International Journal of Engineering Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 10 Issue 8 Series I || August 2021 || PP 48-53

Impact of Feature Selection on Genetic Programming Classifier for Anomaly based Network Intrusion Detection

Ashalata Panigrahi
Roland Institute of Technology, Berhampur, India

ABSTRACT
The objective of intrusion detection system is to detect malicious activities from the network traffic. Although many techniques have been proposed to increase the efficacy of IDS but it is still a problem for existing intrusion detection algorithm to achieve good performance due to many irrelevant features are present in the high dimensional dataset. Irrelevant features may even reduce the performance of the classification algorithm. The objective of feature selection is to select small number of relevant features to achieve better performance than using all features. The objective of this paper is to build a robust and accurate IDS using Genetic Programming classifier. Further six filter based feature selection methods namely, Information Gain, Gain Ratio, Symmetrical Uncertainty, Relief-F, Chi-Squared Attribute Evaluator, One-R and six search based feature selection methods namely Greedy Stepwise Search , Best First Search, Ant Search , Particle Swarm Optimization Search, Genetic Search , and Rank Search have been employed on the dataset to select most relevant features before classification. The performance of the model has been evaluated using ten metrics including Kappa Coefficient, Matthews Correlation Coefficient, Positive Likelihood Ratio.

KEYWORDS: *Anomaly detection, Ant search, Feature selection, Genetic Programming, Geometric Mean*

Date of Submission: 06-08-2021 Date of Acceptance: 20-08-2021

I. INTRODUCTION

Intrusion detection system (IDS) monitors activity to identify malicious or suspicious events. An IDS is a sensor, like a smoke detector, that raises an alarm if specific things occur [1]. Two general types of IDS are signature-based and heuristic or anomaly based. Signature-based IDS perform simple pattern matching and report situations that match a pattern corresponding to a known attack type. All heuristic intrusion detection activity is classified in one of three categories: good or bengin, suspicious, or unknown [2]. Different techniques have been proposed to build intrusion detection systems but the challenge lies in dealing with issues like huge volume of network traffic, identifying boundaries between normal behavior and attacks, imbalanced data distribution, and need for continuous adaptation to a constantly changing network environment.

Patgiri et al. [3] developed a model that using Random Forest and Support Vector Machine. Recursive feature Elimination is used as a feature selection method with SVM. NSL KDD dataset is used for training and evaluation. They have used Cross-validation to evaluate it. Using their model, Random Forest performed better than SVM before feature selection. After feature selection, SVM performed better than RF. Biswas [4] proposed an IDS model that compares performance of different combinations. A subset of significant features is selected using feature selection algorithms and then to train the model using different classifiers. NSL-KDD dataset is used to evaluate the model using 5-fold cross validation. CFS, IGR, PCA, and minimum redundancy maximum-relevance feature selection techniques, and SVM, DT, k-NN, NB, and NN classifiers are used in their model. Through their observations it is proved that the performance of K-NN is superior to other classifiers and performance of IGR is superior to other feature selection methods. Almseidin et al. [5] were performed several experiments to estimate the performance of the ML classifiers namely, Random Tree, J48, MLP, Random Forest, Naive Bayes, Decision Table, and Bayes Network. They have used KDD intrusion detection dataset for testing. Among these RF classifier obtained an accuracy of 93.77% - highest accuracy than others and with the smallest false positive rate and RMSE value. Salih et al. [6] proposed a model with three classifiers: Naive Bayes, Multilayer Perceptron and K-Nearest Neighbors. For Feature selection three methods are used: Gain Ratio, Information Gain, Correlation. The proposed model is analyzed using KDD-CUP 99 data set. The KNN got the highest detection rate than others. Ravale [7] proposed hybrid technique (KMSVM) that combines RBF kernel function of SVM and K-Means clustering for classification. The proposed technique is evaluated using KDD-CUP99 data Set. K-Means clustering is used for feature reduction. The experimental results showed that the performance of the proposed model is superior to others. Accuracy of the proposed method is 92.86%. Mazinia et al. [8] proposed a method for an anomaly based IDS (A-NIDS) using fusion of artificial bee colony

(ABC) and AdaBoost algorithms. ABC algorithm is used for feature selection and AdaBoost is used to evaluate and classify the features. The proposed method is tested on NSL-KDD.

The rest of the paper is structured as follows: Section II presents description of Genetic Programming algorithm. Section III presents the proposed model. Section IV presents dataset description, different feature selection methods and the performance metrics considered in the experimental works of the study. Section V highlights experimental results and analysis. Finally the conclusions is given in Section VI.

II. METHODOLOGY

Genetic Programming

Genetic programming (GP) is an extension of genetic algorithm [9]. GP offers solutions in representations of computer programs. This offers the flexibility to (i) perform operations in a hierarchical way, (ii) perform alternative computations conditioned on the outcome of intermediate calculations, (iii) perform iterations and recursions, (iv) perform computations on variables of different types, (v) define intermediate values and sub-programs so that they can be subsequently reused [10].

GP is a search method that uses analogies from natural selection and evolution. GP encodes multi-potential solutions for specific problems as a population of programs or functions. The programs can be represented as parse trees. The parse trees are composed of internal nodes and leaf nodes. Internal nodes are called primitive functions and leaf nodes are called terminals. The terminals can be viewed as the inputs to the specific problem. The primitive functions are combined with the terminals to form more complex function calls[11].

The GP algorithm can be summed up in the following steps:

Step 1: Create a random population of programs, or rules, using the symbolic expressions provided as the initial population.

Step 2: Evaluate each program or rule by assigning a fitness value according to a predefinedfitness function that can measure the capability of the rule or program to solve the problem.

Step 3: Use reproduction operator to copy existing programs into the new generation.

Step 4: Generate the new population with crossover, mutation, or other operators from a randomly chosen set of parents.

Step 5: Repeat steps 2 onwards for the new population until a predefined termination criterion has been satisfied, or a fixed number of generations have been completed.

Step 6: The solution to the problem is the genetic program with the best fitness within all the generations.

The main operators used in genetic programming are crossover and mutation. In GP, crossover operation is achieved firstly by reproduction of two parent trees; two crossover points are then randomly selected in the two offspring trees. Exchanging sub-trees, which are selected according to the crossover point in the parent trees, generates the final offspring trees. The offspring trees so obtained are usually different from their parents in size and shape. Mutation is also considered in GP which affects an individual in the population. A single parental tree is firstly reproduced. Then a mutation point is randomly selected from the reproduction, which can be either a leaf node or a sub-tree. Finally, the leaf node or the sub-tree is replaced by a new leaf node or sub-tree generated randomly.

The important concept of GP is the fitness function. Fitness functions ensure that the evolution is toward optimization by calculating the fitness value for each individual in the population. The fitness value evaluates the performance of each individual in the population.

III. PROPOSED MODEL OF INTRUSION DETECTION SYSTEM USING GENETIC PROGRAMMING CLASSIFIER

Step 1: Load NSL-KDD dataset with all features.

Step 2: Applying different feature selection methods on the dataset for finding the potential features.

Step 3: Adopting Genetic Programming classifier for classifying the dataset into two classes namely attack and normal.

Step 4: 10-fold cross-validation approach is used to validate the performance of the model.

Step 5: Evaluate the model by comparing the performance of different metrics including Error Rate (ER), Kappa Coefficient (KC), Matthews Correlation Coefficient (MCC).

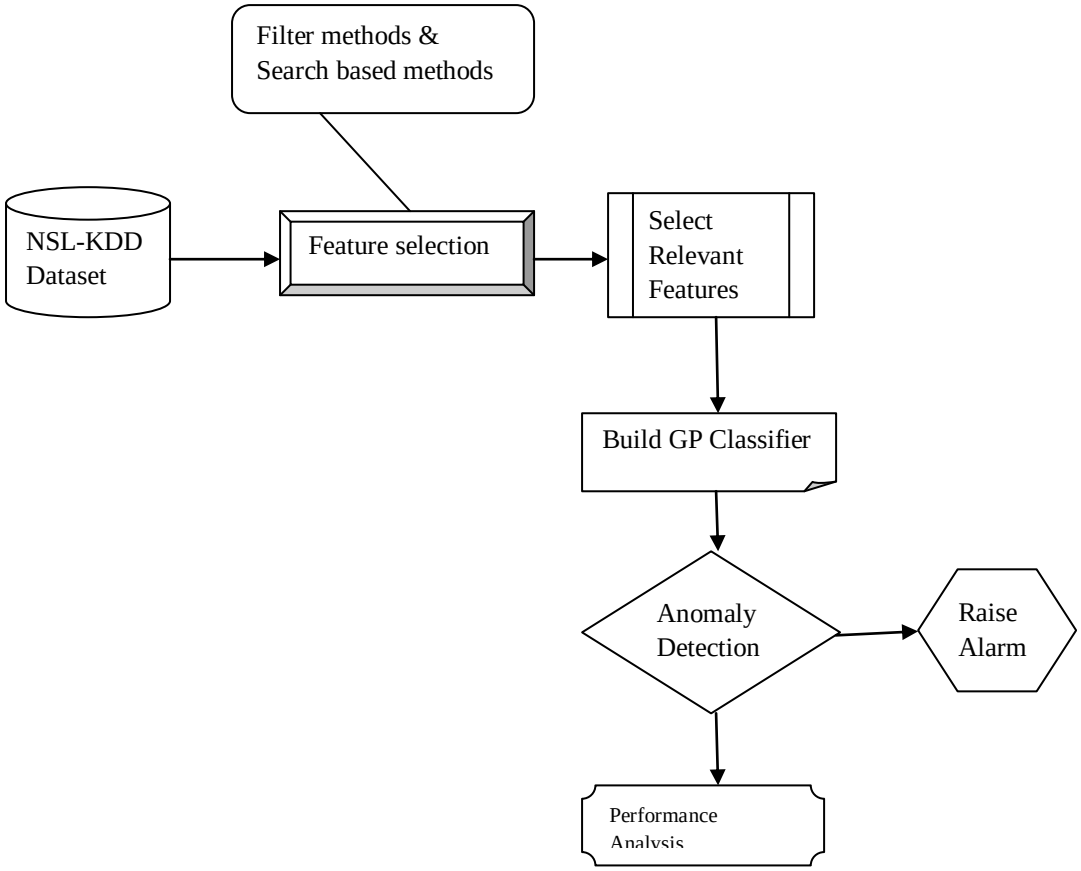

Fig. 1 Proposed Model

IV. EXPERIMENTAL SETUP

4.1 **Dataset Description**

In this paper NSL-KDD dataset is used which is based on KDDCUP 99. NSL-KDD dataset has 41 features for each connection record and one class label. The NSL-KDD dataset remove redundant and duplicate records of KDD CUP 99 dataset and solve the inherent problems of KDD CUP 99 [12]. Therefore it contains reasonable number of instances and the experiment can be implemented on the whole dataset. The dataset contains twenty four different types of attack. All attacks fall into one of the following four categories: Denial of Service (DoS), User to Root (U2R), Probing, Remote to Local (R2L). DoS is an attack that tries to shut down traffic flow to and from the target system. U2R is an attack that starts off with a normal user account and tries to gain access to the system or network, as a super-user (root). Probe or surveillance is an attack that tries to get information from a network. R2L is an attack that tries to gain local access to a remote machine.

4.2 Feature Selection

The main objective of this paper is to experimentally verify the impact of different feature selection techniques on Genetic Programming classification algorithm. Feature selection is helpful to eliminate data redundancy and decrease the computational time and complexity. The feature selection process consists of four basic steps, viz., subset generation, subset evaluation, stopping criterion, and result validation [13]. Feature selection algorithms may be divided into filters [14,], wrappers [15], and embedded approaches [16]. Filter methods evaluate quality of selected features, independent from classification algorithms. Wrapper method require application of a classifier to evaluate the quality. Embedded methods perform feature selection during learning of optimal parameters. In this work six filter methods and six search based feature selection methods are applied on the NSL-KDD dataset to select relevant features. The filter methods are: Information Gain (IG), Gain Ratio (GR), Symmetrical Uncertainty (SU), Relief-F (RF), Chi-Squared (CS), and One-R (OR) and search based feature selection methods are namely, Greedy Stepwise Search (GSS), Best First Search (BFS), Ant Search (AS), Particle Swarm Optimization (PSO) Search, Genetic Search (GS), and Rank Search (RS).

Information Gain attribute evaluation calculates the information gained by the attributes with respect to the classification target. This algorithm sets a threshold value and attributes that are above the threshold will be considered for further processing [17]. Gain Ratio Attribute evaluator evaluates the worth of an attribute by

measuring the gain ratio with respect to the class. Symmetrical Uncertainty ranks attributes by their individual evaluations. Relief-F is an enhancement of the original Relief method. This algorithm randomly selects an instance and its value and compares it with the nearest neighbors to find a relevance score for each attribute. The algorithm tries to create a list of attributes that can differentiate between instances from the class labels [18]. Chi-Squared attribute evaluator evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class. One-R algorithm builds one rule for each attribute in the training data and then selects the rule with the smallest error. It treats all numerically valued features as continuous and uses a method to divide the range of values into several disjoint intervals. It handles missing values by treating "missing" as a legitimate value.

Search Methods

Search methods search the set of all possible features in order to find the best set of features. Greedy Stepwise search [19] performs a greedy forward or backward search through the space of attribute subsets. It may start with no / all attributes or from an arbitrary point in the space and stops when addition/ deletion of any attribute results in decrease in evaluation. This can also produce a ranked list of attributes by traversing the space from one side to other and recording the order that attributes are selected. Best First Search (BFS) [19] uses classifier evaluation model to estimate the merits of attributes. The attributes with high merit values are considered as potential attributes and thus selected for classification. Best first moves through the search space by making local changes to the current feature subset. It searches the space of attribute subsets by augmenting with a backtracking facility. Ant search performs a search using ant colony optimization. For each generation, ants starts off at a random feature and move probabilistically until there is no improvement in their constructed subset quality. The smallest subset found overall with maximum quality is returned. Genetic search performs a search using the simple genetic algorithm. The Rank search method uses an attribute subset evaluator to rank all attributes of the dataset. If a subset evaluator is specified then a forward selection search is used to generate rank of the features. From the ranked list of attributes , subsets of increasing size are evaluated. Finally the best feature set is selected.

4.3 **IDS Evaluation Method**

There are different metrics are used to evaluate the performance of the model. In this work confusion matrix is used to evaluate the performance of the model. Confusion matrix is a tabular representation of true positives (T_P) , true negatives (T_N), false positives (F_P), and false negatives (F_N) (Lippmann et al., 2000) as shown in Table 1.

Table 1: Confusion Matrix

		Predicted Class	
		Normal	Attack
Actual Class	Normal	T_N	F_P
	Attack	F_N	T_P

T_P : The number of actual attack records are classified as attack.
T_N : The number of actual legitimate records are identified as normal.
F_P : The number of actual legitimate records are identified as attacks.
F_N : The number of actual attack records are detected as normal.
Evaluate the performance of the model in terms of True Negative Rate (TNR), Negative Predictive Value (NPV), False Negative Rate (FNR), Error Rate, False Discovery Rate (FDR), Kappa Coefficient (KC), Matthews Correlation Coefficient (MCC), Youden's Index (YI), Geometric Mean (GM), and Positive Likelihood Ratio (PLR).

Specificity or True Negative Rate (TNR) = $T_N / (T_N + F_P)$ …….. (1)
Negative Predictive Value (NPV) = $T_N / (T_N + F_N)$ …… (2)
False Negative Rate (FNR) = $F_N / (F_N + T_P)$ …….. (3)
Error Rate (ER) = $(F_P + F_N) / (T_P + T_N + F_P + F_N)$ …… (4)
False Discovery Rate (FDR) = $F_P / (F_P + T_P)$ …….. (5)
Kappa Coefficient (KC) or Kappa = (Total Accuracy − Random Accuracy) / (1 − Random Accuracy) …(6)
Where Total Accuracy = $(T_P + T_N) / (T_P + T_N + F_P + F_N)$
Random Accuracy = $[(T_N + F_P)(T_N + F_N) + (F_N + T_P)(F_P + T_P)] / (T_P + T_N + F_P + F_N)^2$
Matthews Correlation Coefficient (MCC)
$= [(T_P \times T_N) – (F_P \times F_N)] / \sqrt{[(T_P + F_P) \times (T_P + F_N) \times (T_N + F_P) \times (T_N + F_N)]}$ ……..(7)
Youden's Index (YI) = $[T_P / (T_P + F_N)] + [T_N / (T_N + F_P)] – 1$ ……..(8)
Geometric Mean (GM) = $\sqrt{[T_P / (T_P + F_N)] \times [T_N / (T_N + F_P)]}$ ……(9)
Positive Likelihood Ratio (PLR) = $[T_P / (T_P + F_N)] / [F_P / (F_P + T_N)]$ (10)

V. RESULT ANALYSIS

In this work apply two categories of feature selection methods namely filter based and search based feature selection methods on NSL-KDD dataset to select the most relevant features. The performance of the model has been evaluated using ten metrics namely, True Negative Rate (TNR), Negative Predictive Value (NPV), False Negative Rate (FNR), Error Rate (ER), False Discovery Rate (FDR), Kappa Coefficient (KC), Matthews Correlation Coefficient (MCC), Youden's Index (YI), Geometric Mean (GM), and Positive Likelihood Ratio (PLR). In the experiment 10-fold cross-validation has been applied for evaluation of the proposed model. The results are presented in Table 2,3,4, and 5.. Table 2 and 3 presents the performance score of TNR, NPV, FNR, ER, and FDR . Table 4 and 5 presents the performance score of KC, MCC, YI, GM and PLR.

Table 2 Comparison of, TNR, NPV, FNR, ER, and FDR of GP Classifier using Filter based Feature Selection Method

Feature Selection Method	Classifier Technique	Evaluation Metric				
		TNR	NPV	FNR	ER	FDR
Information Gain	Genetic Programming	0.9541	0.8529	0.189	0.1125	0.0611
Gain Ratio		0.9551	0.8357	0.2157	0.1244	0.0617
Symmetrical Uncertainty		0.9556	0.8626	0.1749	0.1051	0.0582
Relief-F		**0.9655**	0.8698	0.1659	**0.0957**	**0.0453**
Chi-Squared Attribute Evaluator		0.9563	**0.8737**	**0.1587**	0.0972	0.0563
One-R		0.9625	0.8528	0.1909	0.1088	0.0505

Table 3 Comparison of, TNR, NPV, FNR, ER, and FDR of GP Classifier using Search based Feature Selection Method

Feature Selection Method	Classifier Technique	Evaluation Metric				
		TNR	NPV	FNR	ER	FDR
Greedy Stepwise Search	Genetic Programming	0.9409	0.8463	0.1962	0.1229	0.0779
Best First Search		0.9361	0.8869	**0.1371**	**0.098**	0.0784
Ant Search		0.9466	0.8272	0.1933	0.1936	0.0608
PSO Search		0.9581	**0.8543**	0.1876	0.1097	0.0559
Genetic Search		**0.9585**	0.8531	0.1896	0.1104	**0.0555**
Rank Search		0.9537	0.8448	0.2012	0.1183	0.0624

High TNR value indicates the proposed model perfectly classified normal records. Here Genetic Programming technique with Relief-F feature selection method gives highest TNR value of 0.9655. NPV value presents the performance of the prediction. Chi-squared attribute evaluator with GP gives highest NPV value of 0.8737. Low FNR indicates high detection rate and the model is perfectly detected attacks. Best first search with GP technique gives lowest FNR value of 0.1371. Very low error rate is important in intrusion detection system, it indicates the model is better. Here GP with Relief-F feature selection gives lowest error rate of 0.0957. Low FDR value indicates good classification performance. Relief-F feature selection with GP technique gives lowest FDR value of 0.0453.

Table 4 Comparison of KC, MCC, YI, GM and PLR of GP Classifier using Filter based Feature Selection Method

Feature Selection Method	Classifier Technique	Evaluation Metric				
		KC	MCC	YI	GM	PLR
Information Gain	Genetic Programming	0.7718	0.7783	0.765	0.8796	17.6562
Gain Ratio		0.7473	0.7565	0.7394	0.8655	17.4523
Symmetrical Uncertainty		0.787	0.7925	0.7807	0.888	18.5901
Relief-F		**0.8061**	**0.8119**	**0.7996**	**0.8974**	**24.1757**
Chi-Squared Attribute Evaluator		0.8032	0.8074	0.7975	0.8969	19.2367
One-R		0.7791	0.7868	0.7717	0.8825	21.6343

Table 5 Comparison of KC, MCC, YI, GM and PLR of GP Classifier using Search based Feature Selection Method

Feature Selection Method	Classifier Technique	Evaluation Metric				
		KC	MCC	YI	GM	PLR
Greedy Stepwise Search	Genetic Programming	0.7509	0.7565	0.7447	0.8696	13.6031
Best First Search		**0.8023**	**0.8037**	**0.799**	**0.8987**	13.5042
Ant Search		0.6643	0.7599	0.7533	0.8738	15.1201
PSO Search		0.7775	0.7843	0.7705	0.8822	19.3859

Genetic Search		0.7761	0.7832	0.769	0.8814	19.5286
Rank Search		0.7598	0.7674	0.7526	0.8728	**20.2706**

Kappa Coefficient value compares the accuracy of the system to the accuracy of a random system. The coefficient value ranges from 0 to 1. GP classification technique with Relief-F feature selection gives highest value of 0.8061. High score of MCC indicates the classifier able to perfectly predict positive data records and negative data records. The value of MCC ranges from -1 to $+1$. Here GP technique with Relief-F feature selection method gives highest MCC value of 0.8119. Youden's Index (YI) evaluate the algorithms ability to avoid failure. The value of YI ranges from 0 to 1. A high value of YI indicates the classifier performance is good. Here GP technique with Relief-F feature selection method gives highest YI value of 0.7996. Geometric Mean (GM) value is based on accuracy of positive records and accuracy of negative records. Best first search feature selection with GP technique gives highest score of 0.8987. High Positive likelihood Ratio (PLR) value indicates better performance on positive classes. The threshold value of PLR is greater than 10 indicates good. All the feature selection methods combine with GP algorithm gives the value of PLR is more than 13 and Relief-F feature selection with GP technique gives highest value of 24.1757. These results suggest that Relief-F feature selection with GP classifier performs better as compared to other feature selection methods.

VI. CONCLUSIONS

In this paper proposed anomaly based network intrusion detection model based on genetic programming classifier and two categories of feature selection methods. The performance of the model was analyzed along different evaluation criteria on the NSL-KDD dataset. It was observed that Relief-F feature selection gives better result as compared to other feature selection methods.

REFERENCES

[1]. R.Kemmerer and G.Vigna, Intrusion detection: A brief history of overview, IEEE security and Privacy, 1(1), 2002, 27-30.
[2]. C.P.Pfleeger, S.L. Pfleeger, Security in Computing, 4th Edition, Pearson education, 2008.
[3]. R. Patgiri, U. Varshney, T. Akutota, and R. Kunde, An investigation on intrusion detection system using machine learning, IEEE Symposium Series on Computational Intelligence ISSCI 2018.
[4]. S.K.. Biswas , Intrusion detection using machine learning: A comparison study, *International Journal of Pure and Applied Mathematics* , 118(9), 2018, 101-114.
[5]. M. Almseidin, M. Alzubi, S. Kovacs and M. Alkasassbeh, Evaluation of machine learning algorithms for intrusion detection system, SISY 2017 IEEE
[6]. A. A. Salih, M. B. Abdulrazaq, Combining best features selection using three classifiers in intrusion detection system, *International Conference on Advanced Science and Engineering (ICOASE)* 2019, IEEE 2019 .
[7]. U. Ravale, N. Marathe, and P. Padiya, Feature selection based hybrid anomaly intrusion detection system using K- Means and RBF Kernel Function" ICACTA- 2015.
[8]. M. Mazinia , B. Shirazi, I. Mahdavi, Anomaly network-based intrusion detection system using a reliable hybrid artificial bee colony and AdaBoost algorithms, *Journal of King Saud University- Computer and Information Sciences*, 31(4), 2018, 541-553.
[9]. J.R. Koja, (1992), Genetic programming; on the programming of computers by means of Natural Selection, MIT Press, USA, ISBN:10:0262111705, 1992.
[10]. R. Mohanty, V. Ravi, and M.R. Patra, Hybrid intelligent systems for predicting software reliability. *Applied Soft Computing*. 13, 2013, 189-200.
[11]. K.M. Faraoun and A. Boukelif, Genetic programming approach for Multi-Category pattern classification applied to network intrusion detection. World Academy of Science, Engineering and Technology. *International Journal of Computer , Information, systems and Control Engineering*, 1(10), 2007
[12]. Tavallaee, M., Bagheri, E., Lu, W., Ghorbani, A.A., 2009. A detailed analysis of the kddcup99 dataset, in:2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, IEEE.pp.1–6.doi:10.1109/CISDA.2009.5356528
[13]. Dash, M., and Liu, H.(1997), "Feature selection methods for classifications", *Intelligent Data Analysis: An International Journal*, 1 (3) 1997.
[14]. H.Almuallim and T.G. Dietterich. Learning with many irrelevant features. In Proc. AAAI-9I, Anaheim, Ca, 1991, 547-552.
[15]. R.Kohavi and G.H.John. Wrappers for feature subset selection. Artificial intelligence, 97, 1997, 273-324.
[16]. A.I.Blum and P.Langley. Selection of relevant features and examples in machine learning. Artificial Intelligence, 97 (1997), 245-271.
[17]. Mark A. Hall and Lloyd A. Smith. Practical feature subset selection for machine learning. Springer; 1998
[18]. I. Kononenko . Estimating attributes: analysis and extensions of RELIEF. European conference on machine learning. Springer 1994, 171-182.
[19]. E. Rich and K. Knight. Artificial Intelligence, Tata McGraw Hill, 1991.

Ashalata Panigrahi, "Impact of Feature Selection on Genetic Programming Classifier for Anomaly based Network Intrusion Detection." *International Journal of Engineering Science Invention (IJESI)*, Vol. 10(08), 2021, PP 48-53. Journal DOI- 10.35629/6734