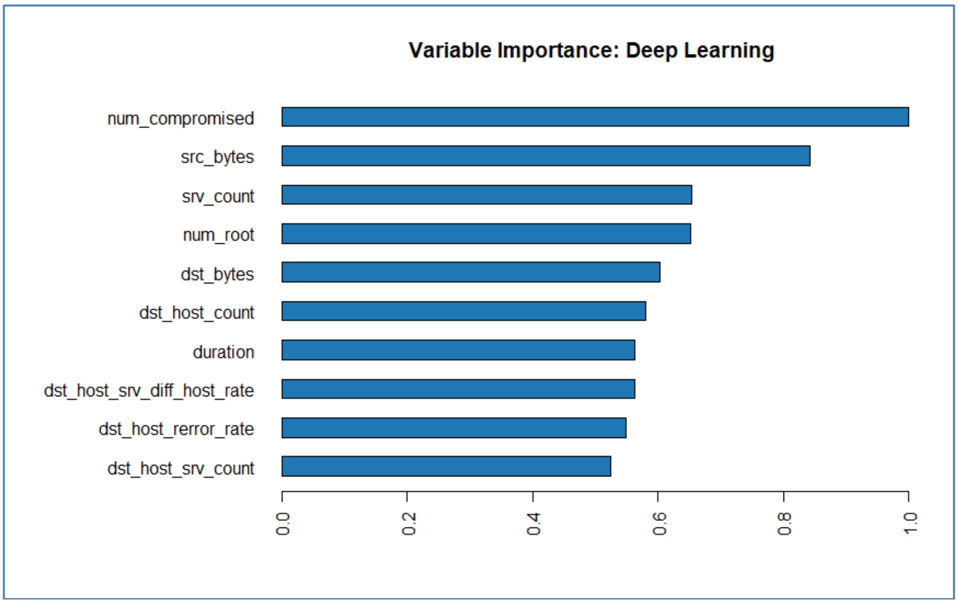

Figure 6. Feature Importance rank from Deep Learning Model

Table 4. Comparison of the Classical Machine Learning Algorithms.

Algorithm	**Accuracy (%)**	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-measure**	**ROC area**	**Time (second)**
J48 on original dataset	99.491	0.995	0.004	0.995	0.995	0.995	0.997	2.64
J48 on CFS subset	99.484	0,995	0,004	0,995	0,995	0,995	0,998	0.44
J48 on Gini subset	99.428	0.994	0.004	0.994	0.994	0.994	0.997	0.81
RF on original dataset	99.753	0.998	0.002	0.997	0.998	0.997	1.000	9.92
RF on CFS subset	***99.972***	1.000	0.000	1.000	1.000	1.000	1.000	***1.34***
RF on Gini subset	99.734	0.997	0.002	0.997	0.997	0.997	1.000	5.01

Firstly, developed J48 DT and RF algorithms have been applied to the NSL-KDD dataset. Then the models have been applied on the train set and then tested. After classification, TP rate, FP rate, precision, recall, F-measure and ROC area have been calculated. Another parameter Time was recorded in seconds during the classification for using in comparison. Comparisons of the results of classical algorithms are tabulated in Table 4.

After classical machine learning algorithms, proposed deep learning algorihm has been applied on NSL-KDD dataset. The results obtained from developed model is presented in Table 5.

Comparison of developed deep learning model and classical machine learning algorithms results for NSL-KDD dataset are presented in Table 6. Accuracy is used as parameter for comparison and the original dataset accuracy values are selected for J48 DT and RF from Table 4.

Table 5. Result of DNN model

Classification Model	**Train Accuracy (%)**	**Test Accuracy (%)**
Deep Learning Model	99.71%	99.64%

Table 6. Comparison of developed deep learning model and classical machine learning algorithms results for original NSL-KDD dataset

Algorithm	**Accuracy (%)**
J48 on original dataset	99.49
RF on original dataset	99.75
Deep learning model on original dataset	99.64

3.2. Discussion

The results are presented as an average for each attack class. When the algorithm performance was examined, the highest accuracy rate was obtained from the Random Forest algorithm applied on subset selected by CFS method with 99.972% accuracy. The second highest result obtained was by the Random Forest algorithm being applied on the original dataset, with an accuracy of 99.753%. The lowest accuracy rate was 99.428% obtained from the J48 Decision Tree algorithm being applied on 15 variables, which decreased with Gini index. The results of other determined comparison parameters are quite similar. However, in terms of processing time, the Random Forest algorithm applied on the CFS dataset was completed in a much shorter time- in 1.34 seconds- than the other Random Forest classification models.

According to the deep learning model classification results, it is found that model gives the accuracy of 99.71% and 99.64% values for training and test datasets, respectively. The negligible difference between train and test classification result is important for excluding overfitting. It is shown that, deep learning algorithm gives high accuracy rate even though any feature selection method wasn't applied on dataset. The most important reason for this, deep learning model select the most relevant features by changing the weights of variables during the traning. In comparison of deep learning and classical machine learning algorithms, it can be considered as deep learning algorithm can be more efficient in big data problems with high interrelated and complex features. One of the future works can be done is using GPU acceleration to reduce the training time of the developed deep learning model and working for a deep learning based real-time NIDS system.

Based on the results, it was determined that Random Forest algorithm applied on variable subset determined by CFS method performed the best. Also, the classification algorithms applied over the reduced number of features resulted in much shorter processing time. The classification of Decision Trees on only one sample limits the reliability of the constructed model. On the other hand, Random Forest algorithm is used to evaluate the results of many decision trees, therefore, it is thought that more reliable prediction can be made with Random Forest algorithm. In addition, using the Random Forest predictions of all decision trees provided a better generalization. However, it was found that the processing times for both classification algorithms applied on the reduced number of variables were quite short. For this reason, it is important to work with a small number of variables- especially when considering the time complexity- in high dimension datasets. Hence, it can be concluded that RF is the best classification algorithm among the classification algorithms including J48 DT and DNN for NSL-KDD dataset in this study.

4. Conclusions and Recommendations

Network intrusion detection systems has a vital importance today because of huge data flow traffic. One of the most important study of today is preventing the cyber attacks. NIDS (Network Intrusion Detection System) is the main security solutions that supports today's military applications, social systems, social networks, cloud services and other critical applications. NIDS automatically monitors network traffic to detect malicious activity and policy violations. The machine learning algorithms can be used for NIDS.

In this study, deep learning algorithm and classical machine learning algorithms' performances are compared for classification of NSL-KDD dataset. The classification model was applied on the original dataset consisting of 41 variables and then on the decreased variables by the Gini index and CFS methods, and the algorithm performances were compared in terms of determined parameters. It is shown that, the Random Forrest algorithm is more successful in general than compared algorithms including J48 DT and deep learning. It is considered that attack information related to network traffic that were gathered during the preprocessing of the dataset - such as the most common types of attacks, attack types according to the protocols - are also very important factors to consider while guiding the measures taken by network administrators against intrusion. Furthermore, important features determined by the feature selection methods in this study can provide information to the network administrators about critical variables that they need to monitor for preventing and detecting attacks.

References

[1] Yan, J., Jin, D., Lee, C. W., & Liu, P. (2018, July). A Comparative Study of Off-Line Deep Learning Based Network Intrusion Detection. In 2018 Tenth International Conference on Ubiquitous and Future Networks (ICUFN) (pp. 299-304).

[2] Horng, S. J., Su, M. Y., Chen, Y. H., Kao, T. W., Chen, R. J., Lai, J. L., & Perkasa, C. D. (2011). A novel intrusion detection system based on hierarchical clustering and support vector machines. Expert systems with Applications, 38(1), 306-313.

[3] Lin, S. W., Ying, K. C., Lee, C. Y., & Lee, Z. J. (2012). An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection. Applied Soft Computing, 12(10), 3285-3290.

[4] Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009, July). A detailed analysis of the KDD CUP 99 dataset. In Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on (pp. 1-6).

[5] Olusola, A. A., Oladele, A. S., & Abosede, D. O. (2010, October). Analysis of KDD'99 intrusion detection dataset for selection of relevance features. In Proceedings of the World Congress on Engineering and Computer Science (Vol. 1, pp. 20-22).

[6] Özgür, A., & Erdem, H. (2016). A review of KDD99 dataset usage in intrusion detection and machine learning between 2010 and 2015. PeerJ PrePrints, 4, e1954v1.

[7] Aljawarneh, S., Aldwairi, M., & Yassein, M. B. (2018). Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. Journal of Computational Science, 25, 152-160.

[8] Shone, N., Ngoc, T. N., Phai, V. D., & Shi, Q. (2018). A deep learning approach to network intrusion detection. IEEE Transactions on Emerging Topics in Computational Intelligence, 2(1), 41-50.

[9] Yin, C., Zhu, Y., Fei, J., & He, X. (2017). A deep learning approach for intrusion detection using recurrent neural networks. Ieee Access, 5, 21954-21961.

[10] Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H.,.. & Wang, C. (2018). Machine Learning and Deep Learning Methods for Cybersecurity. IEEE Access, 6, 35365-35381.

[11] George, A. (2012). Anomaly detection based on machine learning: dimensionality reduction using PCA and classification using SVM. International Journal of Computer Applications, 47(21).

[12] Neethu, B. (2012). Classification of intrusion detection dataset using machine learning approaches. International Journal of Electronics and Computer Science Engineering, 1(3), 1044-1051.

[13] Revathi, S., & Malathi, A. (2013). A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection. International Journal of Engineering Research and Technology. ESRSA Publications.

[14] Siddiqui, M. K., & Naahid, S. (2013). Analysis of KDD CUP 99 dataset using clustering based data mining. International Journal of Database Theory and Application, 6(5), 23-34.

[15] Shrivas, A. K., & Dewangan, A. K. (2014). An ensemble model for classification of attacks with feature selection based on KDD99 and NSL-KDD dataset. International Journal of Computer Applications, 99(15), 8-13.

[16] Al-Jarrah, O. Y., Siddiqui, A., Elsalamouny, M., Yoo, P. D., Muhaidat, S., & Kim, K. (2014, June). Machine-learning-based feature selection techniques for large-scale network intrusion detection. In Distributed Computing Systems Workshops (ICDCSW), 2014 IEEE 34th International Conference on (pp. 177-181).

[17] Hasan, M. A. M., Nasser, M., Pal, B., & Ahmad, S. (2014). Support vector machine and random forest modeling for intrusion detection system (IDS). Journal of Intelligent Learning Systems and Applications, 6(01), 45.

[18] Dhanabal, L., & Shantharajah, S. P. (2015). A study on NSL-KDD dataset for intrusion detection system based on classification algorithms. International Journal of Advanced Research in Computer and Communication Engineering, 4(6), 446-452.

[19] Farnaaz, N., & Jabbar, M. A. (2016). Random forest modeling for network intrusion detection system. Procedia Computer Science, 89, 213-217.

[20] Javaid, A., Niyaz, Q., Sun, W., & Alam, M. (2016, May). A deep learning approach for network intrusion detection system. In Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS) (pp. 21-26).

[21] Tang, T. A., Mhamdi, L., McLernon, D., Zaidi, S. A. R., & Ghogho, M. (2016, October). Deep learning approach for network intrusion detection in software defined networking. In 2016 International Conference on Wireless Networks and Mobile Communications (WINCOM) (pp. 258-263).

[22] Biswas, S. K. (2018). Intrusion Detection Using Machine Learning: A Comparison Study, International Journal of Pure and Applied Mathematics, 118(19), 101-114.

[23] Özgür, A., & Erdem, H. (2018). Saldırı tespit sistemlerinde genetik algoritma kullanarak nitelik seçimi ve çoklu sınıflandırıcı füzyonu. Gazi Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi, 33(1).

[24] Gurung, S., Ghose, M. K., & Subedi, A. (2019). Deep learning approach on network intrusion detection system using NSL-KDD dataset. International Journal of Computer Network and Information Security (IJCNIS), 11(3), 8-14.

[25] Kumar, V., Chauhan, H., & Panwar, D. (2013). K-means clustering approach to analyze NSL-KDD intrusion detection dataset. International Journal of Soft Computing and Engineering (IJSCE).

[26] Kaushik, S. S., & Deshmukh, P. R. (2011). Detection of attacks in an intrusion detection system. International Journal of Computer Science and Information Technologies (IJCSIT), 2(3), 982-986.

[27] Meng, Y. X. (2011, July). The practice on using machine learning for network anomaly intrusion detection. In Machine Learning and Cybernetics (ICMLC), 2011 International Conference on(Vol. 2, pp. 576-581).

[28] Pushpalatha, K. R., & Karegowda, A. G. (2017, December). CFS Based Feature Subset Selection for Enhancing Classification of Similar Looking Food Grains-A Filter Approach. In 2017 2nd International Conference On Emerging Computation and Information Technologies (ICECIT) (pp. 1-6).

[29] Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

[30] Archer, K. J., & Kimes, R. V. (2008). Empirical characterization of random forest variable importance measures. Computational Statistics & Data Analysis, 52(4), 2249-2260.

[31] Calle, M. L., & Urrea, V. (2010). Letter to the editor: stability of random forest importance measures. Briefings in bioinformatics, 12(1), 86-89.

[32] Akman, M., Genç, Y., & Ankarali, H. (2011). Random forests yöntemi ve sağlık alanında bir uygulama. Turkiye Klinikleri Journal of Biostatistics, 3(1), 36-48.

[33] Kawakubo, H., & Yoshida, H. (2012). Rapid feature selection based on random forests for high-dimensional data. Expert Syst Appl, 40, 6241-6252.

[34] Lantz, B. Machine Learning With R. Packt Publishing Ltd, Birmingham, 2013.

[35] Deng, Li & Yu, Dong. (2013). Deep Learning: Methods and Applications. Foundations and Trends in Signal Processing. 7. 10.1561/2000000039.

[36] Bengio, Y.. (2009). Learning Deep Architectures for AI. Foundations. 2. 1-55. 10.1561/2200000006.

[37] Hinton, Geoffrey & Deng, li & Yu, Dong & Dahl, George & Mohamed, Abdel-rahman & Jaitly, Navdeep & Senior, Andrew & Vanhoucke, Vincent & Nguyen, Phuongtrang & Sainath, Tara & Kingsbury, Brian. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. Signal Processing Magazine, IEEE. 29. 82-97. 10.1109/MSP.2012.2205597.