Iterative Feature Selection-Based DDoS attack Prevention Approach in Cloud

Original Scientific Paper

Sarah Naiem
Helwan University,
Faculty of Computers and Artificial Intelligence
Cairo, Egypt
SarahNaiem@fci.helwan.edu.eg

Amira M. Idrees
Fayoum University
Faculty of Computers and Artificial Intelligence
Cairo, Egypt
ami04@fayoum.edu.eg

Ayman E. Khedr
Future University in Egypt
Faculty of Computers and Information Technology
Cairo, Egypt
ayman.khedr@fue.edu.eg

Mohamed Marie
Helwan University,
Faculty of Computers and Artificial Intelligence
Cairo, Egypt
Dr.mmariem@fci.helwan.edu.eg

Abstract *– Distributed Denial of Service (DDOS) attacks aim to exploit the capacity and performance of a network's infrastructure, making the cloud environment one of the biggest targets for attackers. Many efforts are being made in the field of technology to prevent them from disrupting the services provided. Machine Learning techniques are a means to protect against DDOS attacks. Data preprocessing, feature selection, and classifiers are the main components of any prevention framework. The focus of this study is to find and enhance the feature selection approach for increasing the accuracy of the classifiers in detecting DDOS attacks from regular traffic. We used four different techniques, including Pearson Correlation Coefficient (PCC), Random Forest Feature Importance (RFFI), Mutual information (MI), and Chi-squared(X2) measure which we tested on different classifiers. The first selection approach was based on the feature's independency level then the second iteration was based on the feature's importance. We also examined the claim of dropping attacks from the dataset for better accuracy. The best performing set of features was from using PCC and RFFI together for feature selection with average accuracy and precision of 99.27% and 97.60%, which is higher than the use of PCC for both measures by almost 2%. The accuracy is also higher by nearly 12% from the same approach dropping 50% of the attacks.*

Keywords*: DDOS attacks; cloud environment; machine learning; feature selection; random forest; Pearson correlation coefficient; mutual information; chi-square.*

1. INTRODUCTION

Distributed Denials of services (DDOS) is on the list of top attacks jeopardizing the cloud environment, messing with the cloud traffic, and denying benefits to a legitimate user [1]. Recently the cloud computing environment gained massive popularity due to the variety of services it provides, including education, networking, storage, security, elasticity, and migration flexibility, making it a target for cybercrime. [2] [3] . DDoS aims to disrupt a specific server, service, or network's usual traffic by saturating the target or its surrounding infrastructure with non-legitimate traffic. The attacker's DDoS attempts are usually successful because they use several compromised computers as attack traffic sources. Where DDoS attacks operate under the theory of using numerous machines to produce high-intensity-based attack traffic to compromise the integrity of the network [4]. These machines are unaware that the attacker is employing them to harm and are usually referred to as "zombies" who are challenging to detect. [5]. Many efforts have been made to protect the cloud and internet from these attacks with the help of Machine Learning techniques, deep learning, count-base filtering, resource usage, data mining, and other methods. [6]. On the other hand, feature selection techniques are proposed in different research which tackles various issues such as in the education field [7], [8], and construction field [9]. Moreover, features selection techniques also proved their effectiveness in other research directions such as in Recommendation systems sentimental analysis [10], classification [11] [12] [13], query answering [14] [8], decision support systems [15] [16] , and Internet of Things (IoT) [17].

This research focuses on preventing and detecting using machine learning (ML) techniques, including feature engineering and selection, data normalization, and ML classification algorithms. In addition to that, it has been claimed by Tan et al. in [18] that dropping part of the attacks from the dataset would improve the detection accuracy, and we are testing this claim throughout our work. The rest of this paper will include a literature review of the previous related work, a description of the framework applied, including the dataset, the feature engineering approaches followed, including Pearson Correlation, Information Gain, Chi2, Random Forest Feature Importance RFFI, and a comparison between different machine learning classifiers including Random Forest, Decision Tree, and Gaussian Naïve Bayes.

2. LITERATURE REVIEW

In [19] the authors focused on applying the K-fold cross-validation on the CIC-IDS2017 dataset. They tested the model on Random Forest (RF), K- nearest neighbor (KNN), Decision Tree (DT), Gaussian Naïve Bayes (GNB), Support Vector Machine (SVM) on all three kernels, including sigmoid, polynomial, and RBF), and logistic regression. They set the K value equal to 5, and each k was split into 5 sets, one for testing and 4 for training. The results of the classifier's accuracy, precision, and recall were compared, and they concluded that the one with the best results was the DT. It is clear that even though the DT is the best classifier with an accuracy of 99.9,4, the other 6 classifiers, except for the DT, showed an accuracy between 99.78 and 99.88, while the GNB had an accuracy of 61.22.

The authors in [20], used 3 combined datasets, including CSE-CIC-IDS2018-AWS, CICID2017, and CIC DOS 2017 datasets, and selected 7 features. The resulting dataset was tested using RF, DT, GNB, and multilayer perception neural network approaches using different training and testing split sets from 90/10 to 50/50. The results showed that the RF had the highest accuracy from the average of the 5 train test splits at 99.969%, followed by DT at 99.951%, MLP at 98.87,6%, and the least accurate was the GNB at 78.45%. It was also concluded that the different Train-test splits don't affect the model's accuracy feature selection approach discussed was neither mentioned nor the normalization technique followed. Also, the author was not sure about the accuracy of the model due to the selected features or the combination of different datasets. While in [1], in the dataset CICDDOS2019, the authors applied the chi-squared (x2) feature selection approach to select the top 10 features after cleaning the data. The preprocessed dataset was used for the detection using the deep learning technique mixing BILSTM and CNN using different filter size, filter count, and unit size. After using 10 models with varying filter size and count and unit size, it was clear that reducing the unit size increases the accurate model's accuracy while lowering the filter size decreases training time. But even though the hybrid model was compared to other machine learning models, including RF, SVM, KNN, LR, and XBG, proving that it has higher accuracy, the accuracy rate of these models was significantly lower than the average, where it ranged from 64 to 78%. This indicated that the set of selected features using the X2 doesn't represent this dataset's best features.

To improve the feature selection of DDOS in the cloud, the authors of [21] used the CICDDOS2019 dataset and applied the chi-squared (x2) feature selection approach to select the top 10 features after cleaning the data. The preprocessed dataset was used to detect DDOS through a hybrid deep learning technique mixing BILSTM and CNN using different filter size, filter count, and unit size. After using 10 models with varying filter size and count and unit size, it was clear that reducing the unit size increases the accuracy of models and the filter size decreased the training time. Even though the authors compared their hybrid model to other machine learning models, including RF, SVM, KNN, LR, and XBG, proving that it has higher accuracy, the accuracy rate of these models was significantly lower than the average,e where it ranged from 64 to 78%.

The authors in [5] focused on explaining and providing a detailed background for DDOS detection, including the different conceptual ML techniques and types of DDOS attacks. The random forest feature importance (RFFI) technique was used for selecting the critical features resulting in 13 features according to their score. In [21], the authors used 2 different feature selection approaches, including Mutual Information (MI) and RF approach,ches and compared the accuracy by training the dataset using logistic regression (LR), KNN, Gradient Boosting (GB), and RF, and weighted voted ensemble (WVE). They used 3 sets and tested them with the 5 classifiers—16 features with MI, 19 with RFFI, and 23 with MI. The accuracy of the 16 selected features was 9,9.993%, and the 19 features was 99.997%. Even though the set of features chosen had very high accuracy, the authors only stated that their paper proved that MI and RFFI work well with the selected ML classifiers.

The authors in [22], focused on enhancing the accuracy of GNB. The dataset selected for the research was KDD 99, which was cleaned before deciding the essential features using correlation-based feature selection (CSF). The GNB classifier was enhanced using 2 approaches. The first was the elimination of the zeroes from the dataset, and the second was changing the GNB statistical equation from its multiplication form to its addition form, increasing the accuracy by about 4 %. In [23], the author used the CICIDOS-2019 data, where she selected the top 20 features using the Extra Tree Classifier approach. After that, she used Rf, DT, SVM, and NB classifiers for DDOS detection focusing on LDAP and MSSQL DDOS attacks. The accuracy of RF and SVM classifiers was 99.9,9%, while DT was 99.89%, and NB was 99.98%.

The authors in [24] focused on the slow rate DDOS attacks where they integrated CICID2017 and CSE-CIC-IDS2018, extracted the top 30 features using Information gain, and then selected the top 10 using the Chi-square approach. The model was trained on J48, bagging technique, MLP, and KNN classifiers. The attacks under testing included DOS GoldenEye, Slowloris, Slowwhttptest, Hulk, DDoS HOIC, and DDOS LOIC-HTTP. The F1 score, True positive, and true negative were calculated, showing that the feature selection approaches high results for all classifiers. The authors stated that the model's accuracy was almost 99%, with a meager false negative rate.

In [25], the DT model for feature ranking was applied on CICDDOS2019 resulting in a list of the top 30 features, in addition to the use of the person correlation coefficient (PCC) approach resulting in a list of 20 features. They tested the selected features on different ML models, including RF, Light Gradient Boosting, Cat Boost, and CNN. The results of the 20 selected features with the RF and GB and the 30 features with the CatBoost and the CNN were the highest-performing classifiers. Table 1 summarizes the previously mentioned studies in the literature review and their limitations when it comes to the techniques and approaches followed in feature selection which we are trying to overcome in our research.

Table 1. Literature review summary

Reference	Dataset	ML techniques applied	limitation
(Nalayinil and Katiravan 2022) [19]	CIC-IDS2017	K-fold cross-validation on Rf, Knn, GNB, SVM, and LR	The researchers did not focus on the feature selection phase, which would have an impact on the results if considered
(Coelho 2022) [20]	3 combined datasets CSE-CIC-IDS2018-AWS, CICID2017, and CIC DOS 2017 datasets	RF, DT, GNB, and multilayer perception neural network	The feature selection approach discussed was not mentioned, nor was the normalization technique followed. The author needed clarification about the accuracy of the models, whether it was due to the selected features or the combination of different datasets.
(Praveen and Rimal 2020) [1]	CCIDS2017	SVM and NB	The authors only mentioned using 20 features without stating the feature selection criteria.
(Alghazzawi , et al. 2021) [26]	CICDDOS2019	hybrid deep learning technique mixing BILSTM and CNN	The set of selected features using the X2 does not represent the best set of features for this Dataset
(Narote, Zutshi and Potdar 2022) [5]	CCIDS2017	Used RFFI for feature selection	The selected set of features was not tested on any ML techniques, and no accuracy or results were provided to show the success of the feature selection approach chosen.
(Alduailij , et al. 2022) [21]	Not specified	MI and RF for feature selection and compared results using LR, KNN, GB, RF, and WVE)	It should have stated which approach is better and how to choose the appropriate one.
(Kurniawan, et al. 2021) [22]	KDD 99	GNB	The improvement in the GNB classifier was not tested on an up-to-date dataset.
(Mishra 2022) [23]	CICIDOS-2019	Extra Tree Classifier for feature selection and RF, DT, SVM, and NB	Only 2 types of DDOS attacks (LDAP-DDOS MSSQL) were taken into consideration
(Swe, Aung and Hlaing March 7-11, 2021) [24]	CICID2017 and CSE-CIC-IDS2018	RF, DT, SVM	The feature selection approach chosen was only tested for the detection of slow-rate DDOS, and it does not show if it would work for other DDOS attack types
(Alghoson and Abbass 2021) [25]	CICDDOS2019	DT model for feature selection Rf, Light Gradient Boosting, Cat Boost, and CNN	Even though the authors tested the different 2 sets of features on 4 classifiers, they only displayed the results of 4 ML algorithms, and the accuracy of the 8 sets of classifiers should have been provided.

3. PROPOSED FRAMEWORK

This section will describe the components of our framework and methodology. Our Main purpose is to use this dataset most efficiently to be able to detect anomalies in the traffic. An overview of the proposed framework is displayed in Fig 1.

3.1. DATASET

Throughout our research, we targeted some of the available datasets regarding IDs and DDOS attacks. Our focus is on CSE-CIC-ID2018, an open-source dataset made available by the University of New Brunswick UNB. The dataset has 80 features presenting seven attacks, including DDOS, DOS, Web-attacks, infiltration, Brute force, and Botnet attacks, and benign traffic generated through the CICFlowmeter-V3, which we presented in table 2. The data distribution shows that the total number of attacks given in the dataset is less than 20% of the traffic flow. The dataset's traffic is captured and delivered in 7 CSV formatted files classified according to the dates of their occurrences, including Wednesday 14/2/2018, Thursday 15/2/2018, Friday 16/2/2018, Tuesday 20/2, Wednesday 21/2/2018, Thursday 22/2/2018, Friday 23/2/2018 [27] [28] [29]

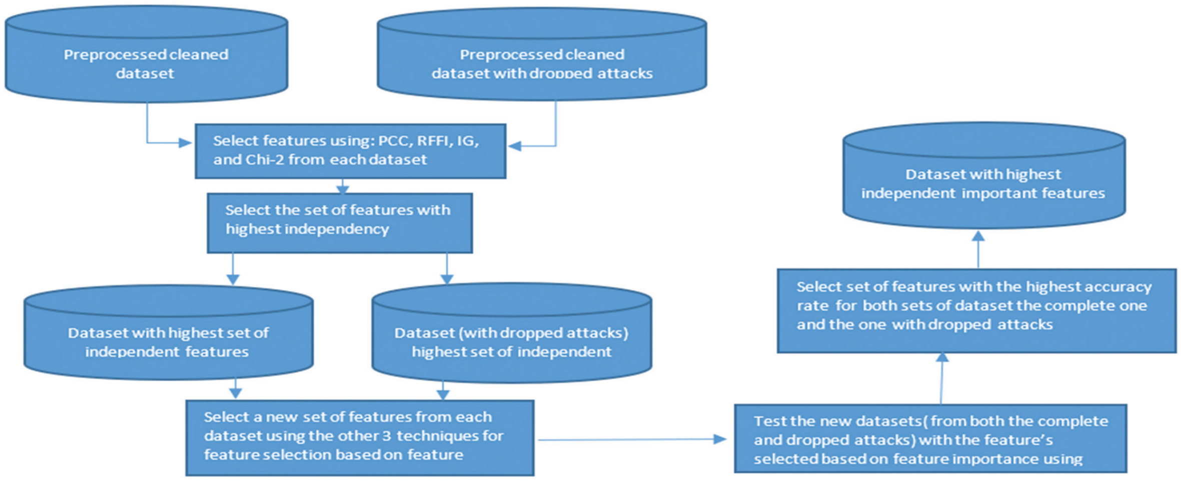
Preprocessed cleaned dataset
Preprocessed cleaned dataset with dropped attacks
Select features using: PCC, RFFI, IG, and Chi-2 from each dataset
Select the set of features with highest independency
Dataset with highest set of independent features
Dataset (with dropped attacks) highest set of independent
Select a new set of features from each dataset using the other 3 techniques for feature selection based on feature
Test the new datasets(from both the complete and dropped attacks) with the feature's selected based on feature importance using
Select set of features with the highest accuracy rate for both sets of dataset the complete one and the one with dropped attacks
Dataset with highest independent important features

Fig 1. framework overview

Table 2. dataset traffic distribution

Type of traffic	Distribution of traffic within the dataset
Benign	83.07%
DDoS	7.79%
DoS	4.03%
Web attacks	0.006 %
Infiltration	0.997%
Brute-force	2.35%
Botnet	1.76 %

3.2. DATA PREPROCESSING

The first step we took was data preprocessing, which is cleaning the data. After that,t the feature selection and normalization of the data is conducted using Machine learning techniques to predict and prevent attacks. Without this phase, any model won't perform as intended, as no machine learning algorithm can handle it to produce predictions and insights.

The most vital part of the data preprocessing is the data cleaning performance, which removes any missing and incomplete data that would result in inaccurate results if considered. The dataset used included some columns with a handful of Zero values, infinite and null values which would have highly impacted the framework's accuracy, so we replaced all the null and infinite values with zeros, and the rows with more than 40% missing values were dropped from the data.

Another thing applied in this phase was dropping 50 percent of the attack traffic presented to represent a more realistic attack. According to their study, this approach was conducted by M.Tan 2019 in their attempts to create a deep learning approach for real-time network intrusion detection and according to their research. Even though they dropped 90% of the attacks, not 50%, to represent a more realistic traffic environment, it still resulted in more accurate results [18]. Through this study, we will present how applying this affects the accuracy of the feature extraction and different supervised and unsupervised ML models to not drop any attack from the dataset.

3.3. FEATURE EXTRACTION

Several different feature selection approaches were conducted to reach the optimum method of generating the best-fitted list of essential features from the 80 features presented through the dataset, including PCC, MI, RFFI, and chi2.

- **Pearson Correlation Coefficient**

The PCC score algorithm calculates each feature's grades regarding the label feature. The features with a threshold higher than or equivalent to 0.08 are selected, and the rest of the 80 features are dropped before the data is used in further steps. PCC score calculates the forte of the linear relationship between variables in a correlative matter [30].

- **Mutual Information**

MI feature selection is based on MI obtained from getting the information gain (IG) and entropy to get the top features with the most IG. The mutual information for variables X and Y and the entropy are represented as follows [31]:

> "I(X; Y) = H(X) – H(X | Y) Where I(X; Y) is the mutual information for X and Y, H(X) is the entropy for X and H(X | Y) is the conditional entropy for X given Y."

It is conducted in python with the help of some sKlearn libraries, including the "mutual_info_classif" and "SelectKBest" along with the "train_test_split." After the data is cleaned and split into train and test, it is passed to the mutual information class, and the features are sorted from the ones with the highest information gain to the least. The top 20 features of the dataset with dropped and without attacks turned out to be the same.

- **Random Forest Feature Importance**

RFFI is one of the most vital feature selections in data science regarding selecting the most significant features. Its approach is based on random forest trees to reduce the Gini impurity. It uses the data after it is cleaned to train using a random forest classifier to select the most vital features. As a result, it creates a subset dataset that is trained again using the RF classifier to compare the accuracy of both datasets. The result of this approach is a list of the essential features.

- **Chi-Squared**

Chi2 followed for feature selection was the Chi2, a statistical approach used to evaluate the correlation between independent categorical variables of a dataset by calculating the p-value and selecting the ones with high correlation reflected through the best chi-square score. It is calculated by subtracting the expected frequency from the observed frequency and divided by the predicted frequency [32].

In our efforts to find the best set of features for the CICID2018 dataset, the RFFI was conducted on four data sets. The first two sets were based on the original dataset with 80 features, once without dropping the attacks and once with dropping 50 % of the attacks. The second two sets were based on the features selected from the first phase result based on the feature dependency. The set of selected features with the highest level of independence was the PCC, which resulted in 24 features dropping 50% of the attacks and 29 without that. We chose a subset of features from the 24 and 29 features based on feature importance achieved from applying the RFFI approach.

Out of all the feature selection methods and approaches, we had 10 sets of derived features, and one of the two sets of selected features using IG was eliminated since they gave the same set of features.

The resulting 9 sets of features are displayed in table 3 and table 4.

Table 3. Set of selected features without dropping attacks

FEATURE SELECTION MODEL	FEATURES
PCC	Dst Port, Protocol, Fw d Pkt Len Max, Fwd Pkt Len Min, Fwd Pkt Len Mean, Fwd Pkt Len Std, Bwd Pkt Len Min, Bwd Pkt Len Mean, Flow Pkts/s, Bwd IAT Tot, Fwd Pkts/s, Bwd Pkts/s, Idle Std, Pkt Len Max, Pkt Len Mean, Pkt Len Std, ACK Flag Cnt, Fwd Seg Size Min, Pkt Size Avg, Fwd Seg Size Avg, Bwd Seg Size Avg, Init Fwd Win Byts, Active Max, TotLen Fwd Pkts, Flow Duration
RFFI	Subflow Fwd Byts, Flow Pkts/s, Init Fwd Win Byts, Flow IAT Std, Active Mean, Fwd IAT Max, Active Max, Bwd Blk Rate Avg, SYN Flag Cnt, Fwd Pkt Len Max, Fwd Blk Rate Avg, Fwd IAT Std, Active Std, Fwd Pkt Len Min, Idle Max
CHI2	Flow Byts/s, Flow IAT Std, Bwd Pkts/s, Fwd IAT Std, Flow Duration, Fwd IAT Tot, Flow IAT Max, Fwd IAT Max, Bwd IAT Mean, Bwd IAT Tot, Flow IAT Mean, Fwd IAT Mean, Bwd IAT Max, Fwd Pkts/s, Bwd IAT Min, Bwd IAT Std, Dst Port, Flow Pkts/s, Fwd Pkt Len Std, Active Mean, Idle Mean, Active Max, Idle Min, Fwd Seg Size Avg, Fwd Pkt Len Mean, Idle Max, Idle Std, Pkt Size Avg, Pkt Len Var, Pkt Len Std
IG	Fwd Seg Size Min, Init Fwd Win Byts, Dst Port, Bwd Pkts/s, Fwd Pkts/s, Flow Pkts/s, Flow IAT Mean, Flow Duration, Init Bwd Win Byts, Flow IAT Max, Fwd Pkt Len Max, Pkt Len Max, Subflow Fwd Byts, TotLen Fwd Pkts, Fwd Seg Size Avg, Fwd Pkt Len Mean, Pkt Len Mean, Pkt Size Avg, Pkt Len Std, Pkt Len Var
RFFI-PCC	Subflow Fwd Byts, Flow Pkts/s, Flow IAT Std, Init Fwd Win Byts, Active Mean, Fwd IAT Max, Active Max, Bwd Blk Rate Avg, SYN Flag Cnt, Fwd Pkt Len Max, Subflow Bwd Byts, Bwd Pkt Len Mean, Fwd URG Flags, Bwd IAT Std, Bwd Pkt Len Std, Fwd Blk Rate Avg

Table 4. Set of selected features with dropping 50% of the attacks

Feature selection Model Name with dropping attacks	Features
PCC-D	Dst Port, Protocol, Fwd Pkt Len Max, Fwd Pkt Len Min, Fwd Pkt Len Mean, Fwd Pkt Len Std, Bwd, Pkt Len Min, Bwd Pkt Len Mean, Flow Pkts/s, Bwd IAT Tot, Fwd Pkts/s, Bwd Pkts/s, Pkt Len Min, Pkt Len Max, Pkt Len Mean, Pkt Len Std, ACK Flag Cnt, URG Flag Cnt, Pkt Size Avg, Fwd Seg Size Avg, Bwd Seg Size Avg, Init Fwd Win Byts, Init Bwd Win Byts, Fwd Seg Size Min
RFFI-D	Subflow Fwd Byts, Flow Pkts/s, Flow IAT Std, Init Fwd Win Byts, Active Mean, Fwd IAT Max, Active Max, Bwd Blk Rate Avg, SYN Flag Cnt, Fwd Pkt Len Max, Subflow Bwd Byts, Bwd Pkt Len Mean, Fwd URG Flags, Bwd IAT Std, Bwd Pkt Len Std, Fwd Blk Rate Avg
Chi2-D	Flow Byts/s, Flow IAT Std, Bwd Pkts/s, IAT Std, Flow Duration, Fwd IAT Tot, Flow IAT Max, Fwd IAT Max, Bwd IAT Mean, Bwd IAT Tot, Flow IAT Mean, Fwd IAT Mean, Bwd IAT Max, Fwd Pkts/s, Bwd IAT Min, Bwd IAT Std, Dst Port, Flow Pkts/s, Fwd Pkt Len Std, Active Mean, Idle, Mean, Active Max, Idle Min, Fwd Seg Size Avg, Fwd Pkt Len Mean, Idle Max, Idle Std, Pkt Size Avg, Pkt Len Var, Pkt Len Std
RFFI-PCC-D	Dst Port, Pkt Len Max, ACK Flag Cnt, Init Fwd Win Byts, Pkt Len Mean, Protocol, Bwd IAT Tot, Pkt Len Std, Pkt Len Min, Bwd Seg Size Avg, Bwd Pkts/s, Bwd Pkt Len Mean, Init Bwd Win Byts, Fwd Seg Size Min, Fwd Pkt Len Std

3.4. NORMALIZATION

This phase is vital to transfer the data into a format that could be used in the training and testing phase without affecting its essence or performance [25]. It is conducted because most of the data represent different types and formats, making it nearly impossible to handle, making it an important step to standardize the data before using it. Several techniques could be used for data normalization, including MinMAxScaler and StandardScaler.

MinMax Scaler is based on representing the maximum value as 1 and the minimum value as 0. Accordingly, it represents all the data between 1 and zero, while the standard scaler scales the data within the maximum and minimum values range. The idea of using the MinMax scaler is based on maintaining the actual distribution and representation of the data [33]

The equation for the MinMax Scaler is

$$X' = \frac{x - min(x)}{(x) - min\,(x)} \tag{1}$$

while the Standard Scaler equation is

$$\text{z=}z = \frac{x-\mu}{\sigma} \tag{2}$$

where x is the score µ the men, and sigma is the standard deviation [33].

3.5. DDOS ATTACKS DETECTION AND DATA CLASSIFICATION

After the data had been preprocessed and normalized, the classification using different machine learning classifiers was applied. The first step in this process is splitting the data into train and testing, done through the train-t testing. For our set of data, we selected three different supervised machine learning algorithms, including the Decision Tree (DT), Random Forest (RF), and Gaussian Naïve Bayes (GNB).

- Decision Tree (DT) is very similar in structure to a flow chart based on a graph with nodes descending from the root or central node. It creates branches in efforts to model the relation between the features of a dataset and its targeted potential output. It is one easy and understandable structure, and normalization of the data is not a necessary step in data preprocessing. DT classifiers use the "Bagging Technique" which trains more DTs in parallel through bootstrapping data samples where the final prediction is based on the results of the trees that are running in parallel [25] [6].
- Random Forest is also a supervised ML classification and regression algorithm based on bagging techniques. It builds several DTs from different illustrations from the datasets and uses their results together. The RF uses the bootstrap data sample and calculates each node's split by subdivision of the features. Using RF ML classifiers results in very accurate results even with imbalanced and missing data. It is also very flexible, has less variance than a single DT, works perfectly with a largamountsnt of data, and resolves the overfitting of data by averaging the results of several DTs. Unfortunately, the main problem with RF is its complexity which results in high computational time and the need for higher computational resources [25] [23].
- Gaussian Naïve Bayes supports continuous data derived from the Gaussian normal distribution, which is also based on Naive Bayes (NB) derived from the Bayes theorem. The NB is based on the hypothesis that features are independent. This classifier is considered one of the simple and easily implementable techniques for supervised machine learning classification [22] [20].

4. PERFORMANCE EVALUATION AND RESULTS DISCUSSION

Several metrics were calculated to evaluate the performance of the model. These metrics support the model analysis and reflect the specific machine learning algorithms' attack detection quality. The metrics mentioned are defined as where TP, TN, FP, and FN are True Positive, True Negative, False Positive, and False Negative [25] [20]:

- Accuracy: represents the overall performance concerning the actual correct predictions calculated by using equation 4

$$\frac{TP+TN}{TP+FP+TN\_FN} \tag{3}$$

$$\frac{TP}{TP+FP}. \tag{4}$$

We examined the accuracy of each model with the different classifiers; the results are displayed in Fig. 2. The model with the highest accuracy for the RF classifiers is the set of features from the PCC-RFFI iterative approach without dropping any attacks from the dataset, followed by the chi-2 model. The DT classifier with the highest accuracy is the chi-2 model, and for the GNB, it's the PCC-RFFI without dropping any attacks.

The accuracy of the GNB is not similar to the other classifiers due to its probabilistic nature, which leads us to calculate the average for all classifiers. Table 5 and Fig. 3 show the average accuracy and precision for each model to be able to identify the best-fitting iterative feature selection approach.

Table 5. Average Accuracy and Precision for the sets of selected features

	accuracy	precision
RFFI-D	71.72%	61.38%
CHI2-D	85.98%	75.67%
PCC-D-RFFI	86.60%	94.67%
RFFI	86.86%	96.00%
CHI2	91.79%	89.64%
IG	92.32%	84.40%
PCC-D	95.90%	83.33%
PCC	96.95%	95.67%
PCC-RFFI	99.27%	97.60%

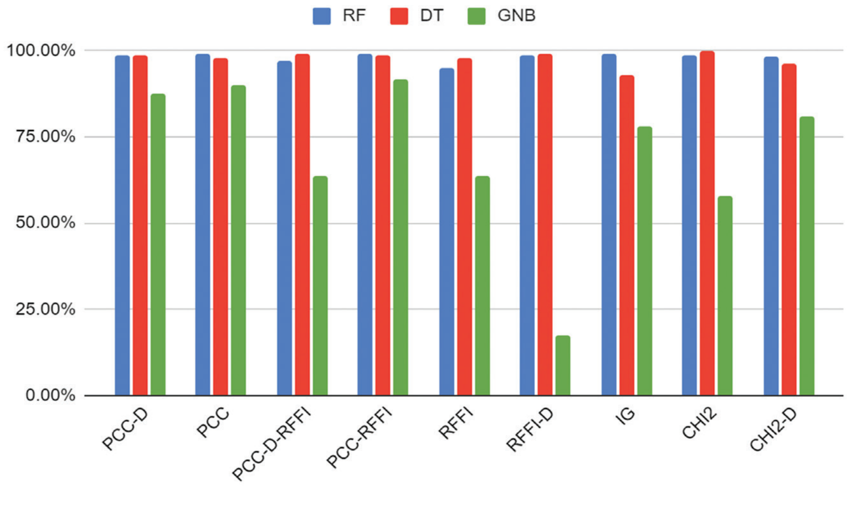

Fig. 2. Accuracy for different models and classifiers

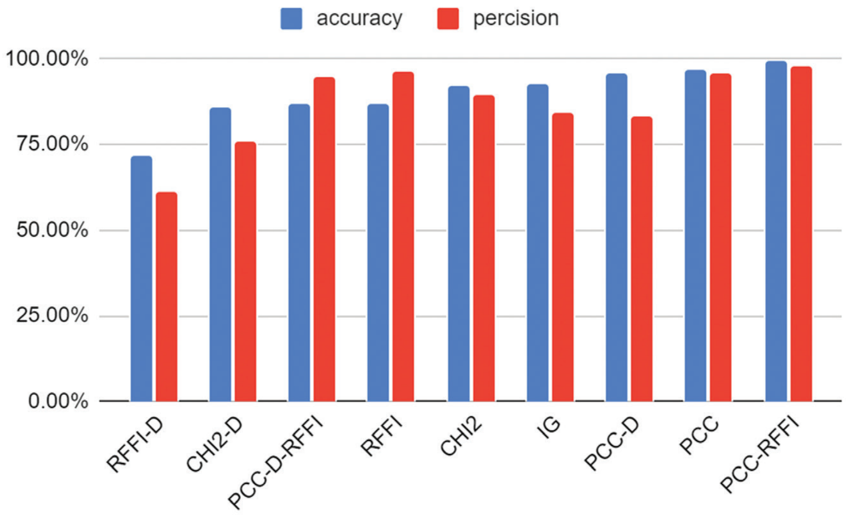

Fig. 3. Average accuracy and precision

5. CONCLUSION AND FUTURE WORK

Distributed Denial of services is one of the biggest problems we face nowadays when it comes to cloud computing and distributed environments in general, where machine learning techniques are considered one of the top ways to protect against them. The prevention and detection process of DDOS attacks using a machine-learning approach is divided into two main stages. The first is the feature selection stage, and the second is the ML classifiers trained to detect these attacks. Our focus through this framework was trying to find a more efficient way for selecting the essential features to increase the accuracy of any classifier and figuring out if the claim of dropping a percentage from the attacks in the datasets would improve the accuracy of selected classifiers or not. We created 9 sets of features from the available features in the dataset CICID2018 using 4 techniques, including PCC, RFFI, MI, and Chi-squared once, dropping 50% of the attacks and once without. This results in creating 2 sets of features for the PCC, 2 for the RFFI, 2 for the Chi2, and one for the MI, as dropping the attacks did not affect the resulting features. In the first round of feature selection, we based our selection process on the dependency of the features selecting the sets of features with the highest independence levels, which resulted frousingof the PCC approach. After that, the second selection iteration was based on the feature importance. We then compared the accuracy and precision of all the models on the DT, RF, and GNB classifiers. We calculated the average accuracy and precision for the three classifiers. In both cases, the highest average for the selected features was for the model RCC-RFFI proving that our iterative feature selection approach resulted in higher prediction accuracy. Our results also showed that, on the contrary, dropping the attacks didn't significantly impact the accuracy of the different classifiers with the other models. In our future work, we aim to test our iterative approach on more classifiers using more feature selection approaches and improve their performance.

6. REFERENCES

[1] D. R. Praveen, A. N. Rimal, "DDOS Attack Detection Using Machine Learning", International Journal of Emerging Technologies and Innovative Research, Vol. 7, No. 6, 2020, pp. 185-188.

[2] A. E. Kheder, A. M. Idress, "Adapting Load Balancing Techniques for Improving the Performance of e-Learning Educational process", Journal of Computers, Vol. 12, No. 3, 2017, pp. 250-257.

[3] A. E. Khedr, A. M. Idrees, "Enhanced e-Learning System for e-Courses Based on cloud computing", Journal of computers, Vol. 12, No. 1, 2017.

[4] S. Naiem, M. MARIE, A. E. Khedr, A. M. Idrees, " Distributed Denial Of Services Attacks And Their Prevention In Cloud Services", Journal of Theoretical and Applied Information Technology, Vol. 100, No. 4, 2022, pp. 1170-1181.

[5] P. A. Narote, V. Zutshi, A. Potdar, "Detection of DDoS Attacks using Concepts of Machine Learning", International Journal for Research in Applied Science & Engineering Technology, Vol. 10, No. VI, 2022, pp. 390-403.

[6] S. Naiem, A. M. Idress, M. Marie, A. E. Khedr, " DDOS Attacks Defense Approaches And Mechanism In Cloud Environment", Journal of Theoretical and Applied Information Technology, Vol. 100, No. 13, 2022, pp. 4632-4642.

[7] A. M. Idrees, M. H. Ibrahim, "A Proposed Framework Targeting the Enhancement of Students' Performance in Fayoum University", International

Journal of Scientific & Engineering Research, Vol. 9, No. 11, 2018.

[8] A. M. Mostafa, Y. M. Helmy, A. E. Khedr, A. M. Idrees, " A Proposed Architectural Framework For Generating Personalized Users' Query Response ", Journal Of Southwest Jiaotong University, Vol. 55, No. 5, 2020.

[9] A. M. Idrees, A. I. ElSeddawy, M. O. Zeidan, "Knowledge Discovery based Framework for Enhancing the House of Quality", International Journal of Advanced Computer Science and Applications, Vol. 10, No. 7, 2019, pp. 324-331.

[10] A. M. Mohsen, H. A. Hassan, A. M. Idrees, "A Proposed Approach for Emotion Lexicon Enrichment.", International Journal of Computer, Electrical, Automation, Control, and Information Engineering, Vol. 10, No. 1, 2016.

[11] H. A. Hassan, M. Y. Dahab, K. Bahnassy, A. M. Idrees, F. Gamal, "Arabic Documents Classification Method a Step towards Efficient Documents Summarization", International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 3, No. 1, 2015, pp. 351-359.

[12] A. E. Khedr, A. M. Idrees, A. Elseddawy, "Adaptive Classification Method Based on Data Decomposition", Journal of Computer Science, Vol. 12, No. 1, 2016, pp. 31-38.

[13] A. E. Khedr, A. M. Idrees, A. I. El Seddawy, "Enhancing Iterative Dichotomiser 3 algorithm for the classification decision tree", WIREs Data Mining Knowledge Discovery, Vol. 6, 2016, pp. 70-79.

[14] H. A. Hassan, M. Y. Dahab, K. Bahnasy, A. M. Idrees, F. Gamal, "Query answering approach based on document summarization", International Open Access Journal of Modern Engineering Research, Vol. 4, No. 12, 2014.

[15] A. M. Idrees, M. H. Ibrahim, A. I. El Seddawy, "Applying spatial intelligence for decision support systems", Future Computing and Informatics Journal, Vol. 3, 2018, pp. e384-e390.

[16] A. M. Idrees, "Towards an Automated Evaluation Approach for E-Procurement", Proceedings of the 13th International Conference on ICT and Knowledge Engineering, Bangkok, Thailand, 18-20 November 2015.

[17] A. M. Idrees, A. E. Khedr, A. A. Almazroi, "Utilizing Data Mining Techniques for Attributes' Intra-Relationship Detection in a Higher Collaborative Environment", International Journal of Human-Computer Interaction, 2022.

[18] M. Tan, A. Iacovazzi, N.-M. Cheung, Y. Elovici, "A Neural Attention Model for Real-Time Network Intrusion Detection", Proceedings of the IEEE 44th Conference on Local Computer Networks, Osnabrueck, Germany, 14-17 October 2019, pp. 291-299.

[19] C. M. Nalayinil, D. J. Katiravan, "Detection of DDoS Attacks using Machine Learning Algorithms", Journal of Engineering technologies and innovative Research, 2022, pp. 223-232.

[20] E. A. R. Coelho, "DDoS Detection using Machine Learning Techniques", Advanced information security, 2022, pp. 1-8,

[21] M. Alduailij , Q. W. Khan, M. Tahir , M. Sardaraz, M. Alduailij, F. Malik, "Machine-Learning-Based DDoS Attack Detection Using Mutual Information and Random Forest Feature Importance Method", symmetry, Vol. 14, No. 1095, 2022.

[22] Y.I. Kurniawan, F. Razi, B. Wijayanto, M. L. Hidayat, "Naive Bayes modification for intrusion detection system classification with zero probability", Bulletin of Electrical Engineering and Informatics, Vol. 10, No. 5, 2021, pp. 2751-2758.

[23] A. Mishra, "Prediction approach against DDOS Attacks Based on Machine Learning Multiclassfier", arXiv:2204.12855, 2022.

[24] Y. M. Swe, P. P. Aung, Hlaing, "A slow DDOS attack Detection Mechanism using Feature Weighing and Ranking", Proceedings of the 11th annual Internation Conference on Industrial Engineering and Operation Managment, Singaphore, 7-11 March, 2021.

[25] E. S. Alghoson, O. Abbass, "Detecting Distributed Denial of Service Attacks using Machine Learning Models", International Journal of Advanced Computer Science and Applications, Vol. 12, No. 12, 2021, pp. 616-622.

[26] D. Alghazzawi , O. Bamasag, H. Ullah, M. Z. Asghar, "Efficient Detection of DDoS Attacks Using a Hybrid Deep Learning Model with Improved Feature

Selection", Applied Science, Vol. 11, No. 11634, 2021, pp. 1-22.

[27] C. I. f. Cybersecurity, "http://www.unb.ca/cic/datasets/ids-2018.html" (accessed: 2018)

[28] I. Sharafaldin, A. H. Lashkari, A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization", Proceedings of the 4th International Conference on Information Systems Security and Privacy, Portugal, 2018. pp 108-116.

[29] A Realistic Cyber Defense Dataset (CSE-CIC-IDS2018), https://registry.opendata.aws/cse-cic-ids2018 (accessed: 2022)

[30] I. M. Nasir, M. A. Khan, M. Yasmin, J. H. Shah, M. Gabryel, R. Scherer, R. Damaševičius, "Pearson Correlation-Based Feature Selection for Document Classification Using Balanced Training", Sensors, Vol. 20, No. 6793, 2020.

[31] J. R. Vergara, P. A. Estevez, "A review of feature selection methods based", Neural Computing and Applications, Vol. 24, 2014, pp. 175-186.

[32] O. S. Bachri, "Feature selection based on Chi Square in Artificial Neural Network to Predict the Accuracy of Student study Period", International Journal of Civil Engineering and Technology, Vol. 8, No. 8, 2017, pp. 731-739.

[33] G. Karatas, "The Effects of Normalization and Standardization an Internet of Things Attack Detection", European Journal of Science and Technology, No. 29, 2021, pp. 187-192,.

[34] A. E. Khedr, A. M. Idrees, R. Salem, "Enhancing the e-learning system based on a novel tasks' classification load-balancing algorithm", PeerJ Computer Science, Vol. 7, 2021, p. e669.