Protecting Accounting Information Systems using Machine Learning Based Intrusion Detection

Biswajit Panja, Phd
bpanja@emich.edu
Associate Professor of Computer Science
Eastern Michigan University, Ypsilanti, MI 48197 USA

Abstract
In general network-based intrusion detection system is designed to detect malicious behavior directed at a network or its resources. The key goal of this paper is to look at network data and identify whether it is normal traffic data or anomaly traffic data specifically for accounting information systems. In today's world, there are a variety of principles for detecting various forms of network-based intrusion. In this paper, we are using supervised machine learning techniques. Classification models are used to train and validate data. Using these algorithms we are training the system using a training dataset then we use this trained system to detect intrusion from the testing dataset. In our proposed method, we will detect whether the network data is normal or an anomaly. Using this method we can avoid unauthorized activity on the network and systems under that network. The Decision Tree and K-Nearest Neighbor are applied to the proposed model to classify abnormal to normal behaviors of network traffic data. In addition to that, Logistic Regression Classifier and Support Vector Classification algorithms are used in our model to support proposed concepts. Furthermore, a feature selection method is used to collect valuable information from the dataset to enhance the efficiency of the proposed approach. Random Forest machine learning algorithm is used, which assists the system to identify crucial aspects and focus on them rather than all the features them. The experimental findings revealed that the suggested method for network intrusion detection has a neglected false alarm rate, with the accuracy of the result expected to be between 95% and 100%. As a result of the high precision rate, this concept can be used to detect network data intrusion and prevent vulnerabilities on the network.

Keywords:
Intrusion detection, Machine learning

1. Introduction

The Internet is being the most obvious choice for everyone to transfer data in today's accounting information systems. The Internet is nothing but a combination of networks. It is difficult to preserve the network's protection against various types of attacks as its demand increases exponentially. There are two types of attacks active and passive. A passive attack is one in which the attacker does not have real-time control over the attack. An active attack is one in which the attacker directs the target's data to be attacked. As a result, attackers use information obtained during a passive attack to manipulate a target during an active attack [12]. Even though security software such as firewalls, anti-virus, and intrusion detection systems are readily accessible (IDS). However, they are unable to avoid a diverse variety of network attacks [1]. IDS attempts to detect attacks as they happen, while firewalls and anti-virus applications attempt to block them. IDS also evaluates network traffic that passes through its ports, but it can't stop it. The primary objective of IDS is to track all of the platform's irregular activity. Detection systems for IDS may be network-based or host-based. IDS based on hosts are made up of computers linked to a single network or a remote server. It detects packets from other systems and notifies administrators if an attack is detected [4].

In this study, we have developed a model to detect network-based intrusion for accounting information systems. So, we can call it a network-based Intrusion detection system (NIDS). Intrusion Detection refers to the task of examining server logs for tracks and determining whether or not there has been any interference. Intrusion occurs when a security system is breached, and intrusion detection is the way of identifying intrusions. It monitors the flow of packets through the network and detects attacks that haven't been predicted. There are many NIDS available to detect attacks using two techniques as misuse and anomaly [13]. In this research, we are detecting two types of data whether it is an anomaly or normal. In the case of anomalies, it compares unfamiliar patterns to known patterns to determine if they are regular or abnormal. Anomaly detection considers that regular user behavior is perfectly measurable and sufficiently distinct from invasive behavior. Following that, it determines what traits they already have based on regular patterns. Since it always has a reference to the normal pattern. Misuse detection has a pattern set of well-known risks which searches for matches in the tracked data and detects an attack if there is a match [3]. The Misuse technique can detect attacks from a predefined pattern with high accuracy. Anomaly detection, contrary, cannot detect attacks based on a predefined pattern and has a high rate of false alarms. Our concept is focused on an anomaly detection model to improve the detection performance like accuracy, consistency, detection, false alarm [3].

The challenge of anomaly detection using various machine learning and data mining techniques is the focus of a lot of similar studies in the field. Many factors must be considered when developing a machine learning-based NIDS, we have to first collect data then the next step is to perform data pre-processing, after that intrusion detection along with analysis and prediction. Data collection is one of the processes that can take a long time. Collecting meaningful data is extremely challenging. There are many approaches, which produce synthetic data. Artificial data are useful in two situations: first when there is no attack data and second there is a small amount of attack data [2]. But in our case, we found data set from one of the popular data science websites which include normal network traffic and anomaly traffic for both

Manuscript received May 5, 2024
Manuscript revised May 20, 2024
https://doi.org/10.22937/IJCSNS.2024.24.5.13

training and testing purposes. Data preprocessing method used to make data appropriate for the research purpose. You can clean, normalize, transform, delete features, and do a lot more with data after it has been pre-processed. We extracted some of the redundant data from our dataset. In addition, we have encoded many categorical as well as numeric fields to make them more useful for the analysis. Data preprocessing can take some time, but it will make the rest of the process go more easily and effectively.

Machine learning algorithms are trained using system experience in intrusion detection systems. There are mainly two techniques in machine learning. Supervised learning and unsupervised learning. Supervised learning enables the generation of data output from existing pattern knowledge. In the testing phase, the feature must forecast the class for unintended instances [9]. Unsupervised learning is used to derive patterns from unlabeled input data. Unsupervised learning aims to extract structure and patterns from input data. In this study, our ultimate goal is to find the network traffic data is normal traffic or irregular. So, we are using the supervised learning method. Our model has learned from the supervised dataset and predicts output accordingly. In the proposed concept section we are going to discuss more on this. Furthermore, there are two concepts in supervised learning classification and regression. The main difference between them is when you try to forecast a volume, use regression. When you want to forecast a class or type, use classification. We have used supervised data with class labels and at the end, we are predicting the type of traffic data to avoid intrusion on the network. So, we have used classification models in the proposed composite model. This paper evaluates and interprets the time required to construct a model using different classifier techniques. Again, in our concept section, we are going to discuss the classification models also. In this study, we are not only evaluating the model over the dataset but we are validating it. Basically in our concept, during the training process, machine learning algorithms attempt to discover associations between feature values and their categories in order to simulate new data, it is also known as testing data or validation [8]. We are evaluating and validating the model to ensure that it makes accurate predictions. In our work, we also calculate the tools that each algorithm uses to achieve high precision.

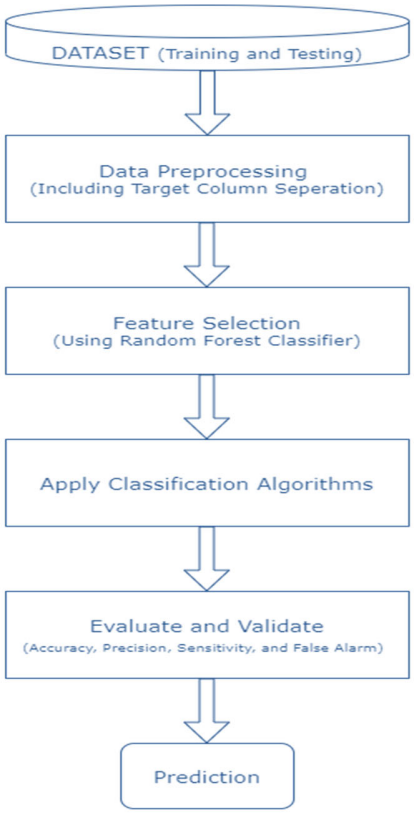
Finally, in this study first data preprocessing used such as removing redundant information, scaling numerical data, encoding categorical features. Besides that, feature selection was used in this analysis because the gathered dataset had a large number of features, rendering the entire process slow and expensive in terms of time and memory. Feature selection facilitates in the accuracy and efficiency of the entire process. Moreover, a hybrid classification approach is used on the dataset for the evaluation and validation. The current literature review is discussed in the following section. After that, we'll go over our proposed concept. Section 4 describes the implementation and outcomes. The conclusion is outlined in the final section.

2. Proposed Concept

The objective of this research is to examine network data and predict the type of data. It can be normal data or irregular attack data. In this paper, we are using different classification models to train the system and detect the types of data. There are different concepts available in today's era to detect the types of network data. We studied different current literature on network-based intrusion detection and came across different available concepts. An intrusion detection system (IDS) is a software program that monitors the network for suspicious activity and analyzes data to detect any intrusion in the network. After the current literature study, we developed our concept to satisfy the primary purpose of this paper.

The ability to identify attacks on the network is a key component in preventing them. Network Intrusion Detection Systems and Host based Intrusion Detection Systems are the two main general classifications. Host-based is an instance of a device that tracks a wide range of operating file types. And a network-based method is one that evaluates approaching traffic on the network. In this study, we have developed a concept to detect intrusion on the network data. Intrusion detection can be divided into two categories: anomaly and misuse detection. The framework administrator defines the normal, or usual, state of the ordered increased traffic, degradation, rule, and common package measure in anomaly detection. This detector screens arrange portions so that they can be compared to a standard baseline and deviations can be found.

The IDS analyzes the data it collects and applies it to massive datasets with attack marks in order to spot misuse. The IDS searches for a specific attack that has already been registered. A misuse detection software, like a virus detection method, is just as good as the database of threat signatures to which it compares packets.



Phases of Intrusion Detection Process

2.1 Dataset (Training and Testing)

We are collecting network traffic data as a notion. The data set should be large enough for the system to be evaluated and validated. The next step is to examine the dataset to learn about the various features of the data and other information. We propose a data mining method for the control of warnings in this paper in order to increase the efficiency of intrusion detection systems. It has the potential to minimize false intrusion alarms. The mechanism of intrusion detection is made up of several processes: first, the system monitors and analyzes data files or network traffic; second, suspicious events are detected; and finally, the system is checked for attack. There are also different data examination techniques available in data mining. Using those techniques, we can identify whether our dataset is supervised or unsupervised. In this paper, we are going to perform supervised machine learning. The supervised learning technique is useful when we have a training dataset well distributed using labels. We used a supervised learning algorithm in which we trained an implementation and then chose the method that better represents and predicts the data input at the final step. We are frequently unable to determine the true mechanism that often allows the right prediction and another factor is that computers only understand the commands given by human beings, so we need to use algorithms. The aim of supervised learning methods is to design dependency relationships here between different recommendation outcome and input features so that we would forecast the expected output for new information using the correlations learned from past data sets. There are various number of supervised learning algorithms to train the machine using the labeled training dataset. These algorithms analyze the training dataset and predict the output of the testing dataset. Supervised learning has mainly two categories of algorithms. Classification algorithms and regression algorithms.

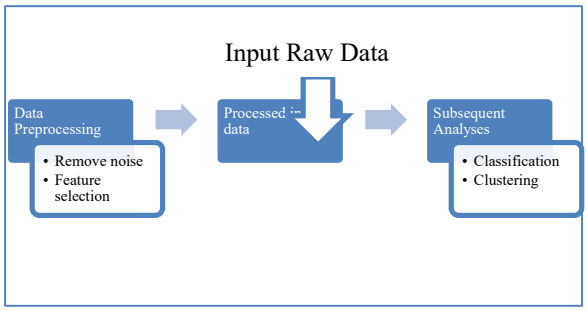
Classification algorithms are used when your output will be based on some categories, such as network data categories; a normal, anomaly, and so on. The term "classification" refers to the process of grouping production into distinct categories. Binary classification is where a process attempts to categorize input into two separate groups. The method of classifying a case into a series of two classes with the help of a classifier is known as binary classification. In an assortment of disciplines, binary classification is commonly utilized. Multiclass grouping refers to choosing from more than two or three classes. For classification, we use a variety of algorithms namely Decision trees, SVM, KNN, and lastly logistic regression. Whereas regression algorithms are used when the output of the trained system is a real value, such as the number of people having cancer, dollar, and so on. Thus, we can say that the regression is a technique for estimating distinct values from a set of individual values.

3. Proposed Algorithms

In our proposed approach we are going to detect whether the network data is normal or irregular. As a result, we've decided to continue with classification algorithms to avoid network intrusion detection. Based on network traffic categories we can prevent having an attack on our network or the systems under that network. The aim of classification is to categorize intrusions based on their characteristics.

3.1 Data Preprocessing

The point of data preprocessing is to turn the raw input data into a format that can be evaluated later. Data preprocessing involves combining information from different repositories, cleaning data to eliminate noise and repeat perceptions, and after that extracting particular perceptions based on the requirements. Preparing data for future review in accordance with the IDS model's requirements is the foremost time-consuming and energetic task. Where a dataset contains redundant records, clustering or classification algorithms take longer and give less accurate results. Dataset should be free of noise and redundant samples to realize a more reliable and effective model. The Data preprocessing block diagram in below Figure shows how data flows from raw data input to preprocess input data for measurable analysis.



3.2 Feature Selection

Moreover, this model is also using the Random Forest Classifier algorithm to select topmost features from the data, which can help the system to learn about important features and focus on them rather than all the features.

Random Forest Classification: Random forests are classified learning strategies that work by building a large number of decision trees during preparation. For decision trees that have over fitted their training collection, extremely randomized forests are right. Multiple decision trees predictions are mixed in random forests, and the ultimate result is chosen by majority voting. The data set must be part into subtrees and accompanied by the proper mix of variables when constructing a decision tree. Finding the best set of variables, on the other hand, is not simple. The aggregate result of this forest of collected trees could be a random forest. Individual decision trees are outperformed by random forest. For both classification and regression, the random forest algorithm is used.

Recursive Feature Elimination (RFE): RFE could be a function selection algorithm with a wrapper. This means that within the heart of the method, a separate machine learning algorithm is given and utilized, which is wrapped by RFE and utilized to assist choose features. Filter-based feature choices, on the other hand, rate each feature and choose the features with the highest (or least) score.

3.3 Classification Algorithms

After reviewing several studies, we have decided to make a hybrid model which integrates the various classification algorithms. There are numbers of classification algorithms to integrate K-Nearest Neighbors (k-NN) Classifier, Logistic

Regression Classifier, Support Vector Classifier (SVC), and Decision Tree (DT) Classifier. DT and k-NN are two of the utmost successful algorithms for machine learning among different classification methods, for distinguishing normal from unethical behaviors of network data [3].

K-NN Classification: The K-NN algorithm is a technique for supervised classification. It uses a collection of labeled points to teach itself how to mark certain values. It looks at the marked point nearest to the new point to mark it. It determines a label depending on which mark has the most neighbors after checking with the K number of closest neighbors.

Decision tree Classification: Decision trees build a classification and regression model arranged in the shape of a tree. One of the foremost common and natural Classification algorithms based on machine learning is decision tree. The aim is to construct a model that forecasts a target value of dependent variable from a set of input variables. The classification issue is broken down into sub-problems utilizing this procedure. It constructs a decision tree in which at that point is utilized to construct a classification model [8]. As a consequence, a tree with decision and their leaf nodes has been formed. A leaf node represents a grouping or judgment, and a node represents two or more divisions. A decision tree is used to deal with numerical results.

Besides that, our proposed composite model also contains Logistic Regression Classifier, and SVC algorithms. Logistic Regression Classifier is the outcome of combining the tree structure and the logistic regression function to make a single Inside the branches of the decision tree, there is a logistic regression method, resulting in a demonstrate of piecewise linear regression that's a real valued function, whereas conventional decision trees with constants at their takes off created the piecewise consistent.

SVC are guided learning models that interpret data for classification and regression analysis. They have associated learning calculations. An SVC training algorithm constructs a demonstration that allots unused cases to one of two categories, Provided a set of training illustrations, each of which is labeled as belonging to one of two divisions, it can be turned into a non-probabilistic binary classification approach. SVM is a widely used machine learning algorithm for a variety of purposes, including intrusion detection, spam filtering, and pattern recognition [8].

3.4 Evaluate and Validate

A binary classification confusion matrix may be a two-by-two table that's created by counting the number of binary classifier's four results. False Positive, False Negative, True Negative, and True Positive are the four types

Accuracy: The rate of accurate forecasts for the test results is known as accuracy. It's simple to figure out by calculating and dividing the overall number of forecasts by the number of true predictions. Precision: Precision is classified as the rate of significant illustrations (true positives) among all the cases expected to be a member of a certain class. Sensitivity (or Recall): The proportion of correct positive forecasts to the total number of positives generates sensitivity. False Alarm: The false positive rate is computed by dividing the total of inaccurate correctly predicted by the overall number of negativity.

3.5 Prediction

This concept will save resource utilization, such as time and memory. Using this concept we are planning to get negligible false alarm rate for the network traffic detection and also expecting the accuracy of the result will be between 95% and 100%, which is comparatively high. With high precision data as a result, we can rely on this principle to detect network data intrusion and avoid vulnerabilities.

As a result, our idea is focused on a network-based intrusion detection scheme, where increased internet use raised the quantity and quality of malicious attacks. Researchers also claimed that old strategies like firewalls and anti-virus couldn't protect against new attacks, according to our posts. So, to provide an accurate and smart intrusion detection system, we combine two leading machine learning algorithms: k-NN and DT. The main idea behind this project is that in future if we get any false alarm or malicious attacks then this technique can quickly catch those abnormal attacks. When designing regulations for network and firewalls, network intrusion system developers can work collaboratively with network and router management to assure that attackers will not use the functionality to restrict access to authorized users. So when we combine two or algorithms, it gives better performance and accuracy than other algorithms.

3.6 Machine Learning in Hybrid Detection

The hybrid detection system utilizes the capability and predictive capacity of an anomaly detection with both the precision and reliable of a misuse detection technique. These hybrid detection models use a variety of random forest variants as their machine learning algorithms. They suggested a method for detecting internal and external attackers in our paper focused on a distributed intrusion detection scheme. We can predict interference in a real world situation from the connection layer to use the clustering method, which can be used with both supervised and unsupervised results. The detection accuracy is used to test and quantify results. If the identification rate falls, the number of unidentified attacks rises. When opposed to misuse and anomaly intrusion detection independently, the identification rate increases.

We not only suggested this idea, but we also put it into practice to see how it works. The next section goes over the implementation of the proposed concept in greater detail, with graphs and tables to support it.

4. Implementation and Results

In this study the concept is not only proposed but also implemented and generating the results to reinforce the concept. For the implementation of the proposed concept Python programming language is used. More specifically, the 3.8.6 version of Python is used. Other than that various libraries are applied such as “sklearn”, “matplotlib”, “numpy”, “pandas”, and so on, which supports required functionality for the program. All implementation were carried out on a 2.20 GHz Intel Core i7 processor with 16 GB of RAM.

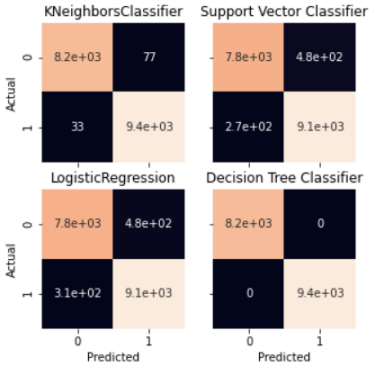
4.1 Definition of the Dataset and data preprocessing

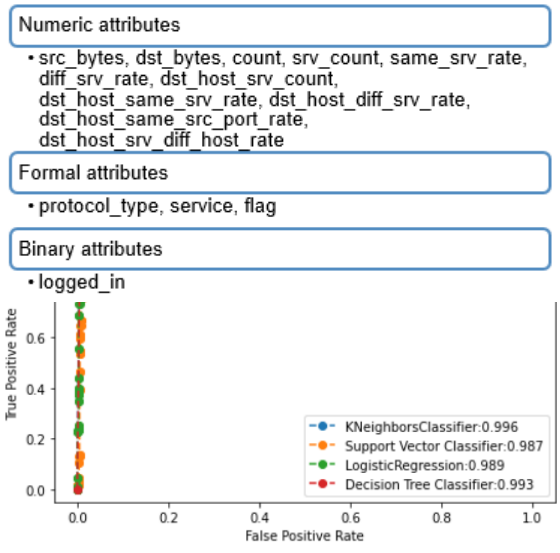
During this research process, the network traffic dataset is used from “www.kaggle.com”. This dataset has around 25000 network traffic records over TCP, ICMP, and UDP protocol with the duration. There are 41 features to identify the attack data from the normal data. Among 41, 3 are qualitative features and others

are quantitative features. Moreover, because the raw data was not ready to fit into the proposed concept data preprocessing was executed. After analyzing the raw data, redundant features are removed, and the dataset is then standardized using the StandardScaler process. The LabelEncoder method is also used for categorical attributes. The target column is defined during the data preprocessing stage. It has two class categories: Normal and Anomalous.

4.2 Feature Selection Phase

Feature selection is the technique of selecting relevant attributes for the model construction. The significance of the features is mapped using the Random Forest Classifier algorithm. The diagram illustrates the significance of different attributes. Subsequently, the most significant attributes are then chosen using the Recursive Feature Elimination (RFE) model.

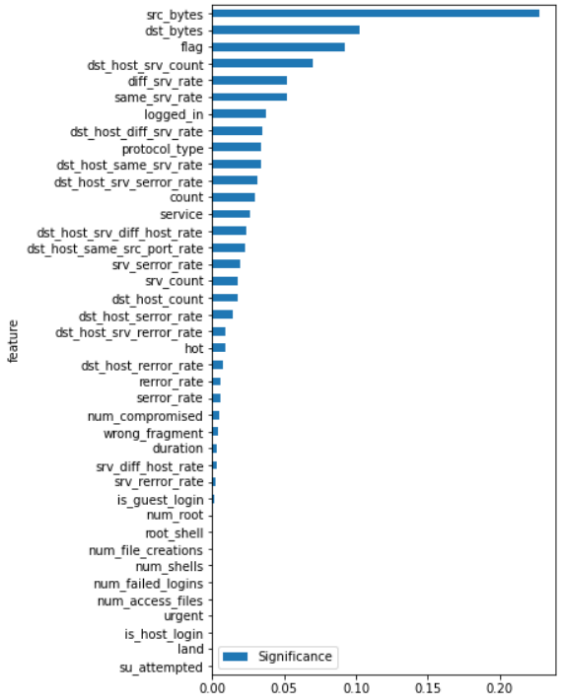




ROC Curve

4.3 Implementation of the Classification Algorithms and Evaluation Result

The classification algorithms (k-NN, LRC, DT, and SVC) have been used in this integrated model. To evaluate and validate these classifiers, the 10-fold cross-validation methodology is used, which divides the dataset into ten non-duplicated sub-datasets. Evaluation and validation are performed on these sub-datasets, implying that the classifiers are qualified

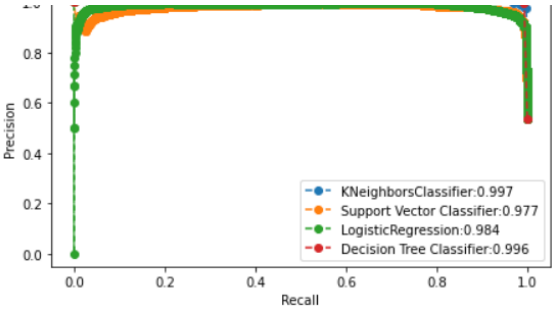


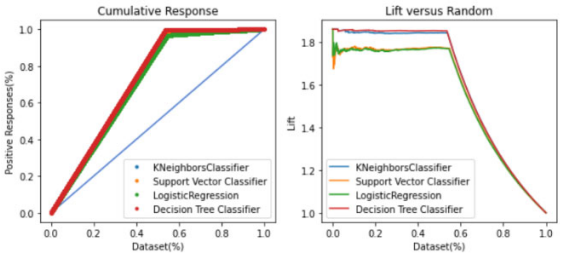
and verified ten times. Furthermore, to detect intrusion on the network accuracy, precision, and sensitivity rate must be high with a low false alarm rate. The evaluation outcome is as follows:

Classifier	*Accuracy*	*Precision*	*Recall*	*False Alarm*
k-NN	99.17%	99.8%	99.8%	00.98%
SVC	95.86%	96.9%	96.9%	04.76%
LRC	95.5%	96.7%	95.7%	04.90%

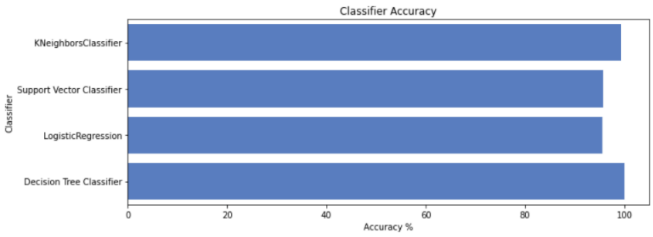
Features Importance

PR Curve





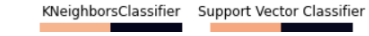
Cumulative Response and Lift Curve



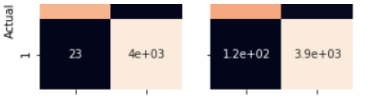
Classifier Accuracy

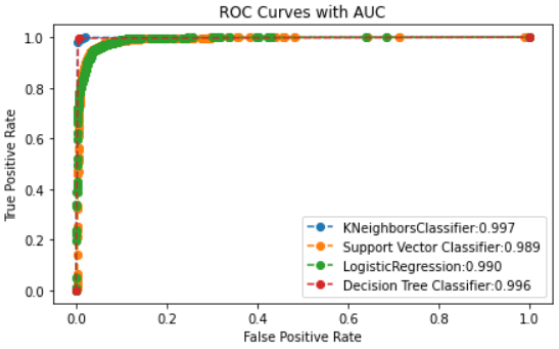
4.4 Validation Result

This research not only evaluates but also validates the proposed concept for detecting intrusion over the network. The proposed concept is validated using the same approach as the evaluation. The validation outcome is as follows:



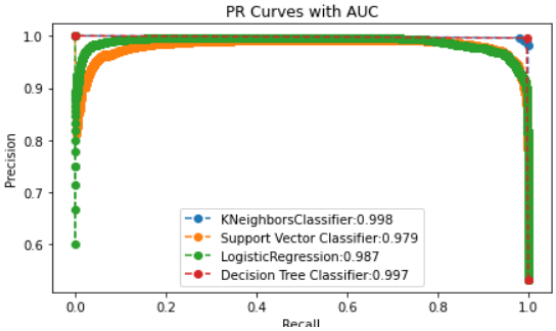
Confusion Matrix in a form of Heat Map





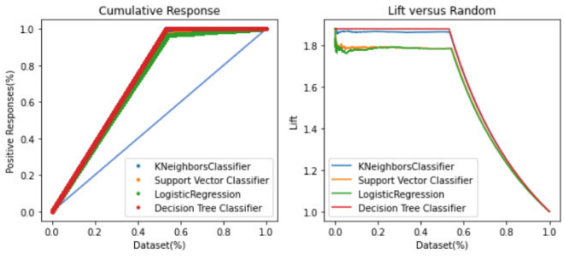
ROC Curve

Classifier	*Accuracy*	*Precision*	*Recall*	*False Alarm*
k-NN	99.14%	99.7%	99.7%	00.82%
SVC	95.75%	96.7%	96.7%	05.0%
LRC	95.49%	96.4%	95.4%	05.04%
DTC	100.00%	100.00%	100.00%	100.00%

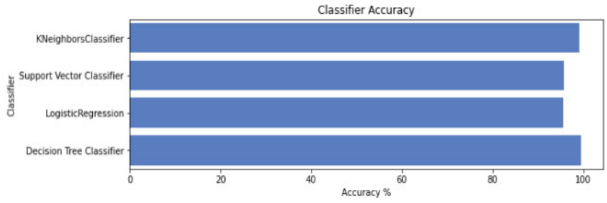


PR Curve

The results of the assessment and validation indicate that the proposed hybrid classification method has a high rate of accuracy, precision, and recall, as well as a low rate of false alarms. As a result, the suggested concept should be used to detect network data intrusion.



Cumulative Response and Lift Curve



Classifier Accuracy

5. Conclusions

A composite network-based intrusion detection approach is proposed in this research study, which classifies network traffic results. If it's regular or unusual traffic. This approach is useful for preventing network vulnerabilities. Prior to classification and after data collection, a data pre-processing and feature extraction method is applied to the dataset. To increase the performance of the suggested methodology, a significant feature extraction process is used. Furthermore, various classification algorithms such as DT, k-NN, SVC, and LRC are trained to test and verify the system. These classification models are used in conjunction with a 10-fold cross-validation technique. The analytical results showed that the proposed system for network intrusion detection has a low false alarm rate, with the accuracy of the result estimated to be between 95% and 100%. Because of the high precision score, this concept can be used to detect network intrusion and avoid potential threats.

References

[1] Nalavade K. and Meshram B. "Intrusion prevention systems: data mining approach." *In proceedings of the International Conference and Workshop on Emerging Trends in Technology*, 2010, pp. 211–214, doi:https://doi-org.ezproxy.emich.edu/10.1145/1741906.1741952

[2] Pham T., et al. "Generating Artificial Attack Data for Intrusion Detection Using Machine Learning." *In proceedings of the Fifth Symposium on Information and Communication Technology*, 2014, pp. 286–291, doi:https://doi-org.ezproxy.emich.edu/10.1145/2676585.2676618

[3] Foroushani A. and Li Y. "Intrusion Detection System by Using Hybrid Algorithm of Data Mining Technique." *In proceedings of the 2018 7th International Conference on Software and Computer Applications*, 2018, pp. 119–123, doi:https://doi-org.ezproxy.emich.edu/10.1145/3185089.3185114

[4] Chapke P. and Deshmukh R. "Intrusion detection system using fuzzy logic and data mining technique." *In proceedings of the 2015 International Conference on Advanced Research in Computer Science Engineering & Technology*, 2015, pp. 1-5, doi:https://doi-org.ezproxy.emich.edu/10.1145/2743065.2743128

[5] Li Y., et al. "Intrusion detection algorithm based on deep learning for industrial control networks." *In proceedings of the 2019 The 2nd International Conference on Robotics, Control and Automation Engineering*, 2019, pp. 40-44, doi:https://doi-org.ezproxy.emich.edu/10.1145/3372047.3372092

[6] Belouch M. and Hadaj S. "Comparison of ensemble learning methods applied to network intrusion detection." *proceedings of the Second International Conference on Internet of things, Data and Cloud Computing*, 2017, pp. 1-4, doi:https:// doi-org.ezproxy.emich.edu/10.1145/3018896.3065830

[7] Yu Y., et al. "Attacks and Defenses towards Machine Learning Based Systems." *proceedings of the 2nd International Conference on Computer Science and Application Engineering*, 2018, pp. 1-7, doi:https:// doi-org.ezproxy.emich.edu/10.1145/3207677.3277988

[8] Tungjaturasopon P. and Piromsopa K. "Performance Analysis of Machine Learning Techniques in Intrusion Detection." *In proceedings of the 2018 VII International Conference on Network, Communication and Computing*, 2018, pp. 6–10., doi:https://doi-org.ezproxy.emich.edu/10.1145/3301326.3301335

[9] Hamid Y., et al. "Machine Learning Techniques for Intrusion Detection." *In Proceedings of the International Conference on Informatics and Analytics*, 2016, Article 53, pp. 1–6, doi:https://doi-org.ezproxy.emich.edu/10.1145/2980258.2980378

[10] Gunupudi R., Mangathayaru N., and Narasimha G. "Intrusion Detection Using Text Processing Techniques." *In proceedings of the The International Conference on Engineering & MIS*, 2015, Article 55, pp. 1-6 doi:https://doi-org.ezproxy.emich.edu/10.1145/2832987.2833067

[11] Khanji S. and Khattak A. "Towards a Novel Intrusion Detection Architecture Using Artificial Intelligence." *In Proceedings of the 2020 9th International Conference on Software and Information Engineering* , 2020, pp. 185–189., doi:https://doi-org.ezproxy.emich.edu/10.1145/3436829.3436842

[12] Anand V. "Intrusion Detection: Tools, Techniques and Strategies." *In proceedings of the 42nd annual ACM SIGUCCS conference on User services*, 2014, pp. 69–73, doi:https://doi-org.ezproxy.emich.edu/10.1145/2661172.2661186

[13] Jafier S. "Utilizing feature selection techniques in intrusion detection system for internet of things." *In Proceedings of the 2nd International Conference on Future Networks and Distributed Systems*, 2018, pp. 1–3, doi:https://doi-org.ezproxy.emich.edu/10.1145/3231053.3234323

[14] Mbarek B., et al. "Enhanced Network Intrusion Detection System Protocol for Internet of Things." *In Proceedings of the 35th Annual ACM Symposium on Applied Computing*, 2020, pp. 1156–1163, doi:https://doi-org.ezproxy.emich.edu/10.1145/3341105.3373867

[15] Aneetha A., et al. "Hybrid Network Intrusion Detection System Using Expert Rule Based Approach." *In proceedings of the Second International Conference on Computational Science, Engineering and Information Technology*, 2012, pp. 47–51, doi:https://doi-org.ezproxy.emich.edu/10.1145/2393216.2393225