symmetry

Article

Machine-Learning-Based DDoS Attack Detection Using Mutual Information and Random Forest Feature Importance Method

Mona Alduailij [1], Qazi Waqas Khan [2,*], Muhammad Tahir [2], Muhammad Sardaraz [2,*], Mai Alduailij [1] and Fazila Malik [2]

[1] Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah Bint Abdulrahman University, Riyadh 11671, Saudi Arabia; maalduailej@pnu.edu.sa (M.A.); maalduailij@pnu.edu.sa (M.A.)

[2] Department of Computer Science, COMSATS University Islamabad, Attock Campus, Attock 43600, Pakistan; m_tahir@cuiatk.edu.pk (M.T.); sp20-rcs-002@cuiatk.edu.pk (F.M.)

* Correspondence: sp20-rcs-006@cuiatk.edu.pk (Q.W.K.); sardaraz@cuiatk.edu.pk (M.S.)

Abstract: Cloud computing facilitates the users with on-demand services over the Internet. The services are accessible from anywhere at any time. Despite the valuable services, the paradigm is, also, prone to security issues. A Distributed Denial of Service (DDoS) attack affects the availability of cloud services and causes security threats to cloud computing. Detection of DDoS attacks is necessary for the availability of services for legitimate users. The topic has been studied by many researchers, with better accuracy for different datasets. This article presents a method for DDoS attack detection in cloud computing. The primary objective of this article is to reduce misclassification error in DDoS detection. In the proposed work, we select the most relevant features, by applying two feature selection techniques, i.e., the Mutual Information (MI) and Random Forest Feature Importance (RFFI) methods. Random Forest (RF), Gradient Boosting (GB), Weighted Voting Ensemble (WVE), K Nearest Neighbor (KNN), and Logistic Regression (LR) are applied to selected features. The experimental results show that the accuracy of RF, GB, WVE, and KNN with 19 features is 0.99. To further study these methods, misclassifications of the methods are analyzed, which lead to more accurate measurements. Extensive experiments conclude that the RF performed well in DDoS attack detection and misclassified only one attack as normal. Comparative results are presented to validate the proposed method.

Keywords: machine learning; mutual information; random forest; DDoS; cloud computing

check for updates

Citation: Alduailij, M.; Khan, Q.W.; Tahir, M.; Sardaraz, M.; Alduailij, M.; Malik, F. Machine-Learning-Based DDoS Attack Detection Using Mutual Information and Random Forest Feature Importance Method. *Symmetry* **2022**, *14*, 1095. https://doi.org/10.3390/sym14061095

Academic Editors: Minxian Xu and Kuo-Hui Yeh

Received: 31 March 2022
Accepted: 23 May 2022
Published: 27 May 2022

1. Introduction

Cloud computing is an Internet-based platform that delivers computing services such as servers, databases, and networking, to users and companies at a large scale, and helps an organization in reducing costs, in terms of infrastructure [1]. A Distributed Denial of Service (DDoS) attack is used by attackers to prevent legitimate users from accessing the services [2]. In this attack, a very high load is put on the victim server by the attackers, by providing multiple requests to the server. This huge number of requests by the attackers fill the bandwidth of the victim server, which makes it unavailable to legitimate users [3]. The DDoS attack is a brute-force attack that affects the devices of the network with malware using Botnet. There are three main categories of DDoS attacks, on the basis of target and behavior. These are bandwidth attacks, traffic attacks, and application attacks. In traffic-based attacks, attackers send a huge volume of TCP or UDP packets to the victim server, and this large number of packets reduces the overall performance of the victim server. The attackers send a large amount of anonymous data in a bandwidth attack and create congestion, by consuming more bandwidth. The application attack is used by attackers to attack a specific system, and it is difficult to mitigate [4]. To detect DDoS, attack-machine-learning-based prediction models are used.

In this modern era of technology, machine learning is an emerging field and has many applications in solving different real-world problems, such as medical images [5], sentiment analysis [6], and cloud-resource-utilization prediction [7]. Machine learning is, also, used in intrusion detection in cloud computing [4,8,9]. The researchers proposed various methods for developing intrusion-detection systems in a cloud environment. Self-adaptive evolutionary extreme learning is used to detect DDoS attacks [9]. Authors in [10] detect a DDoS attack using a Deep Neural Network (DNN), whereas a deep belief neural network is, also, used [11]. The accuracy of the different methods available in the literature is impressive, on different datasets.

In this article, we propose a DDoS-attack-detection method, using different feature-selection and machine learning methods. We use Mutual Information (MI) and Random Forest Feature Importance (RFFI) methods, to select the most relevant feature from CICIDS 2017 [12] and CICDDoS 2019 [13] datasets. K-Nearest Neighbor (KNN), Logistic Regression (LR), Random Forest (RF), Gradient Boosting (GB), and Weighted Voting Ensemble (WVE) methods are used for attack detection. These methods perform better in intrusion detection [14]. The performance of the proposed method is evaluated using precision, recall, F measure, and accuracy. The results show that the proposed method performs better, in terms of accuracy with reduced miss classification errors, compared to existing methods. The main proposal of this study is to reduce miss classification errors in DDoS attack detection, by selecting the most relevant features and performing parameter tuning of the machine learning model.

Key contributions are as follows.

1. In this study, experiments are performed with tree-based methods (RF, GB), distance-based methods (KNN, WVE, and LR), and models based on the CICIDS dataset.
2. This study uses MI and RFFI methods for extraction of the most relevant features.

The rest of the paper is organized as follows. Section 2 presents different methods for intrusion detection, followed by the proposed methodology in Section 3. Sections 4 and 5 present the results and conclusion, respectively.

2. Literature Review

Data security is a widely studied field in computing domain. Many applications use security for different purposes, including access control [15], network security [16], data security [17], availability of services [4], etc. Both symmetric and asymmetric approaches are used, according to the targeted domains. This section covers related work on more relevant domains, i.e., intrusion detection. The field of intrusion detection is widely studied in the literature. Different machine learning approaches are available. Some review articles cover the use of machine learning in cloud computing [18]. Authors in [9] detect DDoS attacks using self-adaptive evolutionary extreme learning. The method has two important features, i.e., the detection of the best crossover operator and the automatic detection of the neurons of the hidden layer. The proposed method is evaluated, with experimental results, which show improved accuracy. Another technique presented in [10] detects DDoS attacks in Software Defined Networks (SDN). The authors used DNN for real-time detection of DDoS attacks. Experimental results show that this method detects DDoS attacks, with better accuracy in less time with less resource usage. Authors in [19] compared machine learning methods for detecting DDoS attacks. Experimental results show that RF detects DDoS attacks with better accuracy. In another study, authors use [20] correlation, information gain, and the relief feature selection method, to select the most relevant features for DDoS attack detection. Comparison of different machine learning methods is presented. Manimurugan et al. [11] developed an intrusion detection system, to detect anomalies in Internet of Things (IoT) systems. They used deep belief network model for attack detection. The experiments are performed on the CICIDS 2017 attacks dataset. The proposed method demonstrates 99.37% accuracy in detecting normal class and 96.67% for DDoS attack detection. To detect all DDoS attacks, Dehkordi et al. [21] presented a model, which detects DDoS attacks in SDN. The proposed model consists of three modules: one is collector, the second is entropy-based, and the third is classification. Three datasets (CTU-13 [22], ISOT [23], and UNB-ISCX [24]) are used for evaluation.

High dimensional data needs huge computing power for processing. For a high dimensional dataset, identification of relevant features plays an important role. Authors in [25] use chi-squared, information gain, gain ratio, ReliefF, and correlation, for selecting the most relevant features from the intrusion dataset for web-attack detection. Web attacks are detected using j48, with 10-fold cross-validation. Experimental results conclude that the j48 with MI features selection method achieved the highest accuracy in web-attack detection. Another study [26] uses Genetic Algorithm (GA) and Principal Component Analysis (PCA), for feature selection from intrusion datasets. Attacks are detected using a Decision Tree (DT) classifier. The experimental results show that PCA-GA with decision tree achieved improved accuracy. Authors in [27] select more crucial features for intrusion detection using the MI, consistency, correlation, and distance methods. The output of the four methods is combined to get the potential features set. The comparison of RF, Naïve Bayes (NB), Support Vector Machine (SVM), Linear Discriminant Analysis (LDA), and DT, with different feature selection methods, is carried out. The experiments were performed on a benchmark intrusion detection dataset.

Another study [28] reduces the execution time and computational resources, by selecting the most relevant 22 and 52 features, using the MI selection method for the intrusion detection system. The experiments were performed on the CICIDS-2017 dataset. Experiment results show that RF achieved the highest accuracy with 22 features, and j48 achieved better accuracy with 52 features. Mohamed et al. [29] proposed the intrusion detection system for IoT networks. The authors use rule-based and decision-tree-based methods, i.e., the JRip algorithm, REP Tree, and Forest PA. The first and second methods are to train on the feature of a dataset and detect the attacks. The third method takes the input from both the original dataset and output of the first and second methods. The experimental results show that this method improves the existing results, in terms of accuracy and detection rate. Authors in [30] proposed an intrusion detection system, to detect attacks in networks. They select the most relevant features, using the NB feature embedding method, and use SVM for the classification and detection of attacks. The experiments were performed on different publicly available datasets for intrusion detection.

To identify malicious traffic and link failure attacks, authors in [31] proposed a novel model, to discover novel features for detecting DDoS attacks. Novel features are listed in a CSV file, and machine learning algorithms are trained on SDN datasets. Non-SDN datasets are used in several studies. The hybrid machine learning model is used to classify data. According to the result, a support vector classifier with the RF hybrid model classifies traffic, with improved testing accuracy and a very low false-alarm rate. Tonkal et al. [32] developed a method that combined machine learning algorithms with Neighborhood Component Analysis (NCA), to classify SDN as normal or a traffic attack. The SDN dataset is used to identify DDoS attacks. For feature selection, the NCA algorithm is used. After pre-processing and feature selection, datasets are classified using classifiers SVM, decision tree, and KNN. Experimental results show that decision trees outperform other algorithms, in terms of accuracy. To categorize network traffic as malicious or benign, authors in [33] proposed a new method for visualizing network activity, using Convolutional Neural Network (CNN) and graphical heat maps. The results of the proposed method are compared with two models, i.e., Long Short-Term Memory (LSTM) and SVM. Based on the results, it is concluded that using CNN to explore network traffic, via graphical heat maps, offers accurate botnet-based DDoS attack detection.

For DDoS attack detection, M. Revathi et al. [34] present a discrete-scalable memory support vector machine (DSM-SVM) and SDN-mitigation framework. Using the spark standardization method, input data is pre-processed, and all unwanted and missing values are removed. The semantic multi-linear component analysis algorithm is used for feature extraction. The proposed DSM-SVM algorithm is used to predict attacks, with high accuracy. The proposed model is trained and used in SDN mitigation and detection. The results show that the presented model surpasses the results of other algorithms, with improved accuracy. Neural network models are, widely, used in developing intrusion detection systems for detecting cyber-attacks. Authors in [35] proposed a DNN-based efficient hybrid method, for anomaly detection in a cloud environment. Improved Genetic Algorithm (IGA) and Simulated Annealing Algorithm (SAA) optimization methods are used to optimize the values of the DNN parameters. Experimental results conclude that this method improves accuracy in anomaly detection.

Another study [36] uses the DNN model to explore the detailed analysis of various intrusion datasets. First, the method trains DNN on the KDDCup 99 attack dataset [37] and learns the hyperparameter of DNN. Then, it applies DNN with the same parameter on another well-known attack dataset . The experimental results show that this method performs well on the CICIDS 2017 [12] dataset. Wenchao et al. [38] select the most relevant 49 features, using the recursive feature elimination method and the proposed attack detection method, based on LeNet-5 CNN. In their architecture, they remove the first pooling layer and the last fully connected layer, to reduce the computational cost. The experimental results show that this method detects attacks with better accuracy.

Authors in [39] use artificial neural network to detect attacks in a cloud environment. The method detects the attacks. with improved accuracy in detecting multiple attacks. To strengthen SDN, authors in [40] proposed an intrusion detection system using a Gated Recurrent Unit Recurrent Neural Network (GRU-RNN). The experiments were performed on NSL-KDD [41] and CICIDS 2017 [12] datasets. The experimental results show that GRU-RNN detects attacks with better accuracy on both datasets.

Hanane et al. [42] use DNN for intrusion detection in SDN. As the number and features of network traffic increase dramatically, traditional machine learning classification of DDoS attack algorithms has become unsuccessful, due to their inability to automatically extract important features. To overcome this limitation, Wei et al. [43] developed a hybrid AE-MLP method, for successful DDoS attack classification. The proposed AE-MLP model component, AE, gives an optimal feature extraction, by identifying the most significant feature sets without the need for human assistance. The multilayer perceptron network component of the proposed model is used to overcome speed and bias issues, which occur with processing large feature sets with noisy data. Experimental results show that the proposed method has high accuracy, surpassing other existing methods. A robust detector for DDoS attack detection, using the Generative Adversarial Network (GAN), is proposed in [44]. The proposed model can detect attack instances that are closer to real scenarios. The network traffic generated by the adversary can be identified using the proposed model. To mitigate the higher-order differential power-analysis attack, a model based on evolutionary computation is proposed in [45]. First, GA is used to split the content into nonuniform shares. Then, the shares are used to compute individual modular components, using the nearest neighborhood algorithm. Authors in [46] detect and classify different types of network flows using machine learning. The proposed model consists of host-intrusion-detection and network-intrusion-detection systems . Comparative experimental results are presented to validate the proposed algorithm.

The literature review shows that the researchers detect the DDoS attack by using complete feature sets of the selected datasets, and some studies performed the detection using other feature selection methods. This study uses the MI and RFFI methods, for the selection of the most relevant features. The existing methods have missed classification errors, and this study reduces the miss classification error, by using MI and RFFI techniques, with different classifiers.

3. Materials and Methods

In this section, the steps of the proposed methodology for DDoS attack detection are discussed. In the first step, we extract the CICIDS 2017 [12] and CICDDoS 2019 [13] datasets . The preprocessing of a dataset is performed in the second step. In the third step, we apply machine learning techniques for the classification of DDoS attacks. Finally, we evaluate the performance of our method, by different evaluation metrics. Figure 1 shows the workflow of the proposed methodology for DDoS attack detection. Each phase of the proposed method is explained in the following subsections.

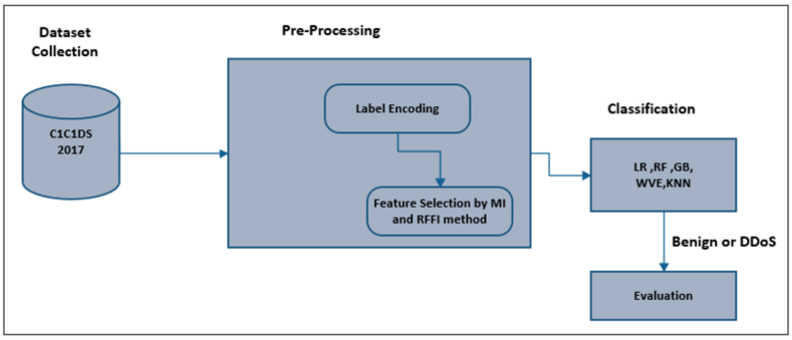

Figure 1. Architecture of the proposed DDoS attack detection model.

3.1. Datasets

The CICIDS 2017 and CICDDoS 2019 datasets are extracted from the respective websites [12,13]. The CICIDS 2017 dataset consists of 3.1 million traffic flow records [47]. This data set has 5 days of log files of traffic flow, Monday to Friday. We experimented on the Friday evening log file of network traffic. This log file has 225,711 instances and 79 features, including the class label. One file (DrDoS_NTP) is selected from the CICDDoS 2019 dataset. The file contains 1,209,961 instances and cleans 84 input features.The class attribute is a binary class label and has two classes, i.e., benign and DDoS. Benign is a normal class, and DDoS is an attack. The other log file contains the dataset of other attacks, and the subject of this study is detecting the DDoS attack. Many other studies in the literature use the same data, for DDoS attack detection. The dataset consists of a large number of samples data that makes it suitable to evaluate the detection accuracy.

3.2. Data Preprocessing

Data preprocessing is a process of converting raw data into a useful form. Convert the categorical class label into discrete form (0,1), by applying label encoding, where 0 is a benign class and 1 is a DDoS attack.

3.2.1. Feature Selection

The selected datasets are high dimensional, and the high-dimensional data increases the training, exponentially, as the dimension of data increase. Different studies have used feature selection on selected dataset for different attacks's detection [20,25]. The second problem with a high-dimensional dataset is that it increases the risk of model overfitting. Several feature selection techniques are used to select the most relevant features from in-hand features. The field of machine learning and data mining has been widely studied in feature selection. A feature is known as an attribute or system that has been evaluated. The goal of feature selection is to find the best feature subset of k features, which cause the least amount of generalization error [48]. There are three main types of feature selection, namely the filter-based method, wrapper method, and embedded method. The filter-based method computes the importance of features, by considering the relationship between the input features and the target attribute. THe wrapper method generates a model on the subset of features and evaluates the performance of the model. The wrapper method takes more time for high-dimensional data, with millions of instances. The embedded method selects the features, by using the insight provided by some machine learning models. MI is a filter-based feature-selection method. The advantage of using MI, compared to other filter-based methods, is that it works well in case of a nonlinear relationship between the input features and the target attribute. The RFFI method is an embedded feature-selection technique. The purpose of using the RFFI method is to give a better result, compared to other embedded feature-selection methods. The reason behind using the MI and RFFI methods is to find the best feature-selection method for intrusion detection, from the filter-based and embedded methods. The main objectives of feature selection are listed below.

(1) Improve generalization performance, when compared to a model with all characteristics.
(2) Provide more robust generalization and faster reaction to unseen data.
(3) Gain a better and simpler understanding of the data-generation process.

The feature-selection approach is used as a preprocessing step, in regression and classification.

3.2.2. Mutual Information

The amount of information that one random variable knows about another random variable is known as MI. Feature selection allows to quantify the importance of a feature subset, in relation to an output vector [49]. Equation (1) shows the calculation of MI.

$$I(X;Y) = H(X) - H(X|Y) \tag{1}$$

where $I(X;Y)$ is MI for X and Y, $H(X)$ is entropy for X, and $H(X|Y)$ is a conditional entropy for attributes X and Y.

3.2.3. Random Forest Feature Importance Method (RFFI)

RF is an ensemble-learning algorithm that grows many decision trees, independently, and combines the output. Decision trees consist of internal and leaf nodes. The selected features are used to make a decision in the internal node, and it divides the dataset into two separate sets, with similar responses. The features in an internal node are selected by the Gini impurity criterion. The feature that has the highest decrease in impurity is selected for the internal node [50].

3.3. DDoS Attack Classification

The following subsections present details of the classification models used. Each model has different parameters that require tuning to achieve better results. This study uses Grid Search (GS) for this purpose.

3.3.1. Logistic Regression

Logistic regression is a machine learning technique that can be used for classification problems. Logistic regression works well on the binary class label. In LR, weights are multiplied with input and pass them to the sigmoid activation function [51]. In the proposed work, we apply LR on selected features for DDoS attack detection. The weights are optimized, using the lbfgs optimizer with C = 0.2.

3.3.2. K Nearest Neighbor

KNN is a classification approach that classifies test data observations, based on how close they are to nearest class neighbors. KNN is used as a semi-supervised learning approach, and KNN is used to identify the nearest neighbors [52]. It is based on a non-parametric approach to classify samples. The distance between separate points on the input vector is determined, and the unlabeled point is, then, allocated to the neighboring class K. K is the main parameter in the KNN classification. If K is large, the prediction neighbors will take a long time to classify, with an effect on prediction accuracy [53]. KNN is easy to understand, when there are few predictor variables. For the creation of models with normal data types, such as text, KNN is used. We set the value of K as 2, by considering the 2 nearest neighbors, and the Minkowski distance metric is used.

3.3.3. Gradient Boosting

GB is one of the most popular prediction algorithms in machine learning [54]. Various ad hoc parameters are used to regulate the algorithm's decision tree evolution. Standard regulatory parameters control tree size and weight magnitude. This creates an optimization routine that is free of parameters. However, a variety of parameters are, mostly, used in training, to adjust tree size and shape. Regulation has shown useful results and makes the algorithm constant. Real extreme gradient boosting is a more regularized framework of GB, which has better control regarding the over-fitting issue [55]. As a result, it helps in the prevention of over-fitting in training data. It is linked to a developed set of tools, under the distributed machine learning architecture, due to its efficiency and improved performance. GB has certain parameters that are used in training for DDoS attack detection. Parameters used for GB are shown in Table 1. The parameters are selected on the basis of the GS method used for parameters tuning.

Table 1. Hyper Parameter of GB.

Hyper Parameter Name	Values
Learning rate	0.5
Max depth	4
Max	2
Min samples	split 2
Random state	0
N estimators	19
Min samples leaf	1

3.3.4. Random Forest

The RF model is comprised of decision trees and can be used for classification or regression. In the classification case, prediction is based on a majority vote of prediction using decision trees, but in the case of regression, the result is the averaging of the tree's output [56]. During the training phase, a training set T_i is created for each tree, based on the samples in the original training set, T, and to build each tree split, m features are randomly selected and, then, analyzed by a measure to determine which one should cause the separation. Due to this randomization, multiple trees are produced, which usually result in better prediction performance, if combined. RF models has several advantages over generally used machine learning methods, including lowest model training time, intensity to handle inconsistent datasets, classification mechanism for embedded features, and inner metrics for determining the impact of features. RF is trained for DDoS attack detection, by using different feature sets. Table 2 shows parameters used for RF.

Table 2. Hyper Parameter used for RF.

Hyper Parameter Name	Values
Bootstrap	True
Criterion	Gini
Min samples	split 2
N estimators	30
Random state	0
Max features	Auto
Min samples leaf	1

3.3.5. Weighted Voting Ensemble Classifier

The first two processes in constructing a classifier ensemble are, usually, selection and combination. Despite the fact that some approaches combine predictions from individual classifiers, the selection of component classifiers is important for the ensemble's performance [57]. The key problem is the variety and precision of the classifiers.

WVE is a representative approach, for combining predictions in paired classification, in which classifiers are not considered equal. On an evaluation set D, each classifier is assigned a weight coefficient, which is typically equal to its classification accuracy. In the proposed work, KNN, RF, and CART decision tree are used as a base learner, predicting the DDoS attack by combining the results of the base learner with WVE.

3.4. Evaluation Measures

Evaluation metrics are used to evaluate the performance of the prediction model. This study used accuracy, precision, recall, and F score to evaluate the performance of machine learning, for DDoS attack detection.

3.4.1. Accuracy

The basic performance metric is accuracy, which is the proportion of correctly predicted observations to all observations. Accuracy is a useful evaluation measure, only when the datasets are uniform, and the false positive and false negative values are almost comparable. Accuracy tells how correctly the classifier is predicting the data points, as shown in Equation (2).

$$Accuracy = \frac{TP}{TP + TN + FP + FN} \tag{2}$$

3.4.2. Precision

Precision is defined as the proportion of accurately predicted positive observations to all expected positive observations. High precision is associated with a low false-positive rate. Precision gives a probability of how correctly the classifier is predicting the positive class. Precision is calculated with Equation (3).

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

3.4.3. Recall

Recall is defined as the ratio of accurately predicted positive observations to all observations in the actual class. Precision gives a probability of how correctly the classifier is predicting the actual positive class, as shown in Equation (4).

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

3.4.4. F1 Score

F1 Score is a normalized average of precision and recall. As a result, this score includes both false positives and false negatives. Although F1 score is simpler than accuracy, it is more useful, especially if class distribution is irregular. F1 score is a harmonic mean of precision and recall, as shown in Equation (5).

$$FMeasure = 2 \times \frac{PR}{(P + R)} \tag{5}$$

4. Results and Discussion

DDoS attack detection and prevention are important problems in a cloud environment. DDoS attack detection is a binary class problem, with benign and DDoS attack class labels. Benign is a normal class. We consider the existence of an attack as a positive class because the interest is in the detection of an attack, and benign is considered as a negative class. MI and RFFI feature selection methods are used. We select 16 features, 19 features, and 23 features, by using the MI and RFFI methods. LR, KNN, GB, RF, and WVE machine learning methods are applied, to selected features. The details of the experimental setup are presented in Table 3. Figures 2–5 show the results of these methods on 16 features, 19 features, 23 features, and all features, respectively, on the CICIDS 2017 dataset. The experimental results show that the overall performance of RF is better, compared to other methods in DDoS attacks detection, with 16 features, 19 features, and 23 features. RF, with these features, has a low miss classification rate, compared to other existing methods.

Table 3. Details of the experimental setup used.

S. No	Components	Description
1	Hardware	Intel Core i3 2nd Gen PC
2	Operating System	Window 10
3	Memory	4 GB RAM
4	Libraries	Pandas, Sk Learn, Matplotlib
5	Storage	MS Excel
6	Core programming language	Python
7	IDE	(Anaconda Jupyter Notebook)

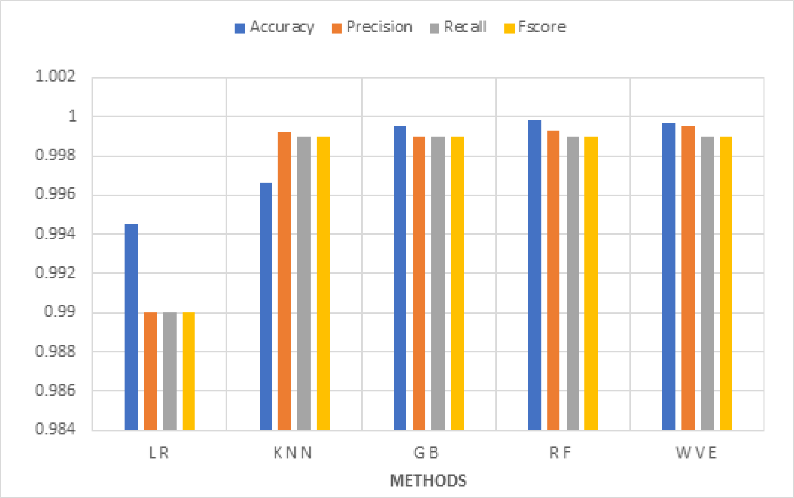

Figure 2. Comparison of different machine learning methods on 16 features.

Figure 2 shows the results of various methods, in DDoS attack detection with 16 features. Sixteen features were selected from the in-hand dataset, using MI and applied machine learning methods, on selected features. RF and WVE methods have the highest prediction accuracy, compared to other methods. All these methods have 99% accuracy and other matrix values. In the case of large datasets, only the measurement of accuracy is not sufficient to measure the performance of the model, since the miss classification of some data points does not affect the accuracy.

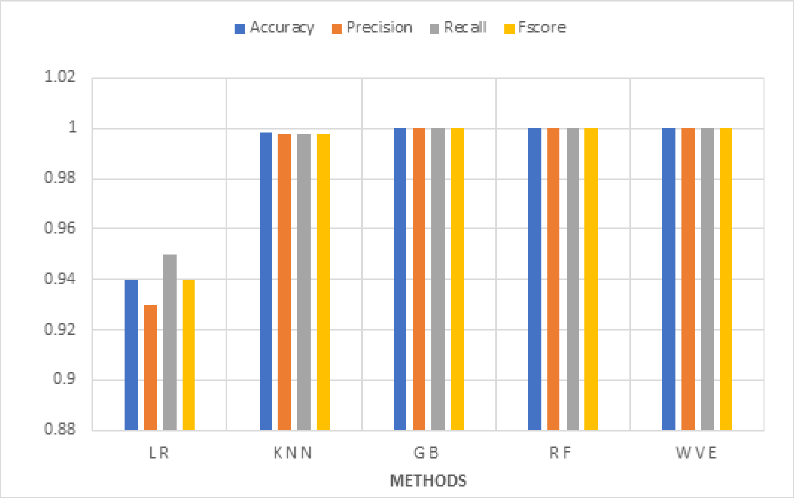

Figure 3. Comparison of different machine learning methods, on 19 features.

Figure 3 shows the results of various machine learning methods in DDoS attack detection, using 19 features that are selected with the RFFI method. The RFFI method is used to select the most relevant feature for DDoS attack detection. The prediction accuracy of the WVE and RF is better, compared to other methods.

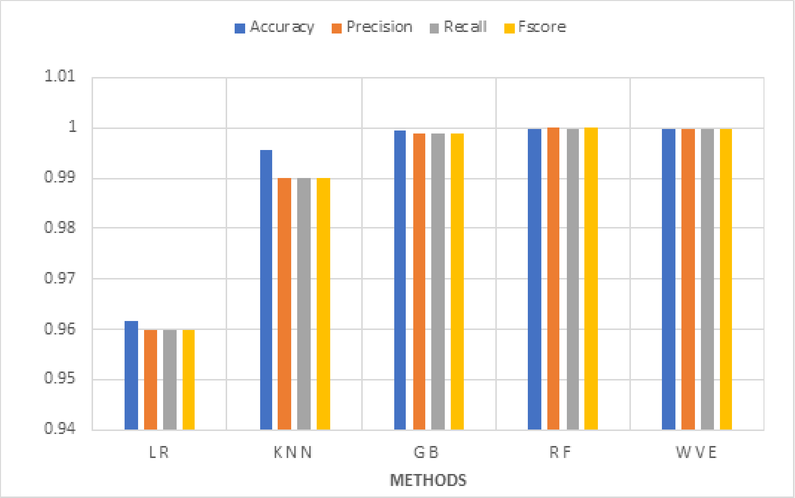

Figure 4. Comparison of different machine learning methods, on 23 features.

Figure 4 shows the results of 23 features, obtained using the MI and applied machine learning methods, on selected features. RF has the highest accuracy, compared to other methods in DDoS attack detection. Twenty-three features were selected from the in-hand dataset, using MI and applied machine learning methods, on selected features. RF has the highest accuracy, compared to other methods in DDoS attack detection.

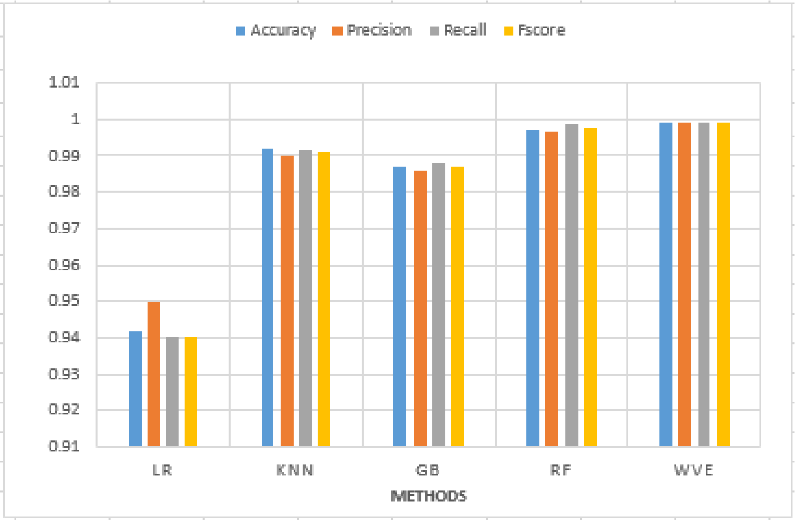

Figure 5. Comparison of different machine learning methods, on complete features.

Figure 5 shows the results of various machine learning methods in DDoS attack detection, using all features. The prediction accuracy of the WVE and RF is highest, compared to other methods.

Tables 4 and 5 show the confusion matrix of different machine learning models, for the CICIDS 2017 and CICDDoS 2019 datasets, respectively. The results show true negative rate, false positive rate, false negative rate, and true positive rate, with different feature sets. The purpose of showing the results of the confusion matrix is to show the miss classification rate of all methods. Low miss classification reflects better accuracy. All these methods have 99% accuracy and other metrics values. For large datasets, measuring accuracy only is not a valid measurement, since miss classification of some data points does not affect the accuracy. This study performed the experiment on the CICIDS 2017 and CIC-DDoS2019 datasets, with a different features set. WVE and RFFI have a low miss classification rate, compared to the other method, with 16 features that are selected by MI. LR has highest miss classification rate, compared to other classifiers, with all feature sets. Tables 4 and 5 demonstrate that using 19 features, which are selected by the RFFI method, has a low miss classification rate. It is concluded that by using all feature sets, we have a high miss classification error, compared to other feature sets.

Table 4. Confusion matrix with 16 features, 19 features, 23 features, and all features, for different models on the CICIDS 2017 dataset.

Method	True Negative	False Positive	False Negative	True Positive
		16 features		
GB	19,535	3	18	25,587
KNN	19,535	3	12	25,593
LR	19,306	232	13	25,592
RF	19,535	3	7	25,598
WVE	19,535	3	6	25,599
		19 features		
GB	19,536	2	2	25,603
KNN	19,509	29	37	25,568
LR	17,237	2301	320	25,285
RF	19,538	0	1	25,604
WVE	19,538	0	1	25,605
		23 features		
GB	19,534	4	13	25,592
KNN	19,461	77	113	25,492
LR	18,675	863	866	24,739
RF	19,538	0	3	25,602
WVE	19,536	2	2	25,603
		All Features		
GB	18,973	565	8	25,597
KNN	19,484	54	21	25,584
LR	17,226	2312	299	25,306
RF	19,531	7	26	25,579
WVE	19,537	1	10	25,595

LR has a high miss classification rate, and WVE has a low miss classification rate, compared to the other methods applied in the detection of a DDoS attack, using 16 features. LR with 19 features, 23 features, and all features has a high miss classification error, compared to GB, KNN, RF, and WVE, for DDoS attack classification. The results show that LR is not performing well, for DDoS attack classification. On the other hand, the RF and WVE models are performing better and have a low miss classification error, using 19 features, 23 features, and all features. The results indicate that these methods are more suitable for detection of DDoS attack classification.

Table 6 shows the comparative results of the proposed method, with the existing methods. The proposed method is superior, compared to the existing methods, in terms of high accuracy and a low miss classification rate. The existing methods have accuracy near to 99%, with more miss classification errors. The proposed method improves the miss classification rate, with only one such attack. The proposed method achieved less miss classification error and high accuracy, by experimenting with the machine learning method on different feature sets and by tuning the parameter of the machine learning classifier.

This study uses the machine learning method for the classification of DDoS attacks. The tree-based methods need less computational time, compared to the distance-based method. KNN is used, which takes more time, compared to the tree-based methods. LR and GB have a high miss classification error, compared to the other methods. These methods need more parameter tuning, to produce fewer miss classification errors. MI features selection takes more time, with an increase in dimensions of data.

Table 5. Confusion matrix with 16 features, 19 features, 23 features, and all features, for different models, on the CICDDoS 2019 dataset.

Method	True Negative	False Positive	False Negative	True Positive
		16 features		
GB	2852	5	2	239,134
KNN	2827	4	27	239,135
LR	2645	726	209	238,413
RF	2853	4	1	239,135
WVE	2854	2	0	239,137
		19 features		
GB	2853	1	1	239,138
KNN	2830	2	24	239,137
LR	2380	425	474	238,714
RF	2849	11	5	239,128
WVE	2848	14	6	239,125
		23 features		
GB	2850	9	4	239,130
KNN	2829	8	25	239,131
LR	2252	439	602	238,700
RF	2850	6	4	239,133
WVE	2853	3	1	239,136
		All Features		
GB	1480	762	1374	238,377
KNN	2830	2	24	239,137
LR	2309	362	545	238,777
RF	2838	15	16	239,124
WVE	2841	10	13	239,129

Table 6. Comparison of the proposed method, with the existing methods, in terms of accuracy.

Method Name	Accuracy
Self-adaptive EEL [9]	99
Deep Neural Network [10]	97.59
Deep Belief Network [11]	96.67
RF [19]	96
Proposed	99.997

5. Conclusions

DDoS attack detection is a common problem in a distributed environment. This type of attack causes the unavailability of cloud service, which makes it essential to detect this attack. A machine learning model can be used to identify this type of attack. The research objective of this work is to detect a DDoS attack, with improved performance. This experiment was performed on the CICIDS 2017 and CICDDoS 2019 datasets. Different files related to DDoS attack were included in experiments, from both datasets. We select the most relevant features, by applying the MI and the RFFI methods. The selected features are fed to machine learning algorithms (RF, GB, WVE, KNN, LR). The overall prediction accuracy of RF with 16 features, is 0.99993, and with 19 features, is 0.999977, which is better, compared to other methods. It is concluded that RF, GB, WVE, KNN, and LR are achieving good results, by using MI and RFFI as feature selection techniques. In the future, we may use wrapper feature selection methods, such as sequential feature selection, with neural networks, for DDoS and other attack detection.

Author Contributions: Conceptualization, Q.W.K. and F.M.; methodology, Q.W.K. and F.M.; software, Q.W.K.; validation, M.A. (Mona Alduailej), M.S., M.A. (Mai Alduailij) and M.T.; formal analysis, M.S., M.T., M.A. (Mai Alduailij), and M.A. (Mona Alduailej); investigation, Q.W.K.; resources, M.S. and M.T.; data curation, F.M., M.T. and M.S.; writing–original draft preparation, Q.W.K. and F.M.; writing–review and editing, M.S., M.T, M.A. (Mona Alduailej) and M.A. (Mai Alduailij).; visualization, M.T. and M.A. (Mona Alduailej) ; supervision, M.S.; project administration, M.S. and M.A. (Mai Alduailij). All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets used are publicly available.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Malik, N.; Sardaraz, M.; Tahir, M.; Shah, B.; Ali, G.; Moreira, F. Energy-efficient load balancing algorithm for workflow scheduling in cloud data centers using queuing and thresholds. *Appl. Sci.* **2021**, *11*, 5849. [CrossRef]
2. Yan, Q.; Yu, F.R. Distributed denial of service attacks in software-defined networking with cloud computing. *IEEE Commun. Mag.* **2015**, *53*, 52–59. [CrossRef]
3. Lau, F.; Rubin, S.H.; Smith, M.H.; Trajkovic, L. Distributed denial of service attacks. In Proceedings of the SMC 2000 Conference Proceedings. 2000 IEEE International Conference on Systems, Man and Cybernetics.'Cybernetics Evolving to Systems, Humans, Organizations, and Their Complex Interactions'(Cat. No. 0), Nashville, TN, USA, 8–11 October 2000; IEEE: Piscataway, NJ, USA, Volume 3, pp. 2275–2280.
4. Sambangi, S.; Gondi, L. A Machine Learning Approach for DDoS (Distributed Denial of Service) Attack Detection Using Multiple Linear Regression. *Proceedings* **2020**, *63*, 51.
5. Erickson, B.J.; Korfiatis, P.; Akkus, Z.; Kline, T.L. Machine learning for medical imaging. *Radiographics* **2017**, *37*, 505–515. [CrossRef]
6. Hasan, A.; Moin, S.; Karim, A.; Shamshirband, S. Machine learning-based sentiment analysis for twitter accounts. *Math. Comput. Appl.* **2018**, *23*, 11. [CrossRef]
7. Malik, S.; Tahir, M.; Sardaraz, M.; Alourani, A. A Resource Utilization Prediction Model for Cloud Data Centers Using Evolutionary Algorithms and Machine Learning Techniques. *Appl. Sci.* **2022**, *12*, 2160. [CrossRef]
8. Aljamal, I.; Tekeoğlu, A.; Bekiroglu, K.; Sengupta, S. Hybrid intrusion detection system using machine learning techniques in cloud computing environments. In Proceedings of the 2019 IEEE 17th International Conference on Software Engineering Research, Management and Applications (SERA), Honolulu, HI, USA, 29–31 May 2019; IEEE: Piscataway, NJ, USA, 2019, pp. 84–89.
9. Kushwah, G.S.; Ranga, V. Optimized extreme learning machine for detecting DDoS attacks in cloud computing. *Comput. Secur.* **2021**, *105*, 102260. [CrossRef]
10. Makuvaza, A.; Jat, D.S.; Gamundani, A.M. Deep Neural Network (DNN) Solution for Real-time Detection of Distributed Denial of Service (DDoS) Attacks in Software Defined Networks (SDNs). *SN Comput. Sci.* **2021**, *2*, 1–10. [CrossRef]
11. Manimurugan, S.; Al-Mutairi, S.; Aborokbah, M.M.; Chilamkurti, N.; Ganesan, S.; Patan, R. Effective attack detection in internet of medical things smart environment using a deep belief neural network. *IEEE Access* **2020**, *8*, 77396–77404. [CrossRef]
12. Intrusion Detection Evaluation Dataset (CIC-IDS2017). Available online: https://www.unb.ca/cic/datasets/ids-2017.html (accessed on 30 September 2021).
13. DDoS Evaluation Dataset (CIC-DDoS2019). Available online: https://www.unb.ca/cic/datasets/ddos-2019.html (accessed on 27 April 2022).
14. Khan, S.; Kifayat, K.; Kashif Bashir, A.; Gurtov, A.; Hassan, M. Intelligent intrusion detection system in smart grid using computational intelligence and machine learning. *Trans. Emerg. Telecommun. Technol.* **2021**, *32*, e4062. [CrossRef]
15. Sandhu, R.S.; Samarati, P. Access control: principle and practice. *IEEE Commun. Mag.* **1994**, *32*, 40–48. [CrossRef]
16. Khan, M.S.; Khan, N.M.; Khan, A.; Aadil, F.; Tahir, M.; Sardaraz, M. A low-complexity, energy-efficient data securing model for wireless sensor network based on linearly complex voice encryption mechanism of GSM technology. *Int. J. Distrib. Sens. Netw.* **2021**, *17*, 15501477211018623. [CrossRef]
17. Sardaraz, M.; Tahir, M. SCA-NGS: Secure compression algorithm for next generation sequencing data using genetic operators and block sorting. *Sci. Prog.* **2021**, *104*, 00368504211023276. [CrossRef]
18. Zhong, Z.; Xu, M.; Rodriguez, M.A.; Xu, C.; Buyya, R. Machine Learning-based Orchestration of Containers: A Taxonomy and Future Directions. *ACM Comput. Surv. (CSUR)* **2021**. [CrossRef]

19. Bindra, N.; Sood, M. Detecting DDoS attacks using machine learning techniques and contemporary intrusion detection dataset. *Autom. Control. Comput. Sci.* **2019**, *53*, 419–428. [CrossRef]
20. Kshirsagar, D.; Kumar, S. An efficient feature reduction method for the detection of DoS attack. *ICT Express* **2021**, *7*, 371–375. [CrossRef]
21. Dehkordi, A.B.; Soltanaghaei, M.; Boroujeni, F.Z. The DDoS attacks detection through machine learning and statistical methods in SDN. *J. Supercomput.* **2021**, *77*, 2383–2415. [CrossRef]
22. The CTU-13 Dataset. A Labeled Dataset with Botnet, Normal and Background traffic. Available online: https://www.stratosphereips.org/datasets-ctu13 (accessed on 27 April 2022).
23. ISOT Research Lab: Botnet and Ransomware Detection Datasets. Available online: https://www.uvic.ca/ecs/ece/isot/datasets/?utm_medium=redirect&utm_source=/engineering/ece/isot/datasets/&utm_campaign=redirect-usage (accessed on 27 April 2022).
24. Canadian Institute for Cybersecurity:UNB-ISCX Datasets. Available online: https://www.unb.ca/cic/datasets/botnet.html (accessed on 27 April 2022).
25. Kshirsagar, D.; Kumar, S. An ensemble feature reduction method for web-attack detection. *J. Discret. Math. Sci. Cryptogr.* **2020**, *23*, 283–291. [CrossRef]
26. Adhao, R.; Pachghare, V. Feature selection using principal component analysis and genetic algorithm. *J. Discret. Math. Sci. Cryptogr.* **2020**, *23*, 595–602. [CrossRef]
27. Binbusayyis, A.; Vaiyapuri, T. Identifying and benchmarking key features for cyber intrusion detection: an ensemble approach. *IEEE Access* **2019**, *7*, 106495–106513. [CrossRef]
28. Stiawan, D.; Idris, M.Y.B.; Bamhdi, A.M.; Budiarto, R. CICIDS-2017 dataset feature analysis with information gain for anomaly detection. *IEEE Access* **2020**, *8*, 132911–132921.
29. Ferrag, M.A.; Maglaras, L.; Ahmim, A.; Derdour, M.; Janicke, H. Rdtids: Rules and decision tree-based intrusion detection system for internet-of-things networks. *Future Internet* **2020**, *12*, 44. [CrossRef]
30. Gu, J.; Lu, S. An effective intrusion detection approach using SVM with naïve Bayes feature embedding. *Comput. Secur.* **2021**, *103*, 102158. [CrossRef]
31. Ahuja, N.; Singal, G.; Mukhopadhyay, D.; Kumar, N. Automated DDOS attack detection in software defined networking. *J. Netw. Comput. Appl.* **2021**, *187*, 103108. [CrossRef]
32. Tonkal, Z.; Polat, H.; Başaran, E.; Cömert, Z.; Kocaoğlu, R. Machine Learning Approach Equipped with Neighbourhood Component Analysis for DDoS Attack Detection in Software-Defined Networking. *Electronics* **2021**, *10*, 1227. [CrossRef]
33. McCullough, E.; Iqbal, R.; Katangur, A. Analysis of Machine Learning Techniques for Lightweight DDoS Attack Detection on IoT Networks. In *Forthcoming Networks and Sustainability in the IoT Era. FoNeS-IoT 2020. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering*; Ever, E., Al-Turjman, F., Eds.; Springer: Berlin/Heidelberg, Germany, 2021; Volume 353, pp. 96–110.
34. Revathi, M.; Ramalingam, V.; Amutha, B. A Machine Learning Based Detection and Mitigation of the DDOS Attack by Using SDN Controller Framework. *Wirel. Pers. Commun.* **2021**, 1–25. [CrossRef]
35. Chiba, Z.; Abghour, N.; Moussaid, K.; Rida, M. Intelligent approach to build a Deep Neural Network based IDS for cloud environment using combination of machine learning algorithms. *Comput. Secur.* **2019**, *86*, 291–317. [CrossRef]
36. Vinayakumar, R.; Alazab, M.; Soman, K.; Poornachandran, P.; Al-Nemrat, A.; Venkatraman, S. Deep learning approach for intelligent intrusion detection system. *IEEE Access* **2019**, *7*, 41525–41550. [CrossRef]
37. University of California, Department of Information and Computer Science: The UCI KDD Archive. Available online: http://kdd.ics.uci.edu/ (accessed on 27 April 2022).
38. Cui, W.; Lu, Q.; Qureshi, A.M.; Li, W.; Wu, K. An adaptive LeNet-5 model for anomaly detection. *Inf. Secur. J. Glob. Perspect.* **2021**, *30*, 19–29. [CrossRef]
39. Choraś, M.; Pawlicki, M. Intrusion detection approach based on optimised artificial neural network. *Neurocomputing* **2021**, *452*, 705–715. [CrossRef]
40. Tang, T.A.; McLernon, D.; Mhamdi, L.; Zaidi, S.A.R.; Ghogho, M., Intrusion detection in sdn-based networks: Deep recurrent neural network approach. In *Deep Learning Applications for Cyber Security*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 175–195.
41. Canadian Institute for Cybersecurity: ISCX NSL-KDD Datasets. Available online: https://www.unb.ca/cic/datasets/nsl.html (accessed on 27 April 2022).
42. Azzaoui, H.; Boukhamla, A.Z.E.; Arroyo, D.; Bensayah, A. Developing new deep-learning model to enhance network intrusion classification. *Evol. Syst.* **2021**, *13*, 1–9. [CrossRef]
43. Wei, Y.; Jang-Jaccard, J.; Sabrina, F.; Singh, A.; Xu, W.; Camtepe, S. Ae-mlp: A hybrid deep learning approach for ddos detection and classification. *IEEE Access* **2021**, *9*, 146810–146821. [CrossRef]
44. Shroff, J.; Walambe, R.; Singh, S.K.; Kotecha, K. Enhanced Security Against Volumetric DDoS Attacks Using Adversarial Machine Learning. *Wirel. Commun. Mob. Comput.* **2022**, *2022*, 5757164. [CrossRef]
45. Mahanta, H.J.; Nath, K.; Roy, A.K.; Kotecha, K.; Varadaranjan, V. Using Genetic Algorithm in Inner Product to Resist Modular Exponentiation from Higher Order DPA Attacks. *IEEE Access* **2021**, *10*, 3238–3251. [CrossRef]

46. Saini, P.S.; Behal, S.; Bhatia, S. Detection of DDoS attacks using machine learning algorithms. In Proceedings of the 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 12–14 March 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 16–21.
47. Injadat, M.; Moubayed, A.; Nassif, A.B.; Shami, A. Multi-stage optimized machine learning framework for network intrusion detection. *IEEE Trans. Netw. Serv. Manag.* **2020**, *18*, 1803–1816.. [CrossRef]
48. Li, J.; Cheng, K.; Wang, S.; Morstatter, F.; Trevino, R.P.; Tang, J.; Liu, H. Feature selection: A data perspective. *ACM Comput. Surv. (CSUR)* **2017**, *50*, 1–45. [CrossRef]
49. Peng, H.; Long, F.; Ding, C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* **2005**, *27*, 1226–1238. [CrossRef]
50. Saeys, Y.; Abeel, T.; Van de Peer, Y. Robust feature selection using ensemble feature selection techniques. In Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2008. Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2008, Vol. 5212, pp. 313–325.
51. Larasati, A.; DeYong, C.; Slevitch, L. The application of neural network and logistics regression models on predicting customer satisfaction in a student-operated restaurant. *Procedia-Soc. Behav. Sci.* **2012**, *65*, 94–99. [CrossRef]
52. Peterson, L.E. K-nearest neighbor. *Scholarpedia* **2009**, *4*, 1883. [CrossRef]
53. Batista, G.; Silva, D.F. How k-nearest neighbor parameters affect its performance. In Proceedings of the Argentine Symposium on Artificial Intelligence ((ASAI), Mar del Plata, Argentina, 24–28 August 2009; Citeseer: Princeton, NJ, USA, 2009; pp. 95–106.
54. Biau, G.; Cadre, B.; Rouvière, L. Accelerated gradient boosting. *Mach. Learn.* **2019**, *108*, 971–992. [CrossRef]
55. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H. Xgboost: Extreme gradient boosting. *R Package Version 0.4-2* **2015**, *1*, 1–4.
56. Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J.C.; Sheridan, R.P.; Feuston, B.P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947–1958. [CrossRef] [PubMed]
57. Kuncheva, L.I.; Rodríguez, J.J. A weighted voting framework for classifiers ensembles. *Knowl. Inf. Syst.* **2014**, *38*, 259–275. [CrossRef]