IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 6 October 2023, accepted 27 October 2023, date of publication 31 October 2023, date of current version 9 November 2023.

*Digital Object Identifier 10.1109/ACCESS.2023.3328951*

RESEARCH ARTICLE

# Enhancing the Efficiency of Gaussian Naïve Bayes Machine Learning Classifier in the Detection of DDOS in Cloud Computing

**SARAH NAIEM[1], AYMAN E. KHEDR[2], AMIRA M. IDREES[2], AND MOHAMED I. MARIE[1]**
[1]Faculty of Computers and Artificial Intelligence, Helwan University, Cairo 11795, Egypt
[2]Faculty of Computers and Information Technology, Future University in Egypt, Cairo 11835, Egypt
Corresponding author: Amira M. Idrees (amira.mohamed@fue.edu.eg)

**ABSTRACT** Distributed Denial of services is one of the most dangerously planned attacks in cloud computing, resulting in huge losses of data and money for both the cloud services providers and the users of these services. Many efforts have been performed to help protect the cloud from these attacks using machine learning techniques. This study focuses on enhancing the efficiency of the Gaussian Naïve Bayes classifier, considered one of the cheapest and fastest classifiers. Still, it has some problems resulting from its equation's statistical nature. The nature of this classifier is based on multiplication, resulting in inaccurate classification due to the zero-frequency issue and the fact that it assumes that features are independent. This research proposed a framework handling the selection of a set of highly independent features following an iterative feature selection approach using the Pearson Correlation Coefficient, Mutual Information, and Chi-squared and then selecting other subsets of features from these sets to reach a set of highly independent features. After that, we used a specific algorithm handling the data pre-processing to handle the zero-frequency problem where we used the Mode to replace the missing values, and if the mode was zero, it used the mean instead. Still, if the record's label is zero, we get the value of the previous record with zero labels. After that, we handled the data imbalances using SMOTE. These enhancements increased both accuracy for the mutual information model by 2% and the average overall accuracy and precision by 1.5%.

**INDEX TERMS** Classification algorithms, machine learning algorithms, Naive Bayes classifiers, cloud computing, algorithmic efficiency, DDoS attacks.

## I. INTRODUCTION

Cloud Computing has changed how we operate daily and helped change how business functions nowadays. It has done that through all its services on demand, including software, infrastructure, and platforms. [1], [2]. Cloud computing is the foundation of our daily lives, providing different services through its infrastructure. It is prone to many security issues targeting different parts of its infrastructure and services [3]. Distributed Denial of Services (DDOS) and Denial of services (DOS) are major threats facing cloud users today. DDOS and DOS are attacks where the attacker overwhelms the targeted network, denying legitimate users from accessing the systems [4]. The attacker's motive can be anything from money, revenge, intellectual challenge, and ideological beliefs to any other reason resulting in catastrophic implications. This type of attack is very challenging to detect, resulting from its distributed nature. It is the idea of creating several bots or zombies and sending them as many requests as possible to the machine, resulting in system failure by preventing others from using the system [5]. According to the latest statistics and research, ''Cisco's analysis of DDoS total attack history and predictions,'' the total number of DDOS attacks has increased from 7.9 million in 2018 to 15.4 million in 2023. The most famous attacks that have happened recently include the Amazon web service (AWS) DDOS attack and the Google Attack in 2020, the GitHub attack in 2018, and the Mirai Krebs and OVH DDoS Attacks

The associate editor coordinating the review of this manuscript and approving it for publication was Seifedine Kadry.

in 2016 [6]. The DDOS attack tricks its victim using cloud features, including auto-scaling and multitenancy, leading them to add more resources [7], [8]. The nature of this type of attack makes it challenging to handle with traditional security solutions such as intrusion detection and firewalls in the cloud [9]. There are a lot of different approaches for handling DDOS attacks in the cloud, including their prevention, detection, and mitigation. DDOS attack prevention approaches include hidden servers/ports, restrictive access, Challenge response, and resource limitation. In contrast, the detection approaches include signature-based, resource usage, anomaly-based, hybrid detection, count-based filtering, and source and spoof tracing [10]. The Anomaly-Based detection technique relies on machine learning and deep learning, which have proven successful in creating a robust detection technique, particularly in cloud computing [11], [12]. There are different types and approaches to machine learning (ML) algorithms. ML approaches include supervised, unsupervised, semi-supervised, and reinforcement learning [13], [14]. Supervised learning is used for training labeled data where the training and validation data are defined; it includes classification where the output has discrete values [15], [16] and regression where the output has continuous values like Random Forest (RF), Support Vector Machine (SVM), Decision Trees (DT), linear and logistic regression, and Gaussian Naïve Bayes (GNB) [17], [18]. In contrast, unsupervised learning is based on having unlabeled and unstructured data clustered to find the unseen behavioral patterns in the data by grouping diverse data patterns for the association technique that figures a way to state the relation between the different data clusters, K-mean clustering, and hierarchical clustering [19], [20]. Semi-supervised learning combines both techniques, using the prediction labels from the unsupervised and supervised techniques. In contrast, the reinforcement technique learns based on the interaction in the environment based on Reward Feedback to capture the behavior or pattern in data as it takes the data to learn it and then adds it to the training Data. Through this paper, we are focusing on improving the efficiency of the GNB ML algorithm for the detection of DDOS in cloud computing, which could be achieved through better selection quality of the features and the data pre-processing phase. The rest of this paper will include a literature review of the related work for different ML approaches and methods, followed by the proposed framework along with the experimental results and discussion, and finally, the conclusion.

## II. LITERATURE REVIEW

Many efforts have been made to help prevent DDOS by detecting them using machine learning techniques. Marvie et al. [21], aimed to enhance detection models in ML by plummeting the number of features used in the model using two feature selection techniques, including filter and embedded. They reduced the features to 20 features of the CICDDoS2019 through the f-test and random forest and trained the model using light gradient boosting algorithms (LGBM). They proved an almost 20% more enhancement than other proposed frameworks and concluded that reducing the feature space to 77% improves the performance [21].

Batchu and Seetha in [22] used hybrid feature selection and hyperparameter tuning to improve their model using the CICDDos2019 dataset. They tested their contribution using five different ML classifiers, including K Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision Tree (DT), Logistic Regression (LR), and Gradient Boosting (GB). They handled data imbalance using SMOTE, and in the data cleaning process, they substituted the negative values with zero and the missing and infinite values with the median. The selected features were by selecting the highest-ranking features from the most independent ones. The model testing and training were conducted on different combinations in the data set from different days, and the results showed that their model increased by at least 0.8% that most of the different state-of-the-art studies. The same authors created a different approach in [23], where they first handled the class imbalance in their data using Adaptive Synthetic Oversampling and using the Shapely Value (SHAP) and the Local Agnostic Methods (LIME) for feature selection. SHAP is used for its ability to global and local interpretations in the feature selection, while LIME is used for its ability in local interpretation. Before the data cleaning step, they removed some of the features irrelevant to the model. In the data cleaning part, they removed the duplicate data and didn't remove any other instances; they added zeros for the negative data and the mean of the feature for any missing and infinity values. They tested their model using the KNORA-E and KNORA-U ensemble classifiers on balanced and imbalanced data, proving that it created an enhanced model based on binary classification. [23].

In another study by [24] the authors used a hybrid feature selection method by applying the Mutual Information (MI) selection approach and RFFI approach, resulting in different sets of 16 and 23 features using MI and 19 using RFFI. These sets were tested on five different classifiers, including RF, GB, KNN, LR, and Weighted Voting Ensemble (WVE) classifiers, generally resulting in a higher misclassification rate where all accuracy was nearly 99% and the highest accuracy rate of 99.993% was achieved from using the RF classifier with the 19 features selected using the RFFI [24].

Dasari and Devarakonda [25] created a framework for detecting different types of DDOS on the CICDDos2019 dataset using six different classifiers, including DT, RF, KNN, Naïve Bayes (NB), AdaBoost, and GB. They removed the socket features and missing values along with encoding the data and checked the accuracy and execution time of each classifier. Even though the authors didn't mention the specific feature selection approach, they concluded from the ROC score that ADA boost and LR are the highest, followed by the NB, RF, and KNN, while the DT had the lowest score.

In [26], the authors created a DDoS model using actual data from attack traffic. The data set had 25 features extracted

using the CICFlowMeter-V3 application. They selected the top features using the forward linear regression (FLR)method based on Mutual information. The model was tested on NB, Random Forest (RF), Neural Network (NN), SVM, and KNN classifiers where RF and NN had the highest accuracy at 98.7%, and the SVM was only 0.3% lower, followed by the NB at 97.96% and KNN at 97.63%. They concluded that the FLR feature selection approach works best with the RF and a 98.4% weighted Average accuracy.

The authors' focus in [27] was Gaussian Naïve Bayes (GNB) accuracy enhancement. First, the KDD 99 dataset was pre-processed using correlation-based feature selection (CSF) to select important features for the research. Then, by eliminating the zeroes from the dataset and changing the GNB statistical equation from multiplication to addition, the GNB classifier was enhanced, increasing accuracy by about 4% without following the enhanced approach. Even though the authors didn't focus on feature selection in their approach, the result from their model fixed the zero-probability problem of the GNB.

The authors of [28] utilized K-fold cross-validation on the CIC-IDS2017 dataset, dividing it into five sets: four for training and one for testing. The value of K was set to 5. They tested the model using various algorithms, such as RF, KNN, DT, GNB, SVM, and LR, including all three kernels (sigmoid, polynomial, and R). The classifier's accuracy, precision, and recall were compared, and the authors concluded that DT provided the best results. Although all other classifiers ranged from 99.78% to 99.88%, DT's accuracy was 99.94%. However, the DT is the best classifier. GNB's accuracy was particularly low at 61.22%, indicating that cross-validation for feature selection is inappropriate for GNB. This may be due to the level of independence of the features or the probabilistic nature of the classifier.

In [29], an iterative feature selection approach was proposed using the CICDDoS2018 dataset where the authors applied four different feature selection approaches, including Pearson Correlation Coefficient PCC, RFFI, MI, and Chi-square and created four different sets of features, selected the set of features with the highest independence which was the PCC set of features and then used the other three techniques to select a subset of features for each technique resulting in 3 more subsets. After that, they tested them using DT, RF, and GNB. The set of features from using the PCC with the RFFI techniques resulted in the highest average accuracy of 99.27% and precision of 97.6%. It was highlighted that although they used the average accuracy and precision for the different classifiers, the GNB classifier had the lowest accuracy.

Shu et al. [30] used the double weighted Naïve Bayes for their fire prediction model, handling the imbalances in the attributes through the help of ''Laplace'' smoothing and logarithmic operations. As mentioned, the logarithmic operation converts the multiplication into addition to avoid the zero-product problem. In contrast, the ''Laplace'' smoothing is used to initialize the values of each attribute to 1 without affecting the final classification problem, where they calculated the coeffect probability for each category with its corresponding attribute probability and the weighted coefficient that are then combined. The Double weighted Naïve Bayes algorithm resulted in an average increase in accuracy from different test runs on different data sets of 2.56%. In [31], the authors used SMOTE and the Genetic algorithm. In the first stage, they used the SMOTE concept to make synthetic data for the minority class to balance the unbalanced data. Then, the genetic algorithm was used for attribute selection, resulting in an increase of 4.8% in accuracy for the GNB, where they tested their work on the German credit data from the UCI machine learning raspatory.

It is quite clear from our literature review that to enhance any ML-based detection approaches; it is important to consider the selected features and their impact on the accuracy of different classifiers. In addition, it was highlighted that the accuracy of the GNB is relatively lower than the rest of the classifiers, whether supervised or unsupervised. Therefore, in the following sections, we will first discuss the nature of the GNB classifier algorithm and why there are still problems with their accuracy. Then, our proposed model and details will be discussed.

## III. GAUSSIAN NAÏVE BAYES CLASSIFIER ALGORITHM

Gaussian NB supports continuous data derived from the Gaussian normal distribution 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, based on the assumption that data is normally distributed.

Figure 1 illustrates how Gaussian Naïve Bayes works [27].

The Bayes theorem is based on the multiplication of the likelihood and prior divided by the evidence, if each selected feature contributes equally and is independent. The assumption that each feature's importance has the same effect on the outcome takes into consideration that they are not dependent on each other. The GNB is the probability of a given event, given that another event has already happened.

$$\text{Option} : P(A \mid B) = \frac{P(B \mid A)\, P(A)}{P(B)} \tag{1}$$

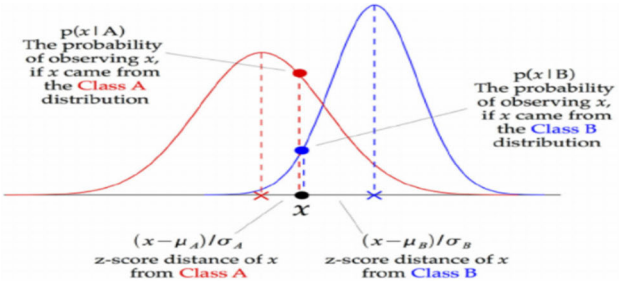


**FIGURE 1.** Gaussian NB equation explanation.

This equation gets the probability that A happens given that B happens, which is the posterior probability, while the probability of B happens given that A happens. P(B|A) is the likelihood multiplied by the prior, which is the probability of A happening P(A) divided by the evidence, which is the probability of B happening P(B).

As for the GNB machine learning classifier, the Bayes theorem is used to predict a class of an unknown record given a set of features related to a specific class. Even though the GNB classifier has many advantages and is used in problems with multi-class prediction, it's fast and has high accuracy with independent features. There is a significant issue with the formula used for predictions, caused by the concept of zero probability. When a value is missing, it is given a zero known as the "zero frequency" phenomenon. This is problematic because the formula includes the multiplication of different probabilities based on the variance and mean of the features. When any of these probabilities are zero, the prediction will also be zero [30], [31], [32], [33].

## IV. PROPOSED FRAMEWORK

Our proposed framework aims to improve the accuracy of the GNB, where we are handling the zero-probability problem through the data pre-processing phase. Moreover, we considered selecting the set of features with the highest level of independence by applying the iterative feature selection approach we proposed in [29]. Each stage of our framework is described in detail in the following sections.

### A. DATA SET

We used the CICD2018 open-source dataset made available through the University of New Brunswick and presented in 7 CSC files presenting different traffic days, including the 15th, 16th,20th, 21st, 22nd, and 23rd of February 2018. These seven files include traffic representing normal traffic and seven different attacks, including DOS, Web-attacks, DDOS, infiltration, Botnet, and Brute force, displayed with 80 attributes representing the dataset's features. The traffic distribution in the dataset is 83.07% benign or normal traffic, and the remaining 16.93% is for the attacks. These attacks are not equally distributed; the two highest attacks are DDOS, with 7.79% of the traffic, and DOS, with 4.03% [34], [35], [36].

### B. DATA PRE-PROCESSING AND CLEANING

This is the most vital part of our framework. We handle this phase to solve the zero-frequency problem in the GNB classification algorithm resulting from the multiplication of zeros from the dataset. The first step is data encoding, replacing the label column, which states if this instance is benign or an attack with 0 if it's benign and one if it's an attack. After that comes the data-cleaning part, which we consider the framework's core. As mentioned before in the literature review section, many efforts have been done to fix this problem, from removing all the zeros in the data to converting the GNB algorithm equation to avoid multiplying by zero and handling imbalanced data [27], [30], [31]. The flow chart for this phase is displayed along with the pseudocode explaining the steps involved in the data cleaning in Figure 2 and Table 1.

The idea of replacing the nan and missing values with the mode of the data instead of zero is to solve the zero-frequency problem occurring from the multiplication in the GNB classifier. In addition, using the mode is preferable to using the mean as the mean encounters the outliers in the data affecting the classification's accuracy [37]. We replaced zeros in the dataset with the mode only if the label was 1; no changes were made if the label was zero. Additionally, if the mode of a column was zero, we used the mean instead of using the zero from the mode as it would contradict the actual purpose of keeping the structure of the data to handle the GNB's accuracy. We used the mode to replace the nan, infinite, and -negative infinite for the non-attack record, while if it was an attack, we replaced the value with the same value of the previous attack record in the dataset. If there are no previous records with values for the attacks, we used the mode or mean if the mode was zero. This helps a lot in keeping the actual shape and meaning of the date and simultaneously solving the zero-frequency problem. In our experiment, we used the mean, the average value for each attribute, the mode, the number repeated the most, and the median, the middle value of the dataset [37]. The mode improved the best accuracy

**TABLE 1. Proposed framework pseudocode.**

| Pseudocode |
|---|
| Replace the label column with zero values for the benign traffic and one for attacks<br>Replace the infinite and – infinite values with nan<br>Get the mode of each column in the data set<br>Check mode value<br>  If mode == 0<br>  Use Mean value for this column<br>Save the value and index of the previous non-nan value for each column<br>Loop on data till a Nan OR ZERO value is found<br>  If value ==0<br>  Check label<br>  If label ==0<br>  Ignore and go to next record<br>  Else<br>  Use mode or mean if mode is zero and replace the zero value<br>  Else if value == nan<br>  Check label<br>  If label ==0<br>  Use mode or mean if mode is zero and replace the zero value<br>  Else<br>  Check the value of the previous non-benign<br>  If value !=0<br>  Use value to replace the nan<br>  Else<br>  Use mode or mean if the mode is zero and replace the zero value<br>Last record<br>End |

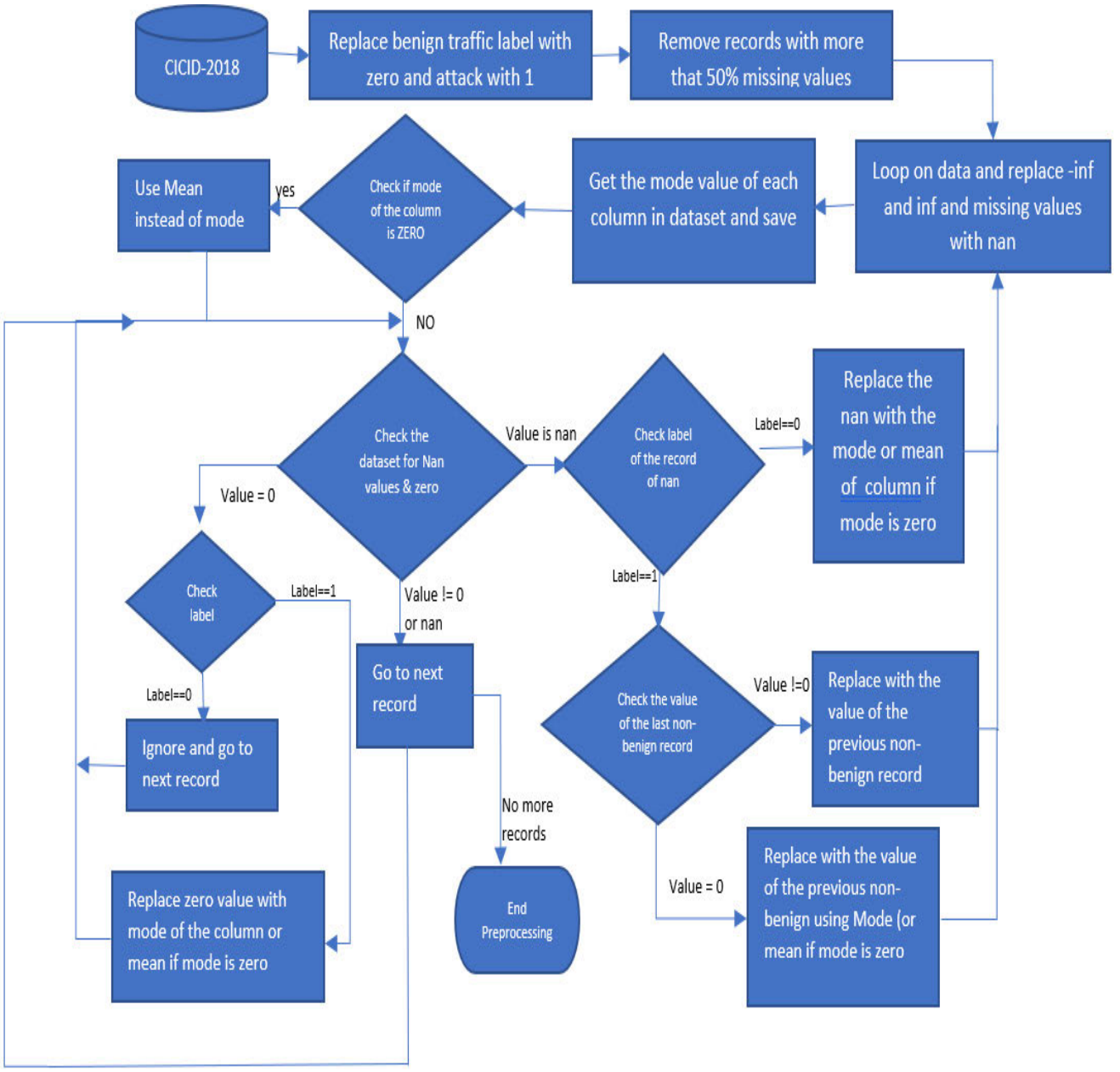


**FIGURE 2.** Data-preprocessing algorithm.

over using the mean or median by almost 0.35%. The average overall accuracy improvement from using the mode in our proposed framework was around 2%.

### C. FEATURE SELECTION

The feature selection stage is a very important step as it handles the problem of overfitting resulting from the data's high dimensional nature; it also helps with the training time of the data as it is directly proportional to the dataset's size [38], [39]. Different feature extraction techniques include wrapper, filter, and embedded methods. The Filter methods are fast and cheap as they are based on statistical measures instead of cross-validation performance, including methods like chi-square, correlation coefficient (CC), information gain, and Fisher's score. The wrapper method explores all feature subsets, trains and evaluates a classifier for each subset to determine its quality. However, this method can be computationally expensive. Some examples of wrapper methods are forward selection, backward selection, exhaustive feature selection, and recursive feature elimination. Finally, the embedded method works in an iterative sense that takes care of each iteration of the model training process and carefully extracts those features that contribute the most to the training for each iteration. This

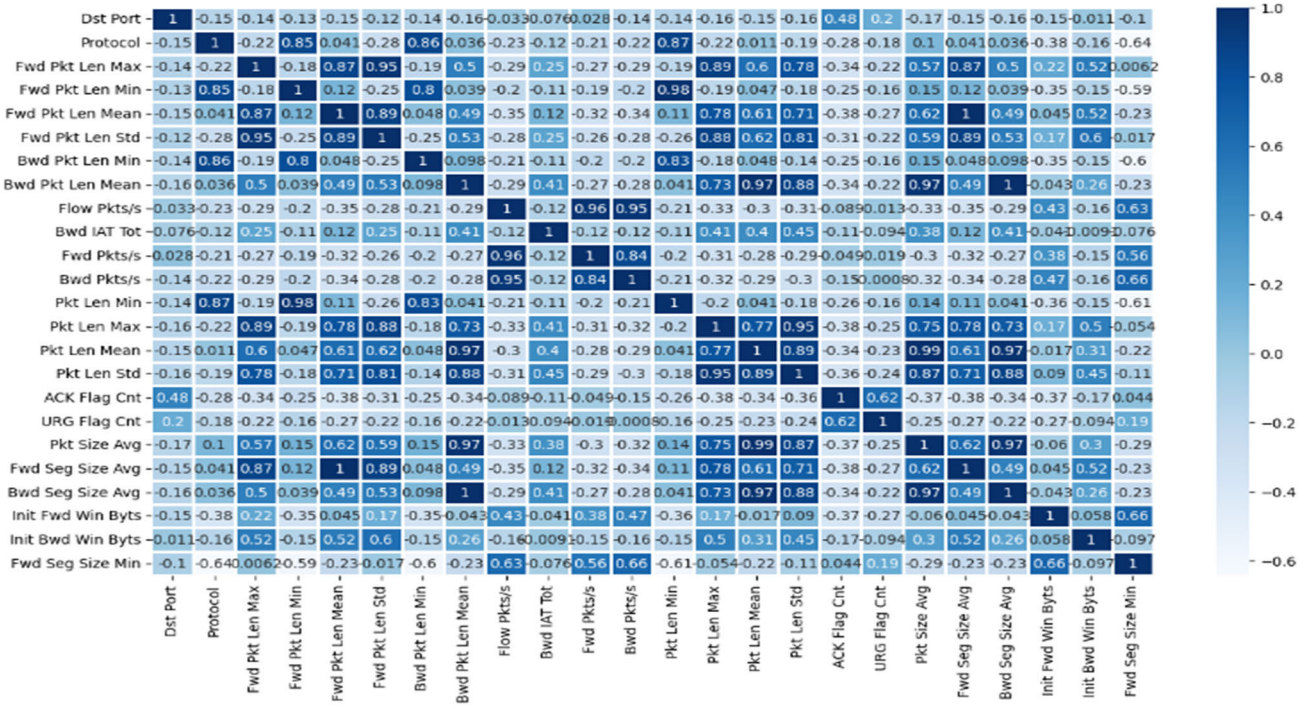


**FIGURE 3.** Heatmap for PCC features dependency.

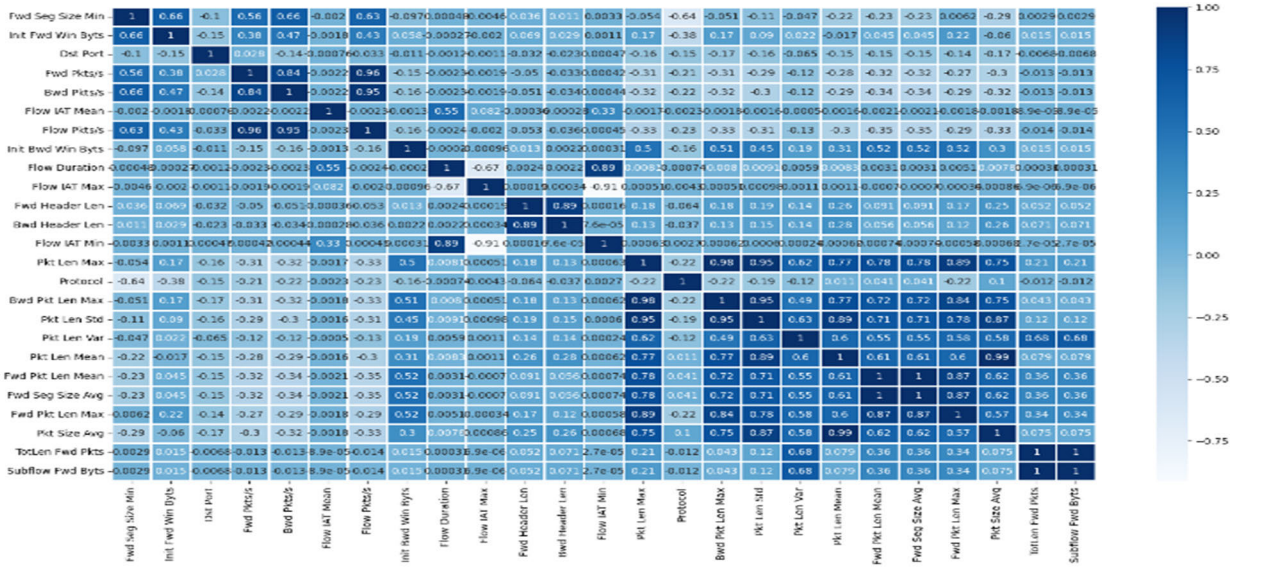


**FIGURE 4.** Heatmap for MI features dependency.

method encompasses the benefits of both the wrapper and filter methods and maintains reasonable computational costs examples include Random Forest Feature Importance (RFFI) and LASSO Regularization (L1) [24], [26], [40], [41].

As stated in the literature review section, authors in [21], [22], [24], and [29] highlighted that the use of more than one approach results in better accuracy and reliable classification. As mentioned in our previous research [29] we used the same iterative approach. Still, we eliminated the part of dropping some of the attacks in the dataset mentioned in [42] by Tan et al. since it didn't result in better accuracy. We first applied PCC, MI, and Chi-square, which resulted in 3 different subsets of features with 24, 20,30. We then applied the MI on this subset of features derived from using the PCC and chi2, resulting in two other subsets of features PCC-MI and Chi2-MI with 16 and 23 features. The set of features

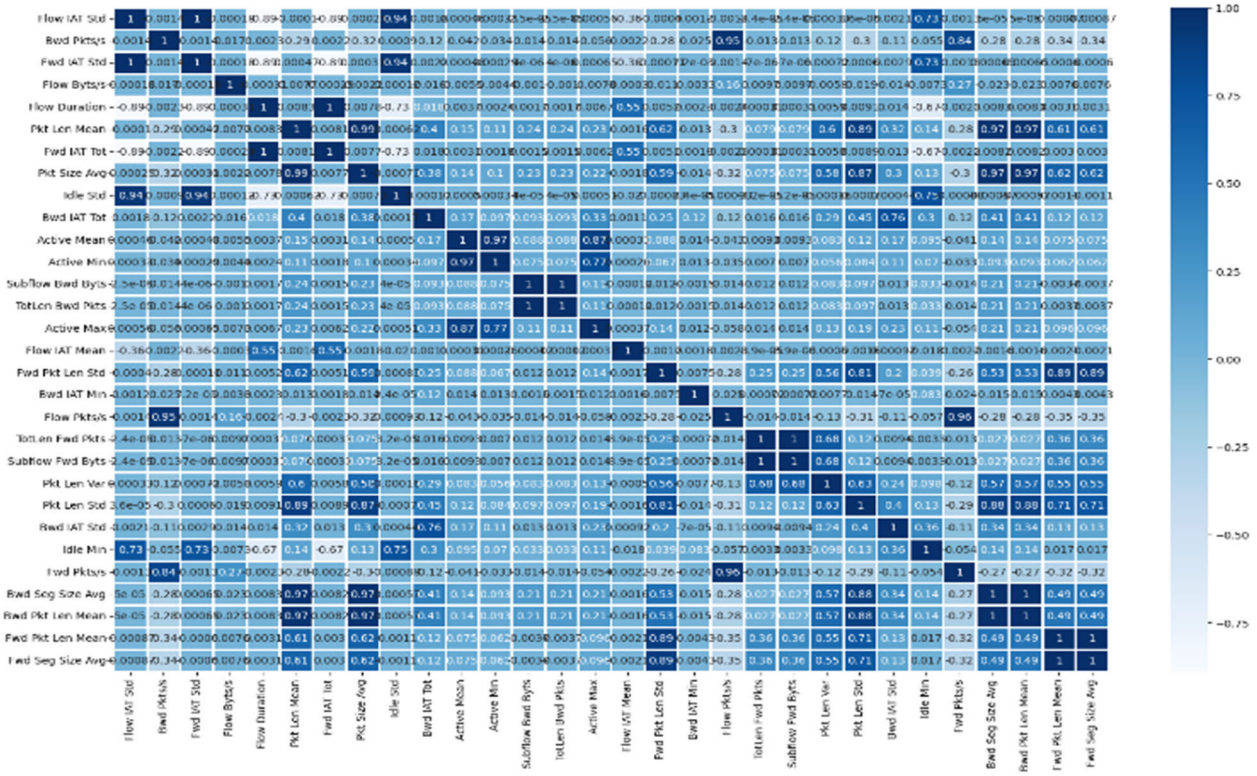

**FIGURE 5.** Heatmap for Chi2 features dependency.

with the highest accuracy was the PCC- MI, with 16 features. Even though in the testing phase, we applied the random forest feature selection (RFFI), which is an embedded feature selection approach that resulted in 12 features and applied the RFFI on the set of features selected from the PCC and the PCC- RFFI resulted in 9 features we decided to not use the RFFI in our testing for the GNB classifier. While the RFFI and PCC-RFFI features resulted in very good accuracy for the DT, KNN, RF, and SVM, the GNB classifier had better results with the filter feature selection approach rather than using filter and embedded. This is due to the nature of the GNB classifier, as mentioned in [27] and [31] where the feature's independency improves the accuracy, and this is achieved through filter-based feature selection techniques [26], [41]. We used the heat map to view the feature's independence to select the best set of features to work on in our model; figures from 3 to 7 shows the results of these features' independence using the heatmap.

### D. DATA NORMALIZATION

This stage comes before the training and testing stage of the classifier, whose main purpose is to transfer the data into a format that scales the features. Since the original data comes in different formats and forms normalization is used to maintain data meaning and performance. MinMAxScaler and StandardScaler are some of the most commonly used for normalization, and both could be used with the GNB, resulting in the same accuracy [32], [33], [43]. In the standard scaler, the data is scaled in a range representing the highest and lowest values. While the MinMax scaler shows the data within the range of one to zero, one representing the highest value and zero representing the lowest value [29], [44], [45].

MinMax Scaler equation

$$X' = \frac{x - min(x)}{(x) - min(x)} \tag{2}$$

Standard Scaler equation

$$z = \frac{x - \mu}{\sigma} \tag{3}$$

where x is the score, is $\mu$ the mean, and $\sigma$ is the standard deviation [44].

### E. HANDLING DATA IMBALANCE

In our framework we applied Synthetic Minority Oversampling Technique- SMOTE to handle data imbalances and over-fitting, which greatly impacts classification correctness and accuracy [46]. As previously mentioned in [23], [22], and [31] that handling imbalances in data using SMOTE showed promising improvements, especially in the GNB classification problem. Data imbalance results from the uneven classification classes, in our case, the distribution of attack. Our dataset's attacks to normal traffic ratio was almost 2:1, resulting in a biased classification model. The improvement resulting from handling the data imbalances can be measured

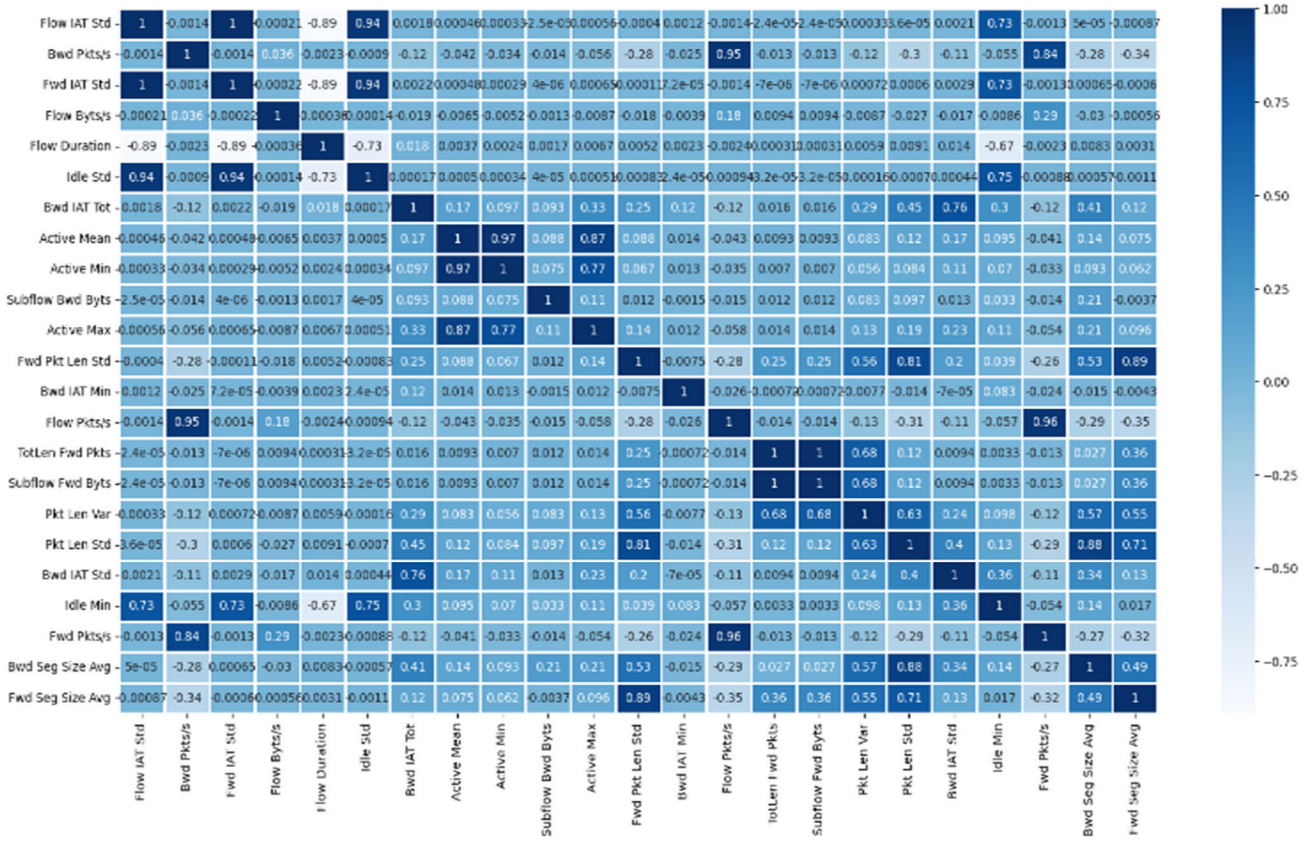


**FIGURE 6.** Heatmap for Chi2-MI features dependency.

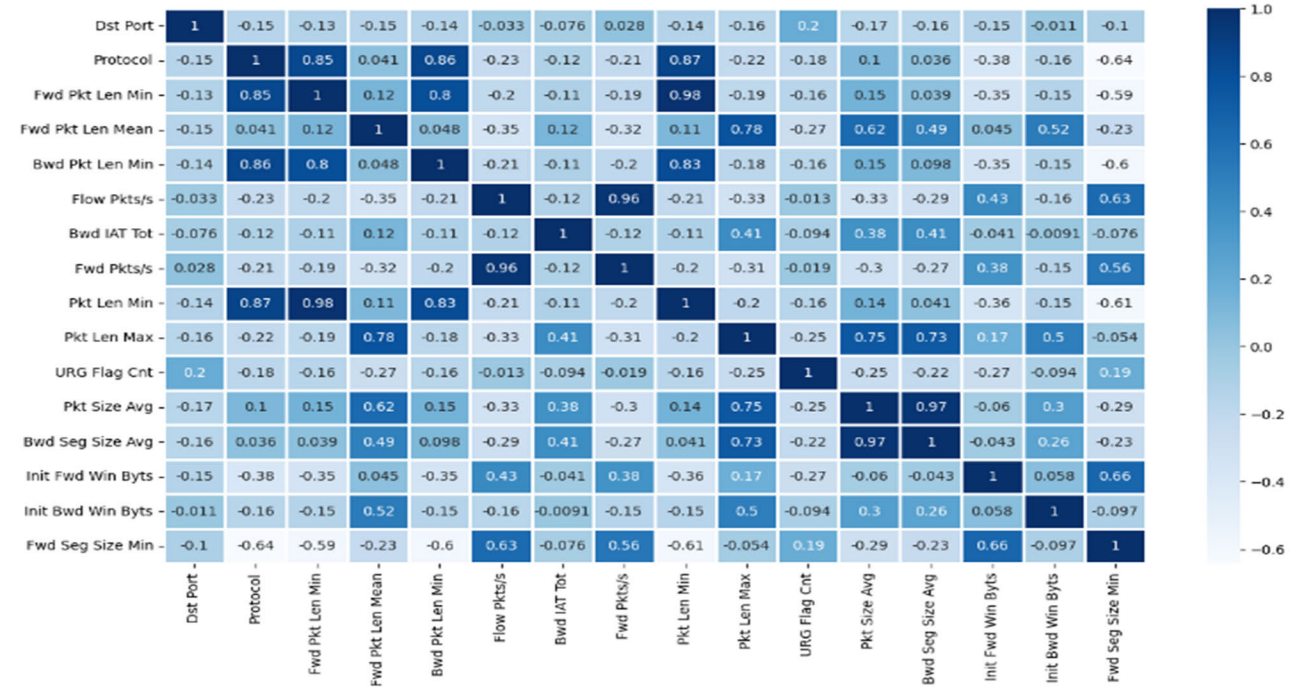


**FIGURE 7.** Heatmap for PCC_MI feature dependency.

**TABLE 2.** Overall evaluation of the model before and after applying the framework to the GNB.

| | | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| Original | PCC | 93.22% | 89.85% | 99.88% | 94.60% |
| | PCC-MI | 96.15% | 96.20% | 99.68% | 97.91% |
| | Chi2 | 76.47% | 64.45% | 99.84% | 78.34% |
| | MI | 79.44% | 68.88% | 99.96% | 81.56% |
| | Chi2-MI | 62.03% | 41.05% | 98.56% | 57.96% |
| Edited | PCC | 94.21% | 91.26% | 99.90% | 95.39% |
| | PCC-MI | 97.57% | 97.44% | 99.98% | 98.69% |
| | Chi2 | 77.89% | 66.35% | 99.85% | 79.73% |
| | MI | 81.51% | 71.18% | 99.97% | 83.16% |
| | Chi2-MI | 62.85% | 41.58% | 99.99% | 58.74% |

**TABLE 3.** Increased improvement in accuracy, precision, recall, and f1-score for each feature model using the GNB classifier.

| | *PCC* | MI | CHi2 | PCC-MI | CHi2-MI |
|---|---|---|---|---|---|
| **Accuracy** | 0.9904% | 2.0668% | 1.4191% | 1.4217% | 0.8245% |
| **Precision** | 1.4111% | 2.3039% | 1.8997% | 1.2411% | 0.5332% |
| **Recall** | 0.0226% | 0.0158% | 0.0155% | 0.2992% | 1.4334% |
| **F1-Score** | 0.7867% | 1.5988% | 1.3920% | 0.7848% | 0.7793% |

by evaluating the precision, recall, and F-score, which are discussed in more detail in the results section.

### *F. GNB CLASSIFICATION*

After applying SMOTE to handle the data imbalance in this framework stage, we applied the proposed data processing approach to different features, including those extracted using the PCC, MI, Chi-2, PCC-MI, and Chi2-MI. We then ran the GNB classifier, which improved accuracy for all the feature sets ranging between 1.2 and 2%, and the results are discussed in the next section. We also applied the proposed framework to other classifiers, including RF, DT, KNN, and SVM using the PCC, MI, and PCC-MI sets of features.

## V. MODEL EVALUATION

We calculated the accuracy and precision of each set of features from the confusion metrics. The confusion metrics are a calculation based on the values of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN). These values help with model evaluation and give insight into the results of the classifiers, helping in the analysis process of any machine learning model [29], [43].

Accuracy: Is calculated using TP, TN, FP, and FN as represented in equation (4) to reflect the precise predictions

$$TP + TN / TP + FP + TN + FN \quad (4)$$

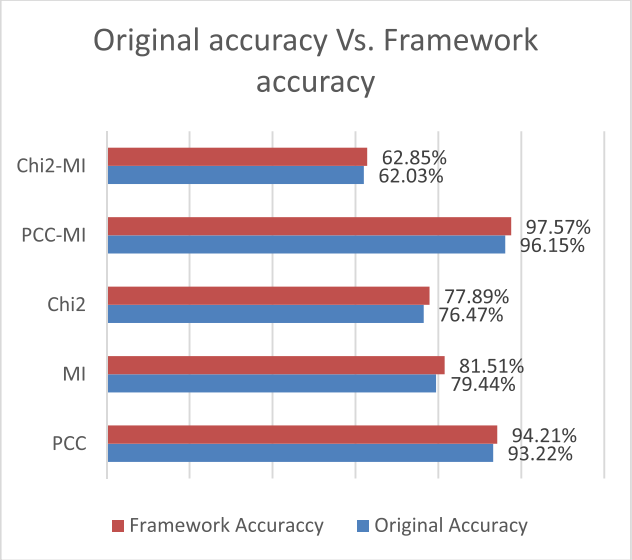


**FIGURE 8.** Original accuracy vs. Frameworks accuracy.

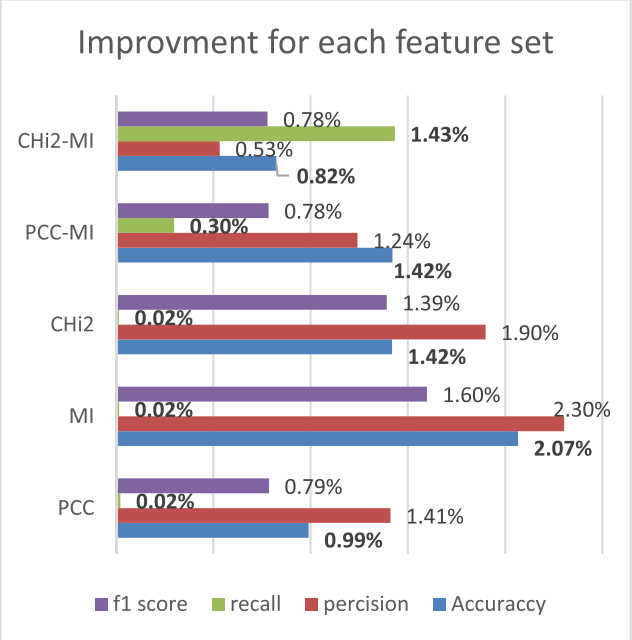


**FIGURE 9.** Improvement resulting from applying the framework for each feature set.

**TABLE 4.** Average overall improvement resulting from our framework on GNB.

| | |
|---|---|
| Accuracy | 1.34% |
| Precision | 1.48% |
| Recall | 0.36% |
| Fl-score | 1.07% |

Precision: Is calculated using the TP and FP as shown in equation (%) to show the ratio of the TP to the total

**TABLE 5.** Applying proposed framework to different classifiers.

| | Original Results | | | Framework Results | | |
|---|---|---|---|---|---|---|
| | PCC | MI | PCC-MI | PCC | MI | PCC-MI |
| RF | 98.98% | 98.07% | 99.38% | 99.24% | 98.53% | 99.89% |
| DT | 97.05% | 96.98% | 97.35% | 97.45% | 97.68% | 97.86% |
| KNN | 97.35% | 96.91% | 97.58% | 97.95% | 97.57% | 98.28% |
| SVM | 94.89% | 94.45% | 95.29% | 95.19% | 94.99% | 96.25% |
| GNB | 93.22% | 79.44% | 96.15% | 94.21% | 81.51% | 97.57% |

**TABLE 6.** Average accuracy improvement for different classifiers resulting from the framework.

| | RF | DT | KNN | SVM | GNB |
|---|---|---|---|---|---|
| Accuracy improvement | 0.41% | 0.54% | 0.65% | 0.60% | 1.49% |

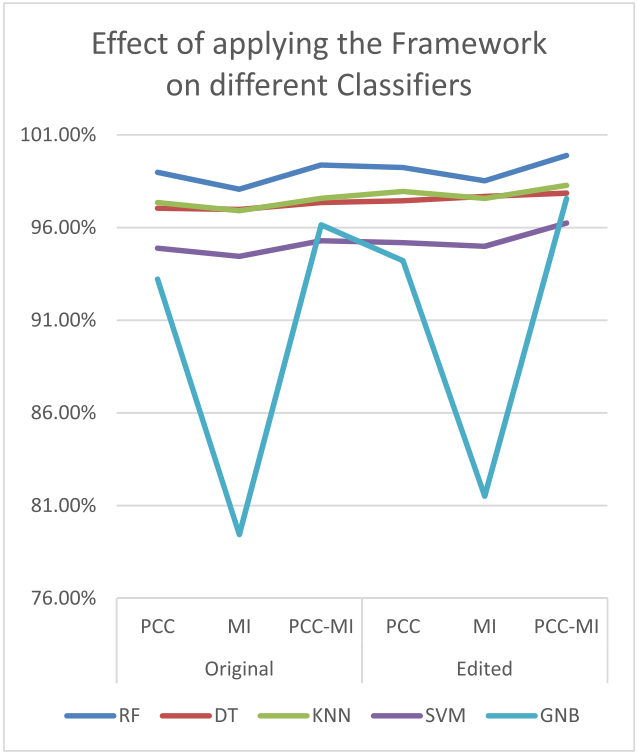


**FIGURE 10.** Applying the framework to different classifiers.

correct predictions

$$TP/TP + FP \tag{5}$$

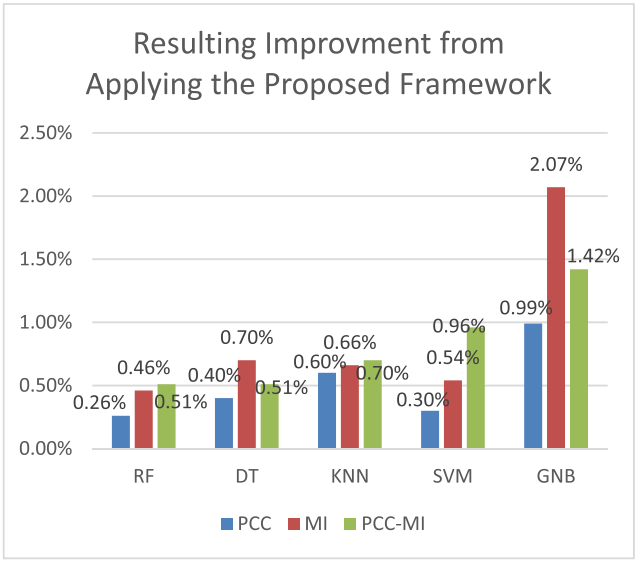


**FIGURE 11.** Resulting improvement from applying the framework to different Classifiers.

Recall: is the ratio of true positives to the total of true positives and false negatives

$$\frac{TP}{TP + FN} \tag{6}$$

F1 Score: is the weighted mean of recall and accuracy. It set the higher the model's performance, the closer the F1 score value is to 1.0.

$$2 * \frac{Percision * Sensitivity}{Percision + Sensitivity} \tag{7}$$

The results displayed in Table 2 show the overall values for accuracy, precision, recall, and F1-score. In contrast, Table 3 reflects the difference or improvement for each set of features,, and Table 4 shows the average improvement for each model for the GNB classifier. Table 5 shows the resulting improvement in accuracy from our framework using RF, DT, SVM, KNN, and GNB using the PCC, MI, and PCC_MI feature selection techniques, and Table 6 shows the average overall improvements in accuracy for different classifiers. Figure. 8 reflects the accuracy before and after applying the framework for each set of features. Fig 9 reflects the improvement in accuracy, precession, recall, and F1socre for each set of features for the GNB. Figures 10 and 11 reflect the results for the framework using different classifiers.

## VI. CONCLUSION

Through this study, we proposed an approach for improving the Gaussian Naïve Bayes machine learning classifier's accuracy for detecting DDOS in the cloud, considered one of the easy and simple classifiers. Still, it suffers from low accuracy when compared to other classifiers. The GNB classifiers are a fast and efficient detection approach with a huge drawback known as the "zero-frequency" phenomenon. In addition, the accuracy in the GNB classifier could be highly affected by the

independence of the features; that's why working with a set of features with high independency highly affected its accuracy. Through our work, we proposed a data pre-processing approach to handle the zero-probability problem where we used the mode for each feature to replace the zeros only if the traffic was not an attack. While for each zero or nan value for an attack, the value of the previous Non-NAN value was used to replace it. Applying this approach resulted in an average increase of around 1.4% in accuracy, 1.5% in precision, 0.35% in recall, and 1.07 for the F1-score. The maximum accuracy improvement was 2.07% for the MI feature set's accuracy. We also used an iterative feature selection approach based on the filter method rather than wrapper and embedded approaches. We selected the set of features with the highest accuracy. Using the approach that resulted in the second-highest accuracy, we selected another set of features from the initial set. Even though the model had the best impact on the features selected using MI, the overall highest accuracy was achieved using the PCC-MI features, indicating that the iterative approach helps select the most relevant and accurate set of features. This proves that the feature's independency greatly impacts the performance, correctness, and accuracy of GNB classifiers, which can only be achieved using filter-based methods along with zero-frequency problem handling. The proposed framework works well with different classifiers, including KNN, RF, DT, and SVM, which are powerful ML classification techniques. As we can see from the results, the GNB classifier always had the lowest accuracy over other classifiers, making it the focus of our research. Applying the framework reflected the highest improvement for the GNB, making it reach a level of accuracy close enough to other classifiers. From the results, we can see that applying the iterative and preprocessing approaches resulted in an accuracy of 96.15 % higher than the SVM, which had 95.29% accuracy.

## REFERENCES


[1] A. E. Kheder and A. M. Idress, "Adapting load balancing techniques for improving the performance of e-learning educational process," *J. Comput.*, vol. 12, no. 3, pp. 250–257, 2017.

[2] A. E. Khedr and A. M. Idrees, "Enhanced e-learning system for e-courses based on cloud computing," *J. Comput.*, vol. 12, no. 1, pp. 10–19, 2017.

[3] M. I. T. Hussan and A. A. Ahmed, "Cloud computing: Study of security issues and research challenges," *Int. J. Adv. Res. Comput. Eng. Technol.*, vol. 7, no. 4, pp. 362–369, Apr. 2018.

[4] D. R. Praveen and A. N. Rimal, "DDOS attack detection using machine learning," *J. Emerg. Technol. Innov. Res.*, vol. 7, no. 6, pp. 185–188, 2020.

[5] S. Naiem, M. Marie, A. E. Khedr, and A. M. Idrees, "Distributed denial of services attacks and their prevention in cloud services," *J. Theor. Appl. Inf. Technol.*, vol. 100, no. 4, pp. 1170–1181, Feb. 2022.

[6] P. Nicholson. (Jan. 21, 2022). *A10 Blog/Network Security/Five Most Famous DDOS Attacks and Then Some*. Accessed: May 10, 2023. [Online]. Available: https://www.a10networks.com/blog/5-most-famous-ddos-attacks/#:~:text=The%20Top-Five%20Most%20Famous%20DDoS%20Attacks%20%28for%20Now%29,CloudFlare%20DDoS%20Attack%20in%202014%20..%20More%20items

[7] H. Abusaimeh, "Distributed denial of service attacks in cloud computing," *Int. J. Adv. Comput. Sci. Appl.*, vol. 11, no. 6, pp. 163–168, 2020.

[8] A. Bakr, A. A. A. El-Aziz, and H. A. Hefny, "A survey on mitigation techniques against DDoS attacks on cloud computing architecture," *Int. J. Adv. Sci. Technol.*, vol. 28, no. 12, pp. 187–200, 2019.

[9] P. A. Narote, V. Zutshi, and A. Potdar, "Detection of DDoS attacks using concepts of machine learning," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 10, no. 6, pp. 390–403, Jun. 2022.

[10] S. Naiem, A. M. Idress, M. Marie, and A. E. Khedr, "DDOS attacks defense approaches and mechanism in cloud enviroment," *J. Theor. Appl. Inf. Technol.*, vol. 100, no. 13, pp. 4632–4642, 2022.

[11] A. M. Idrees, M. H. Ibrahim, and A. I. El Seddawy, "Applying spatial intelligence for decision support systems," *Future Comput. Informat. J.*, vol. 3, no. 2, pp. 384–390, Dec. 2018.

[12] A. M. Idrees, A. I. El Seddawy, and M. Ossama, "Knowledge discovery based framework for enhancing the house of quality," *Int. J. Adv. Comput. Sci. Appl.*, vol. 10, no. 7, pp. 324–331, 2019.

[13] A. E. Khedr, A. M. Idrees, and F. K. Alsheref, "A proposed framework to explore semantic relations for learning process management," *Int. J. e-Collaboration*, vol. 15, no. 4, pp. 46–70, Oct. 2019.

[14] A. A. Qaffas, I. Alharbi, A. M. Idrees, and S. A. Kholeif, "A proposed framework for student's skills-driven personalization of cloud-based course content," *Int. J. Softw. Eng. Knowl. Eng.*, vol. 33, no. 4, pp. 603–617, Apr. 2023.

[15] A. A. Almazroi, A. E. Khedr, and A. M. Idrees, "A proposed customer relationship framework based on information retrieval for effective Firms' competitiveness," *Expert Syst. Appl.*, vol. 176, Aug. 2021, Art. no. 114882.

[16] A. M. Idrees and E. Shaaban, "Reforming home energy consumption behavior based on mining techniques: A collaborative home appliances approach," *Kuwait J. Sci.*, vol. 47, no. 4, pp. 29–38, 2020.

[17] D. H. A. Hassouna, A. E. Khedr, A. M. Idrees, and A. I. El Seddawy, "Intelligent personalized system for enhancing the quality of learning," *J. Theor. Appl. Inf. Technol.*, vol. 98, no. 13, pp. 2199–2213, 2020.

[18] A. Idrees and W. Gomaa, "A proposed method for minimizing mining tasks' data dimensionality," *Int. J. Intell. Eng. Syst.*, vol. 13, no. 2, pp. 182–195, Apr. 2020.

[19] A. M. Idrees and M. H. Ibrahim, "A proposed framework targeting the enhancement of students' performance in Fayoum university," *Int. J. Sci. Eng. Res.*, vol. 9, no. 11, pp. 1–7, 2018.

[20] H. A. Hassan, M. Y. Dahab, K. Bahnassy, A. M. Idrees and F. Gamal, "Arabic documents classification method a step towards efficient documents summarization," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 3, no. 1, pp. 351–359, 2015.

[21] M. Marvi, A. Arfeen, and R. Uddin, "A generalized machine learning-based model for the detection of DDOS attacks," *Int. J. Netw. Managmgment*, vol. 31, no. 6, pp. 1–22, 2020.

[22] R. K. Batchu and H. Seetha, "A generalized machine learning model for DDoS attacks detection using hybrid feature selection and hyperparameter tuning," *Comput. Netw.*, vol. 200, Dec. 2021, Art. no. 108498.

[23] R. K. Batchu and H. Seetha, "An integrated approach explaining the detection of distributed denial of service attacks," *Comput. Netw.*, vol. 216, Oct. 2022, Art. no. 109269.

[24] M. Alduailij, Q. W. Khan, M. Tahir, M. Sardaraz, M. Alduailij, and F. Malik, "Machine-learning-based DDoS attack detection using mutual information and random forest feature importance method," *Symmetry*, vol. 14, no. 6, p. 1095, May 2022.

[25] K. B. Dasari and N. Devarakonda, "Detection of DDoS attacks using machine learning classification algorithms," *Int. J. Comput. Netw. Inf. Secur.*, vol. 14, no. 6, pp. 89–97, Dec. 2022.

[26] A. Maslan, K. M. B. Mohamad, and F. B. M. Foozy, "Feature selection for DDoS detection using classification machine learning techniques," *IAES Int. J. Artif. Intell.*, vol. 9, no. 1, p. 137, Mar. 2020.

[27] Y. I. Kurniawan, F. Razi, N. Nofiyati, B. Wijayanto, and M. L. Hidayat, "Naive Bayes modification for intrusion detection system classification with zero probability," *Bull. Electr. Eng. Informat.*, vol. 10, no. 5, pp. 2751–2758, Oct. 2021.

[28] C. M. NalayiniI and. D. J. Katiravan, "Detection of DDoS attack using machine learning algorithms," *J. Eng. Technol. Innov. Res.*, vol. 9, no. 7, pp. 223–232, Jul. 2022.

[29] S. Naiem, A. E. Khedr, A. M. Idrees, and M. Marie, "Iterative feature selection-based DDoS attack prevention approach in cloud," *Int. J. Electr. Comput. Eng. Syst.*, vol. 14, no. 2, pp. 197–205, Feb. 2023.

[30] L. Shu, H. Zhang, Y. You, Y. Cui, and W. Chen, "Towards fire prediction accuracy enhancements by leveraging an improved Naïve Bayes algorithm," *Symmetry*, vol. 13, no. 4, p. 530, Mar. 2021.

[31] A. R. Safitri and M. A. Muslim, "Improved accuracy of naive Bayes classifier for determination of customer churn uses SMOTE and genetic algorithms," *J. Soft Comput. Explor.*, vol. 1, no. 1, pp. 70–75, 2020.

[32] F. Gharibian and A. A. Ghorbani, "Comparative study of supervised machine learning techniques," in *Proc. 5th Annu. Conf. Commun. Netw. Services Res. (CNSR)*, Fredericton, NB, Canada, May 2007, pp. 350–358.
[33] E. K. Ampomah, G. Nyame, Z. Qin, P. C. Addo, E. O. Gyamfi, and M. Gyan, "Stock market prediction with Gaussian Naïve Bayes machine learning algorithm," *Informatica*, vol. 45, no. 2, pp. 243–256, Jun. 2021.
[34] Canadian Institute for Cybersecurity. (2018). *CSE-CIC-IDS2018 on AWS, A Collaborative Project Between the Communications Security Establishment (CSE) & the Canadian Institute for Cybersecurity (CIC)*. [Online]. Available: http://www.unb.ca/cic/datasets/ids-2018.html
[35] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proc. 4th Int. Conf. Inf. Syst. Secur. Privacy (ICISSP)*, Portugal, 2018, pp. 108–116.
[36] *A Realistic Cyber Defense Dataset (CSE-CIC-IDS2018)*. Accessed: May 1, 2023. [Online]. Available: https://registry.opendata.aws/cse-cic-ids2018/
[37] Z. Wu, "Using machine learning approach to evaluate the excessive financialization risks of trading enterprises," *Comput. Econ.*, vol. 59, no. 4, pp. 1607–1625, Jan. 2021.
[38] S. C. Gupta and V. K. Kapoor, *Fundsments of Mathematical Statistics*, 12th ed. New Delhi, India: Sultan Chad & Sons, 2020.
[39] D. Kshirsagar and S. Kumar, "An ensemble feature reduction method for web-attack detection," *J. Discrete Math. Sci. Cryptogr.*, vol. 23, no. 1, pp. 283–291, Jan. 2020.
[40] D. Kshirsagar and S. Kumar, "An efficient feature reduction method for the detection of DoS attack," *ICT Exp.*, vol. 7, no. 3, pp. 371–375, Sep. 2021.
[41] H. Liu, M. Zhou, and Q. Liu, "An embedded feature selection method for imbalanced data classification," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 3, pp. 703–715, May 2019.
[42] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, and J. Tang, "Feature selection: A data perspective," *ACM Comput. Surv.*, vol. 50, no. 6, pp. 1–45, 2017.
[43] M. Tan, A. Iacovazzi, N. M. Cheung, and Y. Elovici, "A neural attention model for real-time network intrusion detection," in *Proc. IEEE 44th Conf. Local Comput. Netw. (LCN)*, Oct. 2019, pp. 291–299.
[44] E. S. Alghoson and O. Abbass, "Detecting distributed denial of service attacks using machine learning models," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 12, pp. 616–622, 2021.
[45] G. K. Baydoğmuş, "The effects of normalization and standardization an Internet of Things attack detection," *Eur. J. Sci. Technol.*, no. 29, pp. 187–192, Dec. 2021.
[46] K. Gozde, "The effect of normalization and standardization an Internet of Things attach detection," *Eur. J. Sci. Technol.*, vol. 29, pp. pp. 187–192, Dec. 2021.
[47] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[48] A. E. Khedr, A. M. Idrees, and R. Salem, "Enhancing the e-learning system based on a novel tasks' classification load-balancing algorithm," *PeerJ Comput. Sci.*, vol. 7, p. e669, Sep. 2021.

**SARAH NAIEM** is currently pursuing the master's degree with Helwan University. She is also an Assistant Lecturer with Helwan University. Her research interests include knowledge discovery, text mining, opinion mining, cloud computing, E-learning, software engineering, data science, and data warehousing.

**AYMAN E. KHEDR** is currently a Professor with the University of Jeddah. He has been the Vice Dean of Post-Graduation and Research and the Head of the Information Systems Department, Faculty of Computers and Information Technology, Future University in Egypt. He is also a Professor with the Faculty of Computers and Information, Helwan University, Egypt. Previously, he was the General Manager of the Helwan E-Learning Center. His research interests include the themes (scientific) data and model management, data science, big data, the IoT, E-learning, data mining, bioinformatics, and cloud computing.

**AMIRA M. IDREES** is currently a Professor in information systems with the Faculty of Computers and Information Technology, Future University, the Vice Dean for Community Services and Environmental Development, the Head of the IS Department, and the Head of the University Requirements Unit. She has been the Head of Scientific Departments and the Vice Dean of the Community Services and Environmental Development, Faculty of Computers and Information, Fayoum University. Her research interests include knowledge discovery, text mining, opinion mining, cloud computing, E-learning, software engineering, data science, and data warehousing.

**MOHAMED I. MARIE** received the bachelor's degree from Cairo University, in 1994, and the master's and Ph.D. degrees from Helwan University, Cairo, Egypt, in 1999 and 2004, respectively. He is currently an Associate Professor with the Information Systems Department, Faculty of Computers and Artificial Intelligence, Helwan University. He held a lot of positions, such as the Head of the Quality Unit of the Faculty, in 2010. He is a Coordinator of Software Engineering Program, in 2020. He was as Assistant Professor in information systems with the Information Systems Department, Faculty of Computer Science and Information Systems, Jazan University, Saudi Arabia, from 2012 to 2019. He was the Head of Information Systems Curriculum Review Committee, the Head of Information Technology Curriculum Review Committee, the Head of Promotion Committee, the Head of Information Systems Department Staff Members Yearly Testing and Evaluation Committee, the CEO of the Information Systems Department, an Information Systems Department Midterm and Final Exams Reviewer, a member of the Consultation Committee, a member of Quality Assurance Unite (Consultant and Reviewer), a member of Information Systems Department Committee, and a member of Graduation Projects Evaluation Committee. He is also a member of Publons, Web of Science, ORCID, Academia, and Google Scholar. He is a reviewer in many journals and received many certificates from them.