The Pharma Innovation Journal 2019; 8(2): 865-869

www.ThePharmaJournal.com

The Pharma Innovation

ISSN (E): 2277- 7695
ISSN (P): 2349-8242
NAAS Rating: 5.03
TPI 2019; 8(2): 865-869

www.thepharmajournal.com
Received: 14-12-2018
Accepted: 21-01-2019

Priyanka
Assistant Professor, Computer Science & Engineering, Lingaya's Vidyapeeth, Faridabad, Haryana, India

Correspondence Author;
Priyanka
Assistant Professor, Computer Science & Engineering, Lingaya's Vidyapeeth, Faridabad, Haryana, India

Enhanced intrusion detection system through machine learning on NSL-KDD Dataset

Priyanka

DOI: https://doi.org/10.22271/tpi.2019.v8.i2n.25420

Abstract
Cyberattacks that have the potential to bring down a network are continuously being sought after in the field of network security. Malicious acts on the network are also expanding quickly due to the unexpected creation and increased use of the Internet. An effective intrusion detection system (IDS) is necessary to stop unwanted access to network resources in order to identify anomalies in the network and safeguard data. Recently, a number of notable approaches have been put forth as a cure-all for intrusion detection, but it is still difficult to construct a secure system because attackers frequently alter their tactics to get around the system's security measures. In this research, the categorization of data into normal or intrusive categories was accomplished through the application of machine learning (ML) classifiers. A diverse set of classifiers was utilized in this study, encompassing logistic regressions (LR), extra-tree classifiers (ETC), Decision trees (DT), logistic support vector machines (SVM), random forests (RF), Naive Bayes (NB), multi-layer perceptron's (MLP), and K-nearest neighbors (KNN). Four feature subsets from the NSL-KDD dataset were used in the study to evaluate the model's efficacy. A thorough pre-processing of the data was carried out, which included deleting unnecessary attributes from the dataset. This was a critical step because it acknowledged that an intrusion detection system's dimensional aspect are closely related to its effectiveness. The empirical findings revealed that KNN exhibited a performance exceeding 99 percent across all attack classes when applied to various feature subsets. Consequently, through the strategic removal of unnecessary features, the proposed model not only mitigates computational complexity but also attains a notable high prediction accuracy rate.

Keywords: R2L, U2R, SVM, KNN, IDS, machine learning, Probe, DoS

Introduction
Security has emerged as a top worry in computer networks due to the gradual escalation of cyber-attacks. Furthermore, as cutting-edge technology like the cloud emerge, networks are subjected to a variety of intrusive activities that subsequently cause serious impairments. Additionally, they can result in significant financial losses and have a negative effect on an important IT infrastructure, which makes cyberwar data inadequate [1]. Security measures must be implemented to make sure the security of the system is not jeopardized. Sensitive data is protected by a secure system's ability to detect abnormalities. Even with firewalls and encryption techniques installed, a number of attacks are still able to bypass the system's defenses. In order to prevent potential harm to important resources, it is crucial to identify them as soon as possible. Then, appropriate steps can be considered to stop the intrusion.
Intrusion Detection Systems (IDS) are an essential component of cybersecurity. They are designed to monitor network traffic and identify any unauthorized access, malicious activities or attacks. Traditionally, IDS systems have relied on rule-based systems to detect potential threats. However, with the increasing sophistication of cyber-attacks, these systems are becoming inadequate, and there is a need for more advanced solutions [2].
Machine Learning (ML) techniques have been increasingly adopted in IDS systems as they have the ability to learn and adapt to new threats in real-time. An emerging field of study called Automatic Intrusion Detection Systems through Machine Learning (AIDSM) uses ML methods to identify and stop cyberattacks. This research paper aims to provide an overview of AIDSM, its advantages, challenges and limitations, and potential future directions [3].
The inherent challenges of Intrusion Detection Systems (IDS) encompass issues such as inaccurate judgments, false detections, and a lack of real-time responsiveness. In the realm of security, the considerable computational capabilities of computers have ushered in notable progress in machine learning (ML) technology.

ML classifiers have emerged as pivotal tools in addressing these challenges, significantly enhancing system accuracy and resilience. This has led to their widespread adoption for identifying and reporting various attacks in the security domain [4].

In recent research endeavors, scholars have leveraged a spectrum of ML techniques to discern and categorize features within network data. Notable techniques include support vector machines (SVM) K-Nearest neighbor (KNN) random forests (RF) multi-layer Perceptron's (MLP) decision trees (DT) and random forests (RF) among others. These methodologies underscore the dynamic and diverse application of ML in fortifying IDS capabilities against the evolving landscape of cybersecurity threats [5].

Literature Review

An extensive examination of the studies carried out in the field of intrusion detection systems (IDS), which are employed to recognize different kinds of network attacks, has been supplied. IDS is frequently created using ML techniques, which are widely used because they can handle massive amounts of data and enable network administrators to devise suitable countermeasures. Feature selection is essential because it improves the model's overall performance by eliminating superfluous and redundant features from the dataset. The wrapper-based feature selection method was introduced in and selects the most important features from the dataset using an optimizer that draws inspiration from pigeons. Using the UNSW-NB15, NLS-KDD, and KDDCUP 99, it was examined. A comparison of the model's precision performance with other well-known feature selection techniques revealed that it performed better.

Numerous machine learning classifiers, such as KNN, SVM, logistic regression (LR), Nave Bayes (NB), and DT, were used to train and assess the model. When ten important features were taken into account, the accuracy of the model was optimized using particle swarm optimization, and it was discovered to be approximately 99 percent on the NSL-KDD dataset. There are many uses for SVM in intrusion detection; used it and evaluated the results with the UNSW dataset. The accuracy of the model was found to be 94% after a comparative analysis with other classifiers, including RF, RepTree, and MLP. Similar to who discovered that enhanced whale optimized SVM produced 99.86 percent accuracy using three modified variants of SVM.

The Wireshark tool was used to capture network data packets, and a variety of ML classifiers were used to train and evaluate the model in Following the pre-processing stage, accuracy levels of 83.6%, 98.2%, 99.8%, and 95.1% were attained through the application of NB, SVM, RF, and KNN models. Utilizing a hybrid knowledge-based and machine learning approach, various attack types on KDD-99 were identified [6]. The proposed system selected the most suitable model for conducting predictions by navigating through the target classes using a knowledge-based strategy, yielding promising results.

In a comprehensive investigation detailed in reference, three classifiers—namely, MLP (Multi-Layer Perceptron), SVM (Support Vector Machine), and NB (Naive Bayes)—were scrutinized using diverse feature sets. The results of the experiment revealed that MLP exhibited superior capability in distinguishing between benign traffic, malicious windows, and mixed traffic when compared to the other classifiers. Notably, the proposed model demonstrated swiftness in its operations, rendering it particularly advantageous for advanced Intrusion Detection System (IDS) alerts.

Upon integration with traditional IDS, the combined model yielded an impressive accuracy of 92.09 percent and a False Alarm Rate (FAR) of 0.27 percent. The training methodology initially followed a precedent set by where MLP and neural networks were employed to detect network intrusions based on the classification of data as either normal or indicative of an attack. This underscores the effectiveness of MLP in enhancing the accuracy and efficiency of IDS, especially when integrated with conventional approaches.

Methodology

This section has discussed the findings of the suggested model, which used SVM and KNN to categorize the information as intrusive or typical. The NSL-KDD dataset's four unique feature subsets were used to evaluate the model's performance. The process is broken down into steps that are covered in detail in the sections that follow. Preparing the data for processing is the first step. Finding the significant feature subsets within the evaluated model is the second step. In the third stage, the data is tested and trained using multiple machine learning classifiers. The outcomes of the different parameters are then examined.

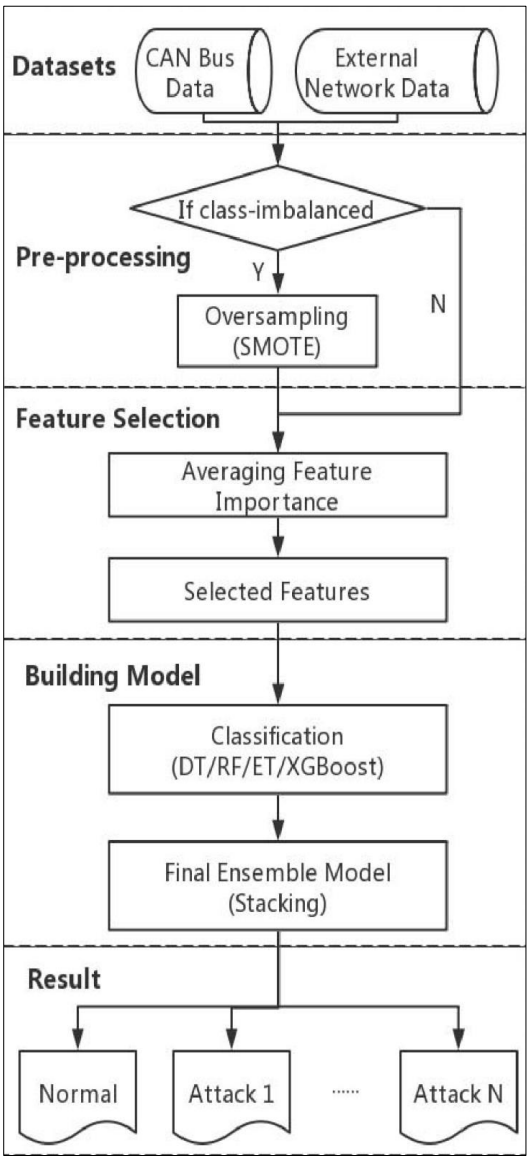

Fig 1: System Framework

A. Data Collection and Data Preprocessing

Preprocessed network traffic data including both attack and normal data can be found in the NSL-KDD dataset. The two groups that were produced from the dataset are the training set and the testing set. The training collection consists of about 125,000 records, while the testing collection consists of 23,000 records. There are a total of 41 features in the dataset, including some fundamental ones like protocol type, service, and source and destination IP addresses, as well as some more sophisticated ones like error rate and the number of unsuccessful login tries.

The dataset has basically 23 types of attacks but these attacks are categoried into main attacks that are important.By performing the transformation, the function allows more easily group attacks by their respective attack class, which can be useful for analyzing and comparing the performance of intrusion detection algorithms on different attack types.

The following entries from the NSL-KDD dataset were removed to add more clarity: 67343, 45927, 11656, 995, 52 entries from Normal, DOS, Probe, R2L, and U2R. The distribution of these records is shown in Fig. 2

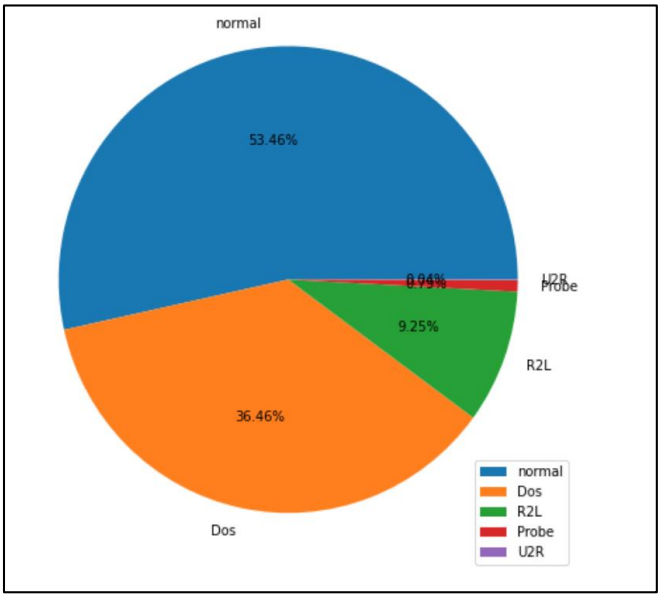

Fig 2: Pie chart distribution of Multi-Class labels

B. Feature Selection

The act of replacing or removing symbolic or non-numeric attributes is known as preprocessing. In the current study to lessen the computational complexity of the model and increase its accuracy at the same time, only the important attributes were used. The model's training and testing times were greatly shortened because only relevant characteristics were used. Pre-processing data is essential for dimension reduction, which reduces the processing time required for the data.

For simplicity purposes, In the NSL-KDD dataset, 67343, and 58630 entries of Normal, and Abnormal attacks were extracted from the dataset, here the attacks were binarily classified. The distribution of these records is shown in Fig.3.

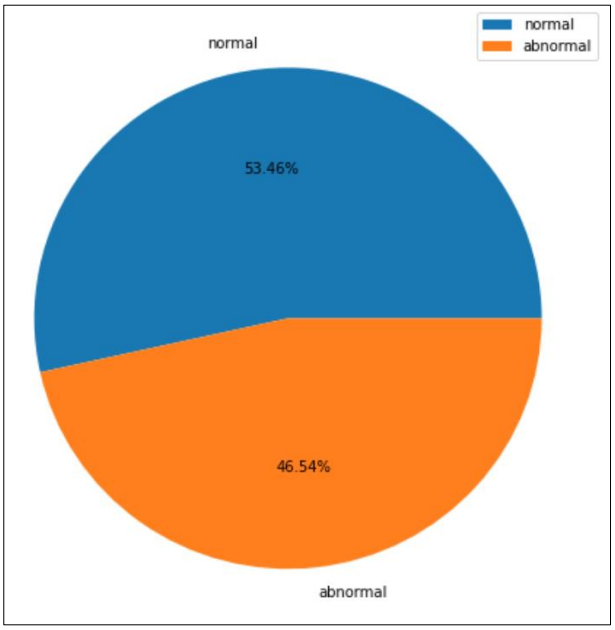

Fig 3: Pie chart distribution of Normal and Abnormal Attacks

Only the numeric columns of the original data framebin_data are selected and creates a new dataframe called numeric_bin. The numeric_col variable contains the names of the numeric columns in bin_data.constructing a data frame that only contains the encoded label attribute and the binary class dataset's numerical attributes. Subsequently determining the characteristics that have a greater than 0 point5 correlation with the encoded attack label attribute. Table 1 below displays a selection of attributes:

Table 1: Various feature subsets that were taken from the NSL-KDD dataset.

count	0.613251
logged_in	0.693770
srv_serror_rate	0.710852
serror_rate	0.712861
dst_host_serror_rate	0.714247
dst_host_same_srv_rate	0.716820
dst_host_srv_serror_rate	0.717387
dst_host_srv_count	0.718579
same_srv_rate	0.798358
intrusion	1.000000

The one-hot-encoded categorical data frame was then joined with the chosen attributes. Next, the encoded, one-hot-encoded, and original attack label attribute—"intrusion," "Dos," "Probe," "R2L," "U2R," "normal," and "label"—were joined. Next, this is used to train the model.

C. Model Building

In this study, the categorization of data into normal or intrusive categories was accomplished through an intrusion analysis engine that utilized Support Vector Machines (SVM) and K-Nearest Neighbors (KNNThe ability of Support Vector Machines to handle both linear and non-linear data is well known; they are known for their binary-class capability. through input vector mapping, SVM identifies the optimal hyperplane, effectively segregating the data based on the maximum distance between support vectors. A noteworthy feature of SVM is its resilience to outliers, as it determines the hyperplane exclusively using support vectors, rather than relying on the entire training set. In this study, the radial basis

kernel function was employed for training, recognizing its efficacy in yielding favorable outcomes, especially in dealing with non-linear data.

Similarly, KNN, a supervised learning algorithm, categorizes data into distinct groups by assessing Euclidean distance and the parameter K. In this study, K was set to 5, with data points located through an evaluation of the model's performance at that value. The iterative process of locating data points and classifying them into different categories continues until convergence, mirroring the functionality of KNN in the context of this research.

Results

In this study, the selection of features from NSL-KDD was contingent upon the specific model under testing or training. The aim was to efficiently decrease the dimensionality of the data by adopting a random feature selection approach. This method has proven to be a practical way to reduce model complexity and enhance training time optimization. While effective in the current model, it is crucial to recognize that the appropriateness of this approach may differ in various contexts.

To assess the performance of the chosen approach, two classifiers, namely SVM and KNN, were trained and tested using 111,386 and 37,129 examples from the NSL-KDD dataset. A comprehensive evaluation of the model's performance was conducted, considering various parameters. The resulting accuracy, precision, recall, and F1-score metrics for distinct subsets are visually represented in the subsequent table and figures.

	precision	recall	f1-score	support
abnormal	0.97	0.96	0.96	14720
normal	0.96	0.97	0.97	16774
accuracy			0.97	31494
macro avg	0.97	0.97	0.97	31494
weighted avg	0.97	0.97	0.97	31494

Fig 4: Evaluation matrix for SVM

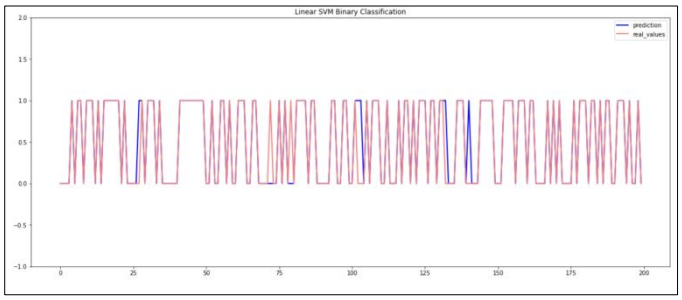

Fig 5: SVM Classification (Real values vs prediction)

On Using KNN the model performs better than SVM. Mean square error for SVM was 0.033 versus for KNN it was 0.014. Accuracy rate even increased a lot from 96% to 98% on switching from SVM model to KNN.

	precision	recall	f1-score	support
abnormal	0.99	0.98	0.98	14720
normal	0.99	0.99	0.99	16774
accuracy			0.99	31494
macro avg	0.99	0.99	0.99	31494
weighted avg	0.99	0.99	0.99	31494

Fig 6: Evaluation matrix for KNN

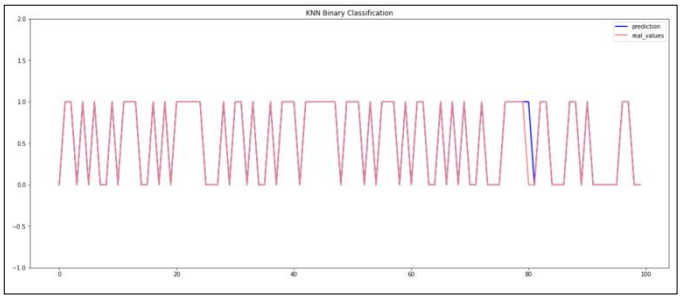

Fig 7: KNN Classification (Real values vs prediction)

Due to the existence of believable entries for this attack class in the NSL-KDD dataset, DoS attack showed the most promising results among the attack classes, while U2L attack produced unsatisfactory results. Overall accuracy was over 99 percent because KNN can handle unevenly distributed data. The bootstrap technique is used by a random forest to detect imbalances in the data, reduce misclassification of the data, and increase the occurrence of minority classes. Similar to this, KNN achieves an exceptional outcome by carrying out an exhaustive, all-encompassing examination of every option along its branches.

Conclusion and future scope

Security is now a top concern due to an increase in unauthorized access and resource exploitation. Attacks are becoming a serious problem, so limiting the damage they do to the system network requires early attack detection. These days, a lot of machine learning models are used in IDS because of their generalization potential, adaptability, and robustness. The effectiveness of the current research's inclusive empirical study in identifying network intrusion using ML classifiers—most notably SVM and KNN—was assessed using NSL-KDD. The dataset was pre-processed before the model was trained and tested using the important attributes. In the future, research efforts could concentrate on the potential application of optimization techniques to provide better distributed systems. Consequently, it is possible to research ensemble-based techniques, which forecast the results by combining the output of multiple algorithms. In addition, to lower the processing complexity of the system, different feature selection strategies can be investigated. Moreover, the intrusion detection problem must be handled by the wireless mobile network domain.

References

1. Kaushik P, Yadav R. Reliability design protocol and block chain locating technique for mobile agent. Journal of Advances in Science and Technology (JAST). 2017;14(1):136-141. https://doi.org/10.29070/JAST
2. Kaushik P, Yadav R. Deployment of Location Management Protocol and Fault Tolerant Technique for Mobile Agents. Journal of Advances and Scholarly Researches in Allied Education (JASRAE). 2018;15(6):590-595. https://doi.org/10.29070/JASRAE
3. Kaushik P, Yadav R. Mobile Image Vision and Image Processing Reliability Design for Fault-Free Tolerance in Traffic Jam. Journal of Advances and Scholarly Researches in Allied Education (JASRAE). 2018;15(6):606-611. https://doi.org/10.29070/JASRAE
4. Kaushik P, Yadav R. Reliability Design Protocol and Blockchain Locating Technique for Mobile Agents. Journal of Advances and Scholarly Researches in Allied

Education (JASRAE). 2018;15(6):590-595. https://doi.org/10.29070/JASRAE
5. Kaushik P, Yadav R. Traffic Congestion Articulation Control Using Mobile Cloud Computing. Journal of Advances and Scholarly Researches in Allied Education (JASRAE). 2018;15(1):1439-1442. https://doi.org/10.29070/JASRAE
6. Sun L, Ho A, Xia Z, Chen J, Zhang M. Development of an Early Warning System for Network Intrusion Detection using Benford's Law Features. In: Security and Privacy in Social Networks and Big Data SocialSec. Singapore: Communications in Computer and Information Science, Springer; 2019:1095:57–73.
7. Darkaie M, Tavoli R. Providing a method to reduce the false alarm rate in network intrusion detection systems using the multilayer Perceptron technique and backpropagation algorithm. In: 5th Conference on Knowledge-Based Engineering and Innovation. Tehran, Iran; 2019:1-6.
8. Wang W, Li Y, Wang X, Liu J, Zhang X. Detecting android malicious apps and categorizing benign apps with ensemble of classifiers. Future Generation Computing Systems. 2018;78:987–994..