**CHAPTER 4**

**PROPOSED METHODOLOGY**

4.1 METHODOLOGY

In this study, we have proposed the work of DDOS Attack detection by using the ensemble based algorithm and machine learning algorithms as Random Forest Regression, K-Nearest Neighbor (K-NN) and Support Vector Machine (SVM). The dataset used here is CIC2019 and it is taken from Kaggle. The optimal feature from dataset is selected using the improved Grey Wolf Optimization algorithm.

4.1.1 Dataset

The DDOS attack dataset was taken from the data repository called Kaggle. This dataset consists of 4000 records of data which were split by applying a 70-30 percent rule to split the training and testing data. The sample data is given below.

* + 1. Dataset description

The data contains the following fields as Source Port, Destination Port, Protocol, Flow Duration, Total Fwd Packets, Total Length of Fwd Packets, Min Packet Length,MaxPacket Length, Average Packet Size, act\_data\_pkt\_fwd, min\_ seg\_ size\_ forward, Attack\_type

Preprocessing prepares the data in such a way that it is ready for the training model. First, delete six socket features which are not influencing the target because they differ from network-to-network values. Then, in order to acquire more accurate results, records with missing or infinite are removed. Some machine learning algorithms [12] work with numerical values, so BENIGN and attack labels are encoded with 0 and 1 binary values respectively. Standardize the data using Standard Scaler to reduce the training time.

The main objective of this research is prediction of categorical values of Benign and DDOS attacks of target labels in the CICDDOS2019 dataset. In this research, machine learning classification algorithms were used to detect DDOS attacks on the CICDDOS2019 dataset. Training and testing are two steps in the classification process. Logistic regression, Decision tree, Random Forest, K-Nearest Neighbor, Naive Bayes, and AdaBoost are some of the most common algorithms in the classification. These methods are significantly more accurate than conventional methods for detecting a DDOS attack, in addition to being faster.

4.2 SUPPORT VECTOR MACHINE

Support Vector Machine (SVM) is one of the supervised learning algorithms which is used for both classification as well as regression. The goal of this SVM algorithm is used to create best fitting line into the model called hyperplane.

It chooses the vectors that are used to creating the hyperplane. The data points that are the closest to the hyperplane and affect the position is known as support

vector.

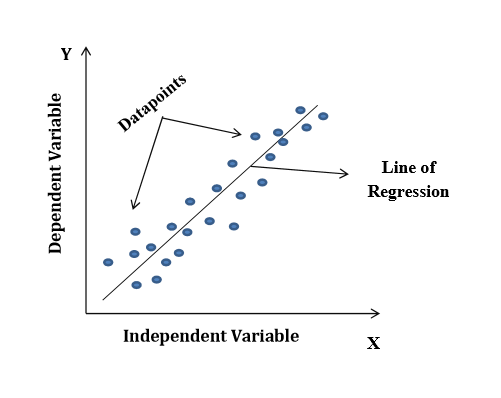


Figure 4.1 Support Vector Machine

The equation for support vector machine

*Y=a+bX (4.1)*

The above equation (4.1) shows that the dependent variable depends on independent variables or predictors. Regression is essentially finding a relationship or association between the dependent variable (Y) and the independent variable (X).



Figure 4.2 Pseudo Code for SVM Algorithms

First, reading and understanding the data. Second, visualizing the data (Exploratory Data Analysis.Third, doing the data preparation, Fourth, splitting the data into training and test sets, Fifth, building a support vector model, Residual analysis of the train data. Finally, making predictions using the final model and evaluation.

4.3 KNN ALGORITHM

KNN is a classiﬁcation approach that classiﬁes test data observations which is based on how close they are to nearest class neighbors. It is used as a semi-supervised learning approach, and KNN is used to identify the nearest neighbors. It is based on 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.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.

*Y=a+b1X1+b2X2 (4.2)*

The above equation (4.2) express that the parameters that are need to obtained from the training data and variables were extracted from the dataset. The model describes a plane in the three dimensional space of Y, X1, X2, Parameter ‘a’ is the intercept of this plane. Parameters b1 and b2 are referred as partial regression coefficients.



Figure 4.3 Pseudo Code for KNN Algorithms

4.4 RANDOM FOREST CLASSIFIER

Random forest is a collection of decision trees trained on different dataset subsets and then averaged to increase predictive accuracy. It is created randomly with a collection of decision trees.

Each node selects a set of features at random to calculate the outcome. The output of individual decision trees is combined in the random forest to produce the outcome.



Figure 4.4 Pseudo Code for Random Forest Regression

4.5 ENSEMBLE METHOD

Ensemble methods are techniques that aim at improving the accuracy of results in models by combining multiple models instead of using a single model. The combined models increase the accuracy of the results significantly. This has boosted the popularity of ensemble methods in machine learning.

The most popular ensemble methods are boosting, bagging and stacking. Ensemble methods are ideal for regression and classification, where they reduce the bias and variance to boost the accuracy of the models.

4.5.1 BAGGING METHOD

Bagging is used to increase the accuracy of models through decision trees. The reduction of variance increases accuracy, eliminating overfitting, which is a challenge to many predictive models.

Bagging is classified into two types, bootstrapping and aggregation. Bootstrapping is a sampling technique where samples are derived from the whole population using the replacement procedure. The sampling with replacement method helps make the selection procedure randomized. The base learning algorithm is run on the samples to complete the procedure.

Aggregation in bagging is done to incorporate all possible outcomes of the prediction and randomize the outcome. Without aggregation, predici\tions will not be accurate because all outcomes are not put into consideration. Therefore, the aggregation is based on the probability bootstrapping procedures or on the basis of all outcomes of the predictive models.

4.6 GREY WOLF OPTIMIZATION

Grey wolf optimizer (GWO) is a population-based meta-heuristics algorithm that simulates the leadership hierarchy and hunting mechanism of grey wolves in nature, and it’s proposed by Seyedali Mirjalili et al. in 2014.

* Grey wolves are considered apex predators, which are at the top of the food chain
* Grey wolves prefer to live in groups (packs), each group containing 5-12 individuals on average.
* All the individuals in the group have a very strict social dominance hierarchy.
* Alpha α wolf is considered the dominant wolf in the pack and his/her orders should be followed by the pack members.
* Beta *β* are subordinate wolves, which help the alpha in decision-making and are considered as the best candidate to be the alpha.
* Delta *δ*wolves have to submit to the alpha and beta, but they dominate the omega. There are different categories of delta-like Scouts, Sentinels, Elders, Hunters, Caretakers etc.
* Omega *ω* wolves are considered as the scapegoat in the pack, are the least important individuals in the pack and are only allowed to eat at last.



4.6 PERFORMANCE METRICS

To evaluate the performance of the model or the quality of the model, different metrics can be used for the comparison between the actual value and the predicted value, to determine the performance of this forecasting model. The parameters are Confusion Matrix, Accuracy, Precision, Recall. These metrics are known as performance metrics.

There are four important terms used in evaluation metrics.

True Positives (TP):

In this case, both the predicted and actual values are Positive.

True Negatives (TN):

In this case, both the predicted, and actual values are Negative.

False Positives (FP):

In this case, the actual value is Negative but the predicted value is Positive.

False Negatives (FN):

In this case, the actual value is Positive but the predicted value is Negative.

4.6.1 CONFUSION MATRIX

The confusion matrix is a key concept in machine learning classification performance. It represents actual and predicted values in tabular form. Predicted and actual values are represented by rows and columns respectively in the table.

4.6.2 ACCURACY

Accuracy is the ratio of the number of correct predictions to the number of all predictions by the classifier. Accuracy tells the proposition of correct predictions out of total predictions.

The below formula 4.3 is used to calculate the Accuracy

Accuracy = (4.3)

4.6.3 PRECISION

Precision is the ratio between the number of True Positives and the number of predicted positives by the classifier. Precision tells the proposition of predicted true are actually. The below formula 4.4 is used to calculate the Precision

Precision = (4.4)

4.6.4 RECALL

Recall or True Positive Rate (TPR) is the ratio between the number of True Positives and the number of all relevant samples. Recall tells the proposition of actually trues are predicted as true. The below formula 4.5 is used to calculate the Recall

Recall = (4.5)

4.6.5 F1 SCORE

F1 score is a harmonic mean of precision and recall. The below formula 4.6 is used to calculate the F1 Score

F1 Score = (4.6)

4.6.6 SPECIFICITY

Specificity is the ratio between the number of True Negatives and the number of all relevant samples. It is also called True Negative Rate (TNR). The below formula 4.7 is used to calculate the Specificity.

Specificity = (4.7)



Figure 4.5 System Flow Model