






Düzce University Journal of Science & Technology

Research Article

Performance Analysis of Machine Learning Algorithms in Intrusion Detection Systems¹

 Fethi Mustafa ÇİMEN^{a,*},  Yusuf SÖNMEZ^b,  Mustafa İLBAŞ^c

^a Department of Energy Systems Engineering, Faculty of Technology, Gazi University, TURKEY

^b Department of Computer Technologies and Cyber Security, Faculty of Information and Telecommunication Technologies, Azerbaijan Technical University, AZERBAIJAN

^c Department of Energy Systems Engineering, Faculty of Technology, Gazi University, TURKEY

* Corresponding author's e-mail address: fethimustafacimen@gazi.edu.tr

DOI:10.29130/dubited.1018229

ABSTRACT

With the developing technology, the need for the dissemination and protection of information is becoming increasingly important. Recently, attacks on information systems have increased significantly. In addition to the rise in the number of attacks, attacks of different types pose a great threat to systems. As a result of these attacks, institutions and users suffer serious damages. At this point, Intrusion Detection Systems (IDS) have a very important position. The pre-detection of these attacks on the systems and the preparation of the necessary reports can reduce the impact of the threats that may be encountered in the future. Recent studies are carried out so as to increase the performance of IDS. In this paper, classification was made using NSL-KDD dataset and SVM, KNN, Bayesnet, NavieBayes, J48 and Random Forest algorithms, and it was aimed to compare performance of these classifications by using WEKA. Consequently, it has been reached that the KNN algorithm had the best performance with an accuracy rate of 98.1237 %. In addition, the effect of increasing the number of folds and neighborhoods on the classification result has been examined comparatively.

Keywords: Machine learning, IDS, KNN

Saldırı Tespit Sistemlerinde Makine Öğrenimi Algoritmalarının Performans Analizi

ÖZ

Gelişen teknoloji ile bilginin yayılması ve korunması gereksinimi giderek önem kazanmaktadır. Son dönemde bilişim sistemlerine yönelik saldırılar önemli düzeyde artış göstermiştir. Saldırı sayısındaki artışın yanı sıra farklı türlerde meydana gelen saldırılar sistemler üzerinde büyük bir tehdit oluşturmaktadır. Yapılan bu saldırılar neticesinde kurumlar ve kullanıcılar ciddi zararlar görmektedir. Bu noktada Saldırı Tespit Sistemleri (IDS) oldukça önemli bir konuma sahiptir. Sistemlere yönelik yapılan bu saldırıların önceden tespit edilip gerekli raporların hazırlanması ileride karşılaşılabilecek tehditlerin etkisini azaltabilmektedir. Son zamanlarda yapılan çalışmalar Saldırı Tespit Sistemlerinin performanslarını artırma doğrultusunda gerçekleştirilmektedir. Bu çalışmada NSL-KDD veri seti ile SVM, KNN, Bayesnet, NavieBayes, J48 ve Random Forest algoritmaları kullanılarak sınıflandırma yapılmış, özellik çıkarımı gerçekleştirilerek sonuçların iyileştirilmesi amaçlanmıştır. Çalışma sonucunda KNN algoritması % 98.1237 doğruluk oranı ile en iyi performansa sahip olduğu gözlemlenmiştir. Bunun yanı sıra artan fold ve komşuluk sayısının sınıflandırma sonucuna etkisi karşılaştırmalı olarak incelenmiştir.

Anahtar Kelimeler: Makine öğrenimi, IDS, KNN

¹This paper has been presented at the 3rd International Conference on Artificial Intelligence and Applied Mathematics in Engineering (ICAIAME 2021).

I. INTRODUCTION

With the development of technology and the internet, our lives have become more comfortable, and information has become much easier to use and accessible. However, the internet has been the ultimate driver of globalization in recent years; a lot of data can be easily outsourced over the internet anytime and anywhere. Accessing data covering comprehensive information on passwords, systems, individuals and companies has become quite simple. Thanks to the Internet, instant access to all kinds of events and data around the world has become possible. However, this convenience always brings with it potential dangers. Abusing this convenience brought by technology, attackers attack systems and networks, posing a risk to data.

Intrusion Detection System (IDS) are methods that can help detect what type of attack is trying to damage data and systems. These determinations can be made by monitoring computer systems and network traffic. With the increasing awareness of attacks on the system, the success rate of detection systems can be increased by developing better classifiers. There are many studies in the literature using different data sets. However, it is seen that different types of classifiers give variable results for different data sets.

Mandal et al. [1] focused on security difficulties to detect intrusion in IoT system in their study. They carried out a study on the machine learning classification algorithm applied in the IoT system network to analyze diverse attacks for the IDS on effects, types, and surface attacks over the IoT network and to increase the efficiency in identifying attacks. The accuracy of the classification model was determined as 94.57 percent by applying the model to the datasets with the machine learning algorithm to classify normal attacks and abnormality of the TCP packets. Model accuracy is predicted by utilizing Machine Learning over the IoT network.

Yu et al. combined traditional intrusion detection systems with deep learning theory to improve network security and the accuracy of intrusion detection. As a result of the experiments, it has been revealed that intrusion detection systems based on multi-scale convolutional neural networks have faster convergence speed and reach higher detection accuracy [2].

Dina and Manivannan made a compilation of the problem of detecting intrusion into networks, which researchers have been dealing with for a long time. In the study, the problems in which machine learning techniques were applied in the IDS made in the last 10 years were discussed and the problems were discussed. [3].

Roy et al. have suggested a model for the Internet of Things, which is the subject of increasing study day by day. In the model they created, they used machine learning in intrusion detection systems. So as to evaluate the proposed approach, NSL-KDD and CICIDS2017 datasets have been used to carry out experiments.. With the optimization processes, they have achieved a shorter training time in their models. In addition, less training data is used thanks to the proposed approach [4].

Sharma and Yadav aimed to monitor and secure traffic in intrusion detection systems. In their studies, they have improved the error detection rate in IDS by using machine learning in which feature elimination technique is applied. Thanks to the technique they use, it is aimed to eliminate all kinds of redundancy in the KDD CUP99 dataset. They argued that the current data set of their study, in which different classification methods were applied, had a successful classification rate for intrusion detection systems [5].

Nimbalkar and Kshirsagar have proposed an intrusion detection system using Information Gain and Gain Ratio in Internet of Things network traffic, which is the target of attacks due to many security vulnerabilities in devices. The proposed system has been validated by applying it to KDD Cup 1999 and IoT-BoT datasets. As a result of the study, it was emphasized that by using 16 and 19 features, a better performance was achieved than traditional systems [6].

Fang et al. in their study, proposed a machine learning technique to resolve and detect the intrusion threat. With the proposed system, the advantages of Elman neural network and robust Support Vector Machine (SVM) noise data elimination are utilized. It has been suggested that the system detection rate can go up to 100% if the false alarm rate reaches 2.8% [7].

Verma and Ranga [8] classified the CIDD-001 dataset by using k-means clustering and k-nearest neighbor classification algorithms. Openstack and external servers data were individually evaluated for classification. It was concluded from the results that both k-means clustering and k-nearest neighbor classification outperformed the CIDD-001 dataset from the point of the salient metrics used.

Gu and Lu proposed an effective IDS based on the SVM with feature embedding in the Navie Bayes algorithm in their study to ensure network security. By using the transformed data, the SVM classifier was trained and as a result of the experiments, the proposed method reached 99.35% accuracy on the NSL-KDD dataset [9].

The aim of this study is to classify the NSL-KDD dataset, which is widely utilized in the literature, with different algorithms, as well as to examine the feature extraction, fold and neighborhood effects. Also, it has been investigated the changes in the results as a result of the improvement on performance. Thus, it has been aimed to contribute to the literature.

In this study, classification process has been performed with SVM, KNN, Random Forest, J48 and Bayes algorithms using NSL-KDD data set. After the classification was completed, the process has been repeated by performing feature extraction to improve the results and the KNN algorithm has been applied by repeating for different neighborhood values.

II. MATERIAL & METHOD

As stated before, the NSL-KDD dataset was used in this study. NSL-KDD dataset, which is one of the foremost commonly used datasets in IDS, contains the records of internet traffic seen by a simple intrusion detection network and expresses the estimation of the traffic encountered by a real IDS. The NSL-KDD dataset has been created by removing some records from the KDD-CUP 99 dataset and it includes 41 features [10]. In the first part of the study, classification was carried out for all features, and in the following stages, as given in Table 1., the 10 features with the lowest impact were extracted.

Table 1. NSL-KDD dataset attribute table

Attributes	
wrong_fragment	num_access_file
root_shell	num_outbound_cms
su_attempted	count
num_root	srv_count
num_file creations	dst_houst_count

Weka is a extensive open source software that allows to preprocess big data, applying different machine learning algorithms to big data and contrast various outputs. This software facilitates working with big data and many operations using machine learning algorithms. Depending on the type of machine learning model that is being developed, automatic selection for features is provided to form a dataset with reduced selection features by choosing from options such as Cluster, Classify or Associate. In the study, the classification of the NSL-KDD data set with various algorithms was carried out through WEKA.

III. RESULTS & DISCUSSIONS

The data set was primarily classified by SVM, KNN, Random Forest, J48, Bayesnet and NaiveBayes algorithms. In the next stages, performance analysis was made for the case of Fold value of 6, 10 and 20. In the continuation of the study, after feature extraction, the classification was repeated for different neighborhood values in the KNN algorithm, which gave the best accuracy.

A. ACCURACY ACCORDING TO CLASSIFIERS

Table 2. and Figure 1. show the accuracy rates and classification times according to different classifier types for the 10 fold value. When the graph is examined, it is seen that the Random Forest algorithm, which correctly classifies 22252 samples with 98.70% accuracy, reaches the highest accuracy rate. However, it is clear that the algorithm with the lowest percentage is the NaiveBayes algorithm with an accuracy rate of 80.74%. In addition, when the classification times are compared, SVM reaches the highest time. The KNN algorithm has the lowest classification time with 0.05 seconds. When the results were examined, it has been seen that they showed parallelism with the study of Tavallae et al. [10], however, different accuracy rates have been obtained due to the difference in the data set and software used.

Table 2. Results of classifier effect

Classifier	Correctly Classified Instances	Accuracy (%)	Time
Bayesnet	21446	95.13	0.92
KNN	21999	97.58	0.05
J48	22228	98.60	1.8
NaiveBayes	18201	80.74	0.19
Random Forest	22252	98.70	7.51
SMO	21327	94.60	80.75

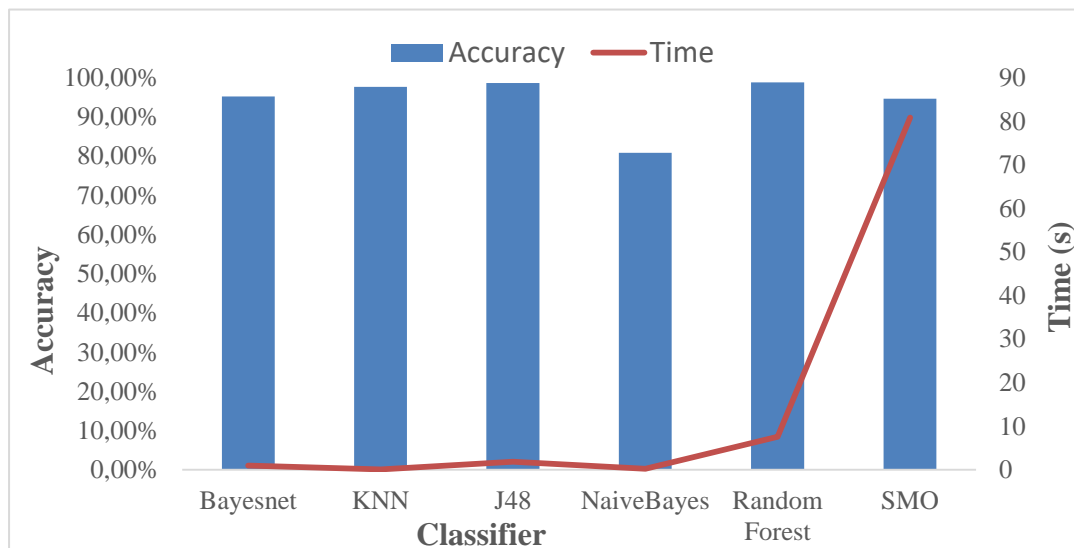


Figure 1. Accuracy and model setup time according to classifiers

B. FOLD EFFECT

Table 3. and Figure 2. show the effect of increasing fold value on different algorithms. It cannot be said that the changing fold value has a generally accepted effect on the algorithms. In the Bayesnet algorithm, there is an increase in the accuracy rate in the transition from 6 fold to 10 fold value, while a decrease is observed in the accuracy rate when changing 10 fold-20 fold. Except for SVM, there is a decrease and then an increase in others. In SVM, on the other hand, a result could not be obtained when the fold value is 20.

Table 3. Results of fold effect

Classifier	Accuracy (%)		
	Fold Value		
	6	10	20
Bayesnet	95.063	95.1295	95.1073
KNN	97.6047	97.58	97.5869
J48	98.6826	98.5983	98.6914
NaiveBayes	80.833	80.7355	80.7443
Random Forest	98.7181	98.7048	98.7136
SMO	94.5795	94.6017	

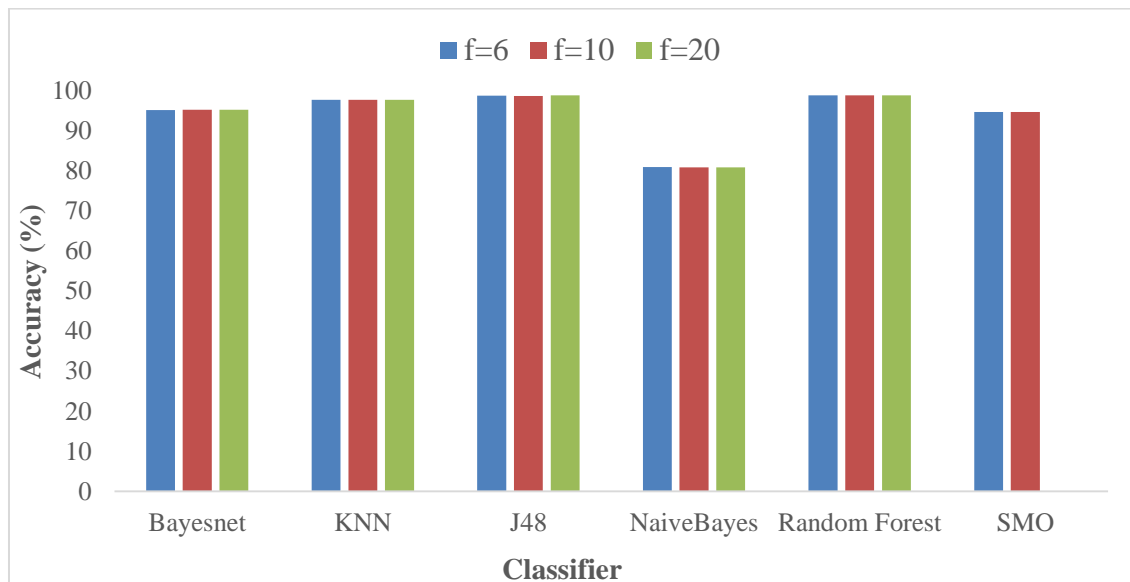


Figure 2. Fold value effect

C. FEATURE EXTRACTION EFFECT

In this section, the analyzes have been made for the 10 fold value. The classification results after the extraction of the 10 features with the lowest effect are shown in Table 4. While Random Forest algorithm had the highest accuracy rate before feature extraction, KNN algorithm reached the highest rate with 98.1273% with feature extraction. While there is a decrease in the accuracy rate for the J48 algorithm, it is seen a significant increase in the accuracy percentage of the NaiveBayes algorithm.

Table 4. Feature extraction effect

Classifier	42 Features		Total Number Instances	32 Features	
	Correctly Classified Instances	Accuracy (%)		Correctly Classified Instances	Accuracy (%)
Bayesnet	21446	95.13	22544	21466	95.2182
KNN	21999	97.58		22121	98.1237
J48	22228	98.60		21959	97.4051
NaiveBayes	18201	80.74		21436	95.0852
Random Forest	22252	98.70		---	---
SMO	21327	94.60		---	---

D. NEIGHBORHOOD EFFECT

After feature extraction, reclassification was made in different neighborhood values in order to improve the KNN algorithm, which has the highest accuracy rate. Figure 3. shows the classification results for 1,3,5,7 and 9 neighborhood values. As can be clearly seen from the graph, the increase in the number of neighborhoods decreased the accuracy rate in the KNN algorithm. While the accuracy rate for 1 neighborhood was 98.1237%, this rate was 97.6535% for 9 neighborhoods.

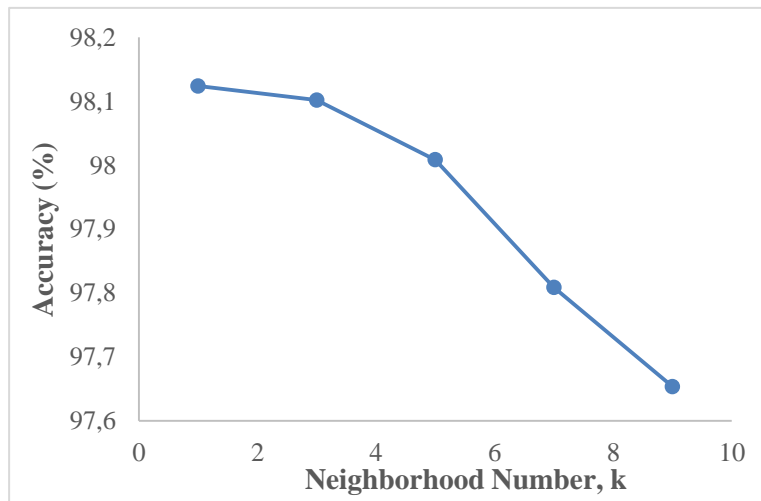


Figure 3. k value for KNN

E. DISCUSSION

In order to evaluate the results of the study, the study outputs of Tavallae et al. are discussed in this section by comparing them with the present results. Figure 4 shows the performance values of the classifiers as a result of the study performed by Tavallae et al. [10] using the NSL-KDD data set.

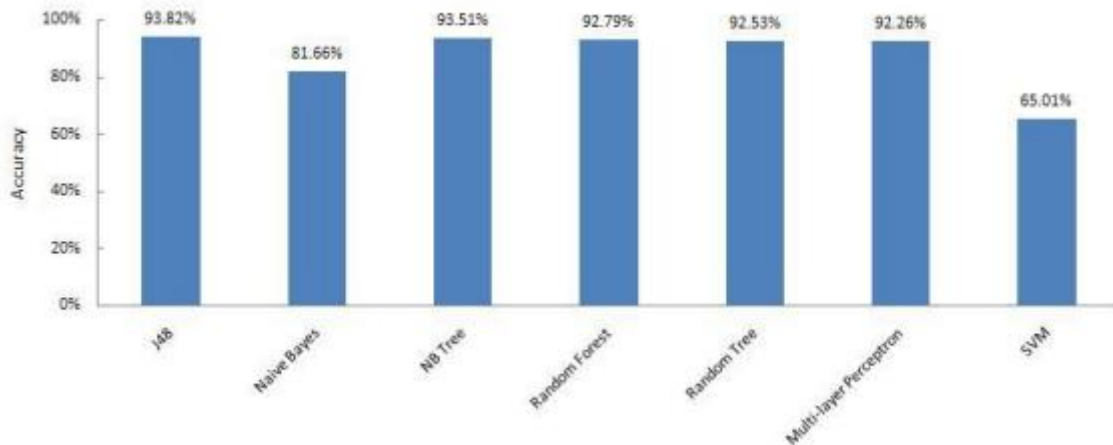


Figure 4. The performance of machine learning algorithms on the data set in the study selected for validation

As can be clearly seen from the figure, there is a concordance between the present results and the results selected in the validation. When the J48, Naive Bayes, Random Forest, and SVM classifiers are examined, it is seen that the accuracy rates of 93.82%, 81.66%, 82.79% and 65.01% are reached, respectively. In the present results, 98.60%, 80.74%, 98.70% and 94.60% accuracy rates were achieved in J48, Naive Bayes, Random Forest and SVM classifiers, respectively. This is because the studies have a different approach. It is thought that the widespread use of machine learning algorithms in studies will allow future studies to increase and to be used in different areas.

IV. CONCLUSION

In this study, it is aimed to see the effect of different algorithms on the classification result for the NSL-KDD dataset and to determine the most appropriate algorithm.

As a result of classification, the accuracy rates were compared for 6 different algorithms proposed for the current data set, and it was seen that the Random Forest algorithm reached the highest rate with 98.70%. As a result of feature extraction, it is seen that the KNN algorithm achieves the highest accuracy in the repeated classification process by removing the 10 features that have the least effect, while there is an improvement of about 15% in the accuracy rate for NaiveBayes, one of the other classifier types. However, it has been concluded that the increase in the number of neighborhoods in the KNN algorithm has a negative effect on the accuracy rate and the highest accuracy rate is reached in the number of 1 neighborhood.

In future studies, it is planned to contribute to the literature by comparing the effects of fold number, feature extraction and neighborhood number values on performance for different data sets.

V. REFERENCES

- [1] K. Mandal, M. Rajkumar, P. Ezhumalai, D. Jayakumar, and R. Yuvarani, "Improved security using machine learning for IoT intrusion detection system", *Materials Today: Proceedings*, doi: <https://doi.org/10.1016/j.matpr.2020.10.187>.
- [2] J. Yu, X. Ye, and H. Li, "A high precision intrusion detection system for network security communication based on multi-scale convolutional neural network", *Future Generation Computer Systems*, doi: <https://doi.org/10.1016/j.future.2021.10.018>.

- [3] A.S. Dina and D. Manivannan, "Intrusion detection based on machine learning techniques in computer networks", *Internet of Things*, vol. 16, no. 100462, 2021.
- [4] S. Roy, J. Li, B.J. Choi, and Y. Bai, "A lightweight supervised intrusion detection mechanism for IoT networks," *Future Generation Computer Systems*, vol. 127, pp. 276-285, 2022.
- [5] N. V. Sharma and N. S. Yadav. "An optimal intrusion detection system using recursive feature elimination and ensemble of classifiers," *Microprocessors and Microsystems*, vol. 85, 2021.
- [6] P. Nimbalkar and D. Kshirsagar, "Feature selection for intrusion detection system in Internet of-Things (IoT)," *ICT Express*, vol. 7, pp. 177-181, 2021.
- [7] W. Fang, X. Tan, and D. Wilbur, "Application of intrusion detection technology in network safety based on machine learning," *Safety Science*, vol. 124, 2020.
- [8] A. Verma and V. Ranga, "Statistical analysis of CIDDS-001 dataset for Network Intrusion Detection Systems using Distance-based Machine Learning", *Procedia Computer Science*, vol. 125, pp. 709-716, 2018.
- [9] J. Gu and S. Lu, "An effective intrusion detection approach using SVM with naïve Bayes feature embedding," *Computers & Security*, vol. 103, 2021.
- [10] M. Tavallaei, E. Bagheri, W. Lu, and A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," *Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA)*, 2009.