Protecting Accounting Information Systems Using Machine Learning Based Intrusion Detection

D. Naik
Eastern Michigan University

M. Krishnappa
Eastern Michigan University

K. Kaur
Eastern Michigan University

A. Almohsen
Eastern Michigan University

P. Meharia
Eastern Michigan University

B. Panja
Eastern Michigan University

The key goal of this paper is to look at network data and identify whether it is normal traffic data or anomaly traffic data. 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 detect whether the network data is normal or an anomaly. 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, Logistic Regression Classifier and Support Vector Classification algorithms are used in our model to support proposed concepts. A feature selection method is used to collect valuable information from the dataset to enhance the efficiency of the proposed approach. The experimental findings revealed that the suggested method has a neglected false alarm rate, with the accuracy expected to be between 95% and 100%.

Keywords: machine learning, intrusion detection, classification algorithm

INTRODUCTION

The Internet is being the most obvious choice for everyone to transfer data. Like everything these days, depends on the internet. The Internet spreads all over the world. 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. 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. 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.

In this study, we have developed a model to detect network-based intrusion. 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. 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. 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.

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. 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 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. 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. 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.

EXISTING WORK

The key objective of this research is to handle and reduce the rising number of network traffic warnings. To resolve the increasing number of network traffic alarms, the researchers have proposed a system that integrates intrusion detection and the exploration of information to decrease the number of alarms of attacks on the network traffic. They claimed that the proposed approach is effective for many vulnerable activities, such as assaults using brute force, weaknesses recognition, identification of protocol irregularity, and many more.

After reviewing current literature, the researchers found out that firewalls and anti-virus applications typically aim to suppress attacks, and Intrusion Detection Systems (IDS) attempt to detect vulnerabilities as they arise. That is the reason anti-virus systems, firewalls, and IDS are not capable of dealing with a broad diversity of suspicious activities on the network. Moreover, because of the globalization of internet connection, network security came in the light of research and improvement. An IDS is a software application that controls malicious attacks on the network. Currently, new network security technology is known as Intrusion Prevention System. It is a network protection approach to detect suspicious behavior on the network or system. It also acts in real-time to avoid or prevent malicious activities. However, its major drawback is the detection of false positives or false negatives alarms of network activities. As a result, the authors implemented two concepts together (IPS and Data Mining) to provide security to the network. After reviewing Nguyen, and Nguyen's article, we found that the proposed approach is strong enough to resolve the issue. Although, this is just a theory. The authors have not done any experiment or implementation to test their approach. So, we are not sure that this method will succeed in the real situation or not.

We need normal data and attack data to implement machine learning techniques for intrusion detection. Collecting normal data is a straightforward process, but collecting attack data is very challenging and expensive. The primary purpose of the study is to propose a solution that can produce synthetic attack data for machine learning approaches.

Nowadays, network-based intrusion detection is the most demanding system due to the extensive use of network traffic. According to researchers, if we want to research and develop machine learning-based network-based intrusion detection systems, we need huge amounts of data to train the machine. Apart from that, we also need up-to-date data to compete in the real world. In this study, the authors have proposed a system that can generate artificial attack data. They have proposed a method by focusing on two situations. In the first case, no attack data gathered from the network. So, in this situation, the suggested algorithm will generate attack data based on collected normal data. In the second case, the number of attack data is small. In this situation, the projected algorithm will create more artificial attack data based on collected attack data. This artificial data method has not only been suggested but also tested and the majority of the machine learning techniques can obtain satisfactory outcomes on the synthetically produced data according to experimental results. The researchers have also claimed that many developing studies have emerged from the proposed technique. In the future, they are going to focus on the efficiency of the method by comparing similar approaches. This approach sounds useful for startups and also useful even after reaching at some point when we do not have enough data to train machines for network-based intrusion detection.

The primary focus of the study is the flaws of the current intrusion detection system such as; insufficient accuracy, time-consuming and high rate of false alarms. To solve these problems, the authors have proposed a fusion approach for intrusion detection, which is based on decision tree and k-Nearest Neighbor classifiers.

The increasing use of the Internet also boosted the quantity and quality of malicious attacks. The researchers believed that old techniques like firewalls and anti-virus cannot defend against modern attacks. That is the reason for the requirement of an accurate and smart intrusion detection system. The authors have proposed an intrusion detection system that is the mix of two leading machine learning models (DT & k-NN). Along with that, they have used feature selection methods to upgrade the performance of the presented system. In this approach, they have used k-NN and DT as data classifiers and using voting techniques to merge DT and k-NN results. They have also done implementation and tests and as a result, they found that the false alarm rate of the proposed method was 0.2%, which is negligible. Also, they have examined that accuracy and positive detection rate were between 99% & 100%. The proposed approach seems attractive to us because of its accuracy and positive detection rate.

The authors discussed that an enormous number of people are using the internet in recent years and that will impact the security of devices. The system that discovered - intrusion detection system (IDS) is to find the unusual action in the user's devices. There are some features to detect that intrusion such as ongoing operation, afford errors, any ruin will try to keep it out, and can be adjustable. However, these systems work on devices and give a warning if there is any intrusion.

The researchers have expressed the algorithm to expose an intrusion it builds on deep learning to power networks. They found a program to detect auto-encoder-extreme the learning machine to produce data also checked its effectiveness. This manufacturing is utilized to control the system.

In 2010, the Stuxnet Virus had been a revolution in the global industry since that time there were offensive aversions to industrial control systems, and several serious security accidents happened too. Because of the increase of network attacks which is unlimited, the national security, economic development and social stability and the public are faced with a large warning. However, the security issues of industrial control networks can be brief in few points:

1. Intranet security of industrial enterprises.
2. Manufacturing resource access security
3. Network boundary security
4. Identity authentication security
5. ID resolution security

This study is done to confirm that this process can develop the system and minimize the risk. By merging the deep learning theory and the extreme learning machine in the neural network. Since the accuracy will be more precision. Also can be developed for detection production.

The authors discussed critical services highly determined by Industrial Control Systems (ICS). Since that it needs to be improved, and with advances in technology the ICS developed its features. This development provides strict security performance and risk analysis because this is a big challenge for that system. They also suggest an algorithm and attack diagram confirm between cyber and physical domains system. For the analysis and attack diagrams after operating the allowing to reach analysis on DFG S, will have a set of attacks.

In addition, that will impact Cy2Phy interfaces. For instance, Pump 205 operates that has pair signals, and for that cause, the attacker needs a minimum of two attack points.to trade off the target successfully. The attack diagrams can be utilized to adopt a double schema that takes in a various number of attack points for choosing targets. That's will lead to a decrease in the explore area to those that have more Opportunities to attack targets. In the end, in the future, the authors will search for assorted information that allows them to conclude the attack vectors in an ICS.

The authors researched for developing models for physical attacks in cyber-physical systems. They do examine the security of cyber-physical systems by utilizing Adversary View Security Evaluation (ADVISE). To create a system they have to characterize the system components and the relationships between them.

The cyber-physical systems have security risks that need to be examined, and that will be through designing the cyber aspect.

So, the central point of his study is designing physical attacks and their outcomes. However, the practitioners require to catalog the potential attack steps for each system component. And, that will be by recognizing the biggest threat to the system, to define that will be by the total cost of implementing attacks and the number of steps. Every attack step has its own cost. In the end, the creator many times rates the security vulnerabilities of cyber-physical systems without thinking of the negative effect of the system.

The main focus of this study is to help administrators to make better decisions on selecting proper hardware for Intrusion detection in various environments. Here researchers are focusing on measuring time to build a model as a performance measurement for detecting network intrusion based on various supervised machine learning algorithms such as.

- Decision tree
- Support vector machine (SVM)
- Multi- Layer Perceptron (MLP)

Google cloud platform was used to perform experiments with Windows server 2016 OS, Intel CPU processor and RAM 16.00 GB. Data source used for the study was the “NSL-KDD dataset” WEKA-3.6 is used as a data mining tool to validate and test the algorithms. Experiments were conducted on multiple scenarios and various sets of data sources (starts with 10000 records and increases every 10000 until 120000 records) run across 1,2,4,8 & 16 core CPU. From results of the experiments performed on all 3 algorithms, authors came up with the following conclusion.

Increasing the number of CPU hardly effects the process time. Based on the space plots of Decision tree and SVM when the number of CPU is greater than 2, there was no change in process time. Multi-layer perceptron Model (MLP) was the slowest among the three algorithms as the process time for building the model was linear function of workload and the number of CPU used, whereas in other two models process time for building the model was exponentially function of workload and the number of CPU used. Even though in the Decision tree model and SVM process time for building the model was exponentially function of workload and the number of CPU, Decision tree was preferred because the time taken to create the model in a larger workload scenario was lesser than SVM.

The authors of this paper compare all of Weka's machine learning algorithms to the standard data set for intrusion detection, to be specific kddCup99 and they are essentially concentrating on the examination of intrusion detection in systems utilizing different expectation methods such as Random forests, One R, Naïve Bayes, and Multilayer Perception beneath supervised learning.

Weka is a set of information analysis machine learning methodologies specifically to information from the data named from the Java program. The emphasis of this paper is on a multi-class classifier, with a dataset that includes both mathematical and conceptual features, as a result, only certain machine learning methods that are suitable for multi-class identification and charge on both functional and mathematical factors can be considered. The performance matrix further used to test the strategies under examination is derived from the confusion matrix. The tests were conducted on a complete data set with training data and a limited data set with just 11 feature characteristics. According to the findings, “classification methods do not depend on all 41 traits, so the methodology may possibly get excellent performance with a significant reduction in resources required by running on the same dataset with a smaller dataset.”

Some of the priority areas of studies in the field of networks and communication protection has been the method of intrusion detection using text analysis. The device calls are used as a basis for mining and forecasting any possibility of attack in this method to intrusion detection.

The authors use text mining to conduct an exhaustive survey on intrusion detection and verify the worthiness of different feature interventions obtained in this study. We assume that the results of this poll will be beneficial for authors focusing on intrusion detection utilizing text analysis. The misuse-based intrusion detection system comes first, followed by the anomaly-based intrusion detection system. It is essential to build a base of knowledge for a misuse-based intrusion detection system in order to determine if an incoming message is natural or intrusive. If there is an attack, the intrusion detection system will trigger an alert. The second approach is anomaly-based, in which the intrusion detection system observes the device's actions and generates a warning automatically if it deviates from normal activity. This paper explores the different ways that can be used to detect an intruder. It also goes into the testing questions that should be considered when using text analysis to detect intrusions.

Network intrusion monitoring is important for detecting internet in potential threats, where AI can be used to capture and analyze network traffic in real time. The code will use the Random Forest machine learning technique to provide more precision in discovering new attacks and ensuring that the system is well trained to detect them in the future.

With modern technological advancements and the advent of concepts such as Cloud Technology, Mobile Technology, and Deep Learning, a massive volume of data is exchanged over a server, and intrusion detection mechanisms must continually capture and interpret this data to identify potential risk and computer hackers. Machine learning methods are used to recognise and respond to external attackers through identifying attributes of data traffic to determine if it is suspicious or not. AI is an evolving technology that consists of many strategies used in judgment processes, such as machine learning algorithms applied to recognize and respond to external attackers by classifying attributes of network activity to determine whether it is suspicious or not. All of the approaches outlined above can be integrated in a hybrid AI design to create the next wave of research and intellectual ability technologies in fields as diverse as science, medicine, and cybersecurity. The aim is to provide a scalable security framework that can be learned to anticipate the occurrence of a new form of attack in network traffic if it eventually happens. The remaining paper is structured in a way that narrowly discusses structures and strategies in the research that combine IDS and AI. Besides that, the computer new framework host-based intrusion system is implemented to identify anomalies and secure the application's data. A further method is suggested, in which the authors used Recursive Function Inclusion in a network-based intrusion detection scheme and other techniques to encrypt message string models to improve feature extraction. Our suggested solution is intended to automatically classify potential threats and track intrusions in network activity. The outlying identification challenge occurs when something is hard to determine a network's class attribute when it differs across clients and programs. Machine learning is an AI technology that allows machines to develop on their own. Furthermore, we built an ad blocking software to identify messages and avoid spam from accessing the network.

To provide secure and trustworthy computing network intrusion detection use many tools, techniques, and strategies to detect threats. There are two techniques – signature pattern matching and Anomaly detection. Signature pattern matching has a sequence of known threat events if a new event matches with the threat pattern then it is processed for countermeasures. Anomaly detection is based on the unknown attack signature. Normal events are tested using artificial intelligence and prediction model techniques to identify threats in the cyber system then again it checks by signature pattern. In this paper, they define two formats- network intrusion detection and host intrusion detection system.

Cyber enabled infrastructure which is controlled, and access remotely can be damaged by threat attacks. There are two types of attacks -passive and active. Which can be detected by IDS. There are different types of IDS. To handle these attacks, they set up an environment with intrusion detection system (IDS) where a number of homogeneous computer software needs just one IDS and another needs to match with a server. If this deals with no end-to-end encryption then after decryption if any threat attacks then it cannot solve it and then it needs Host intrusion detection system (HIDS). There are two elements to solve it. The first is a firewall which can distinguish traffic between end points with some rule-based policy. Another is anti-malware software which checks an application with a known signature pattern. When antivirus cannot find attacks then IDS comes where all the data passed through the detection engine. HIDS checks traffic at host level and log traffic. Another IDs is signature based which is very fast and provides accurate results. Drawback of this detection is fragmentation and it always needs to be updated. Host ID can detect encrypted data which is missed by Network IDS. Thus, HIDS is an important intrusion detection system to find attacks and detects a variety of intrusion in a system including cryptographic type intrusions.

Intrusion detection system used to keep all data together and system occurrence from threats. Where it studies all fields to get harmful things related knowledge that occurs while testing and checking for attacks. Data mining has its own different kinds of techniques which are used to find attacks in networks like clustering and classification. It used very popular and important methods to convert large sets of forms into small sets and save the space and memory. And also keep the original data without loss. Thus the internet on things intrusion detection gets knowledge about network attack by SVD technique. This technique is applied to separate the tough variables for each class independently. In this technique, input is given by a two-dimensional matrix and output is just a dot product and list of signature patterns. For training, resulting actions of all outside activities and their actions are registered.

As the need for the internet increases many devices are connected to servers to provide service. Thus, another problem of threat attacks, intrusion detection, and security is also a biggest challenge for network administrators. Here, they provide a new detection algorithm of the network of intrusion detection system (NIDS) protocol. It sends many test branches into the network to provide loyalty and increase the efficiency and achievement of previous network based detection systems. They provide information about a popular and dangerous attack is clone attack which is also known as node replication attack. This attack grabs some data, makes it duplicates and posts through the network. If these attacks are not found then it may get the location of antiviruses, duplicate data using the same properties, convert it with wrong information, and can also shut off the correct information. Thus, security is very important in internet related issues.

There are different kinds of attacks which are monitored by Network intrusion detection systems (NIDS). However, To detect duplicate data NIDS used many techniques- it starts with checking neighbor signals information and detecting the irregularity to detect abnormal activities on the internet, and check something out routing attacks such as cloning attacks. Second method provides the hierarchical trust discovery technique which confirms identification states, along with result performance based on the power from the root nodes. If they know that one node is successfully forward then another node trusts that node and continues forwards. It also works for machine learning problems. They offer a solution with which it can find hidden corrupted data. And also test the performance that after corruption when it is fixed, how it works. However, if the matching pattern data does not protect its carbon copy and sometimes send a report about the duplicacy, the system will not be able to find that virus. To overcome this kind of problems, without the knowledge of anyone about testing, witness nodes are chosen. This protocol takes more power while searching an unauthorized entrance and it saves more energy when they find an attack in the network.

To secure the transaction information and communication among people, it provides a new framework to catch the known and unknown incursion. Which is divided into different phases such as data collection, finding any missing information, rule based, data pre-processing, clustering techniques and expert system. This framework can detect attacks in an execution time with the link layer of the network by combining rule-based technique with clustering technique.

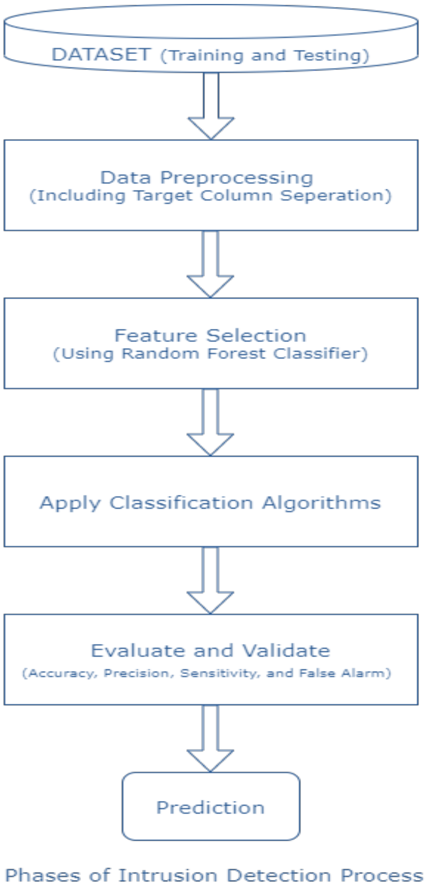
Using the knowledge of expert databases new attacks have been reforms in the rule base for missing information knowledge. It validates detection rates and reduces false alarm rates. Intrusion detection systems have two variants: misuse and anomaly-based approach. For misuse or signature, they code the rule for known attack patterns which gives a high detection rate; in an anomaly-based approach, any changes in normal behavior are considered as an attack. Performance of the system is estimated with detection rate. If the known interference is more than unknown then attack detection rate will be higher. When known interference falls off, detection rate also goes down. The detection rate of the combination of multiple software improved with comparison of misuse and anomaly individually. In future different protocols can be examined for some more unknown occurrences with some other techniques.

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.

FIGURE 1
INTRUSION DETECTION PROCESS


DATASET (Training and Testing)
Data Preprocessing
(Including Target Column Seperation)
Feature Selection
(Using Random Forest Classifier)
Apply Classification Algorithms
Evaluate and Validate
(Accuracy, Precision, Sensitivity, and False Alarm)
Prediction
Phases of Intrusion Detection Process

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.

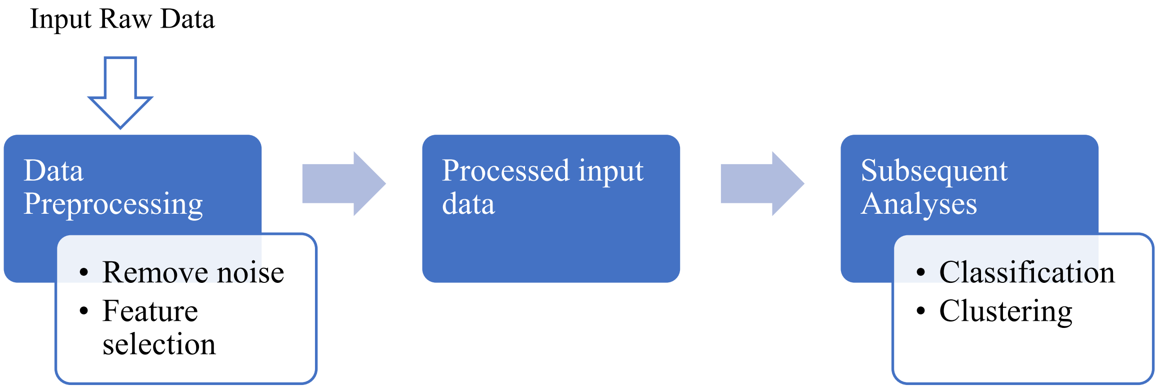
Classification algorithms are used when your output will be based on some categories, such as network data categories; a normal, anomaly, and so on. The term "classification" refers to the process of grouping production into distinct categories. Binary classification is where a process attempts to categorize 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.

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.

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.

FIGURE 2
DATA PREPROCESSING



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.

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.

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. 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.

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.

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.

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.

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.

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.

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.

FIGURE 3
SIGNIFICANT FEATURES

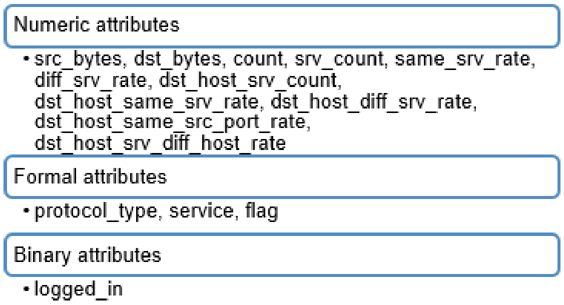
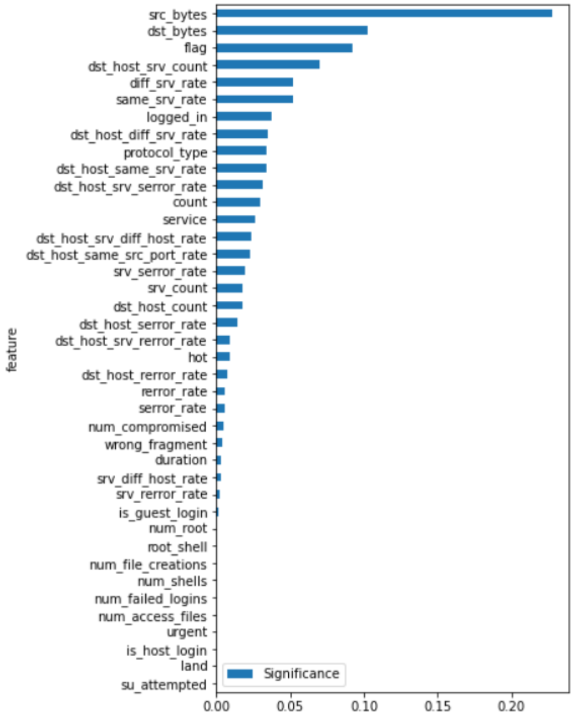

Numeric attributes
• src_bytes, dst_bytes, count, srv_count, same_srv_rate, diff_srv_rate, dst_host_srv_count, dst_host_same_srv_rate, dst_host_diff_srv_rate, dst_host_same_src_port_rate, dst_host_srv_diff_host_rate
Formal attributes
• protocol_type, service, flag
Binary attributes
• logged_in

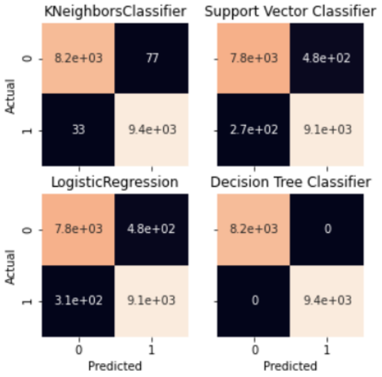
FIGURE 4
SIGNIFICANT FEATURES, TYPES



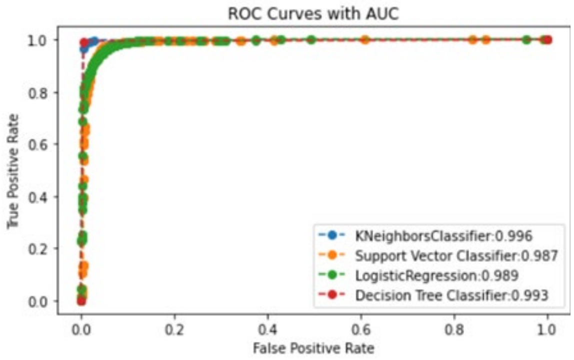
Implementation of the Classification Algorithms and Evaluation Result

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:

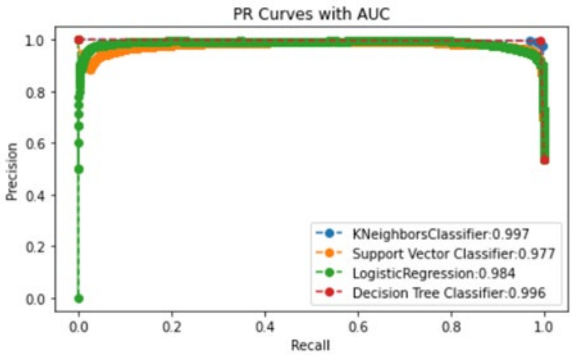
FIGURE 5
CONFUSION MATRIX AND HEATMAP



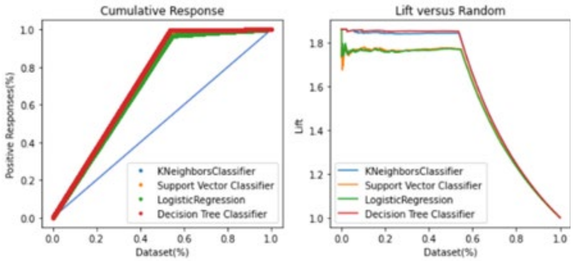
Classifier	***Accuracy***	***Precision***	***Recall***	***False Alarm***
k-NN	99.17%	99.8%	99.8%	00.98%
SVC	95.86%	96.9%	96.9%	04.76%
LRC	95.5%	96.7%	95.7%	04.90%
DTC	99.47%	99.7%	99.7%	00.37%



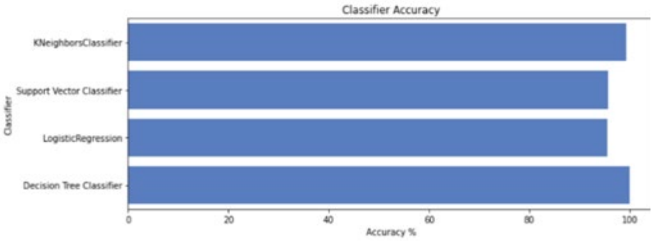
ROC Curve



PR Curve



Cumulative Response and Lift Curve

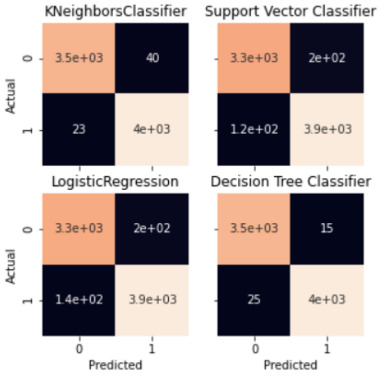


Classifier Accuracy

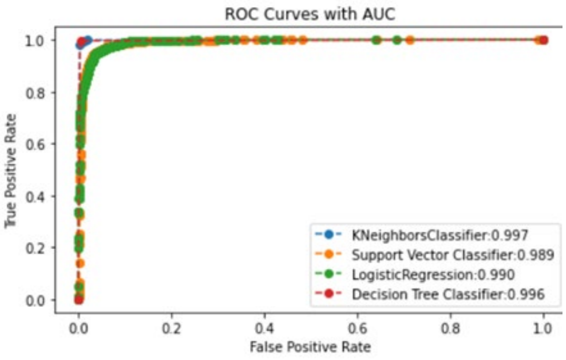
Validation Result

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:

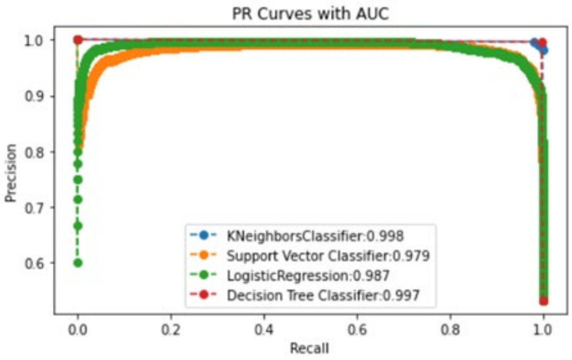
FIGURE 6
CONFUSION MATRIX AND HEATMAP



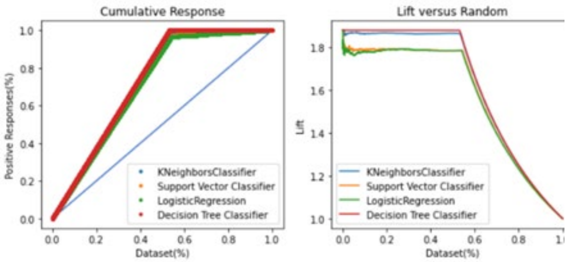
Classifier	*Accuracy*	*Precision*	*Recall*	*False Alarm*
k-NN	99.14%	99.7%	99.7%	00.82%
SVC	95.75%	96.7%	96.7%	05.0%
LRC	95.49%	96.4%	95.4%	05.04%
DTC	100.00%	100.00%	100.00%	100.00%



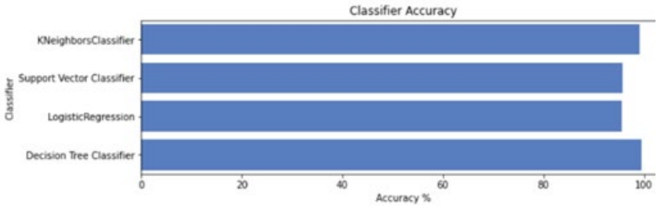
ROC Curve



PR Curve



Cumulative Response and Lift Curve



Classifier Accuracy

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.

CONCLUSION

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

Anand V. (2014). Intrusion Detection: Tools, Techniques and Strategies. In *Proceedings of the 42nd Annual ACM SIGUCCS Conference on User services* (pp. 69–73). https://doi-org.ezproxy.emich.edu/10.1145/2661172.2661186

Aneetha, A.S., Indhu, T.S., & Bose, S. (2012). Hybrid network intrusion detection system using expert rule based approach. In *Proceedings of the Second International Conference on Computational Science, Engineering and Information Technology* (pp. 47–51). https://doi-org.ezproxy.emich.edu/10.1145/2393216.2393225

Belouch M., & Hadaj S. (2017). Comparison of ensemble learning methods applied to network intrusion detection. In *Proceedings of the Second International Conference on Internet of things, Data and Cloud Computing* (pp. 1–4). https:// doi-org.ezproxy.emich.edu/10.1145/3018896.3065830

Chapke P., & Deshmukh, R. (2015). 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* (pp. 1–5). https://doi-org.ezproxy.emich.edu/10.1145/2743065.2743128

Foroushani A., & Li, Y. (2018). Intrusion detection system by using hybrid algorithm of data mining technique. In P*roceedings of the 2018 7th International Conference on Software and Computer Applications* (pp. 119–123). https://doi-org.ezproxy.emich.edu/10.1145/3185089.3185114

Hamid, Y., Sugumaran, M., & Journaux, L. (2016). Machine Learning Techniques for Intrusion Detection. In *Proceedings of the International Conference on Informatics and Analytics*, (Article 53, pp. 1–6). https://doi-org.ezproxy.emich.edu/10.1145/2980258.2980378

Jafier, S. (2018). 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* (pp. 1–3). https://doi-org.ezproxy.emich.edu/10.1145/3231053.3234323

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

Kumar, G.R., Mangathayaru, N., & Narasimha, G. (2015). Intrusion Detection Using Text Processing Techniques. In *Proceedings of the International Conference on Engineering & MIS*, (Article 55, pp. 1–6). https://doi-org.ezproxy.emich.edu/10.1145/2832987.2833067

Li, Y., Li, Y., & Zhang, S. (2019). 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* (pp. 40–44). https://doi-org.ezproxy.emich.edu/10.1145/3372047.3372092

Mbarek, B., Ge, M., & Pitner, T. (2020). Enhanced network intrusion detection system protocol for internet of things. In *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, (pp. 1156–1163). https://doi-org.ezproxy.emich.edu/10.1145/3341105.3373867

Nalavade, K., & Meshram B. (2010). Intrusion prevention systems: Data mining approach. In *Proceedings of the International Conference and Workshop on Emerging Trends in Technology* (pp. 211–214). https://doi-org.ezproxy.emich.edu/10.1145/1741906.1741952

Pham, T.S., Nguyen, Q.U., & Nguyen, X.H. (2014). Generating artificial attack data for intrusion detection using machine learning. In *Proceedings of the Fifth Symposium on Information and Communication Technology* (pp. 286–291). https://doi-org.ezproxy.emich.edu/10.1145/2676585.2676618

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

Yu, Y., Liu, X., & Chen, Z. (2018). Attacks and Defenses towards Machine Learning Based Systems. *Proceedings of the 2nd International Conference on Computer Science and Application Engineering* (pp. 1–7). https:// doi-org.ezproxy.emich.edu/10.1145/3207677.3277988