International Journal of Applied Engineering & Technology

Generating an Optimal Feature Subset for Performance Enrichment of Intrusion Detection System using GINI

Syeda Khubroo Hashmi[1]

[1]*Research Scholar, Dept. of CSE, Sagar Institute of Science & Technology (SISTec), Bhopal, India;*

Dr. Ghanshyam Prasad Dubey[2], Vishal Chourasia[4]

[2]*Associate Professor, Assistant Professor*[4]*, Department of CSE, Sagar Institute of Science and Technology (SISTec), India;*

Anita Yadav[3]

[3]*Assistant Professor, Department of CSE, Tulsiramji Gaikwad-Patil College of Engineering and Technology, Nagpur, India;*

Corresponding Author: Dr. Ghanshyam Prasad Dubey; email: ghanshyam_dubey2@yahoo.com;

Date of Submission: 22nd November 2022 Revised: 12th January 2023 Accepted: 26th January 2023

How to Cite: **Dr. Ghanshyam Prasad Dubey (2023), Generating an Optimal Feature Subset for Performance Enrichment of Intrusion Detection System using GINI, International Journal of Applied Engineering and Technology 5(1), pp. 60-70.**

***Abstract*-- In the creation of a Machine Learning (ML) based Classifier, particularly for Intrusion Detection Systems (IDS), Feature Engineering is critical. It reduces the size of available datasets, Minimize training-time, and reduces large computation while uplifting the model's accuracy and detection correctness. The most popular strategy for lowering the dimensionality of the available dataset is feature selection. The more dimensions in the dataset, the more training time the Machine Learning model will need to process the dataset. Gini Impurity or the Gini Index assesses the possibility of a certain feature being erroneously identified when randomly selected. The minimum value of Gini impurity helps to select the important features. This work suggests an approach for constructing a reduced feature subset to reduce the dimensions of the dataset based on Gini Impurity. The reduced feature set is used to train the various ML based IDS models. The results of this experiment have demonstrated that the proposed feature selection approach not only operates more quickly but also more accurately.**

***Keywords*--Intrusion Detection system, Anomaly Detection, Gini Index, Security attacks, Machine Learning.**

INTRODUCTION

When discussing about the Internet, computers, electronic devices, and networking, viruses, Trojan horses, and other ransomware threats are frequently brought up as security issues. Antivirus, antimalware, firewalls, cryptographic security systems, and tools to prevent unwanted access are just a few of the security systems and tools that are necessary to provide security against these threats.

An intrusion occurs when an element, person, tool, or other entity enters a system, territory, or environment without authorization. For these operations, an intrusion detection system is utilized to provide security. The usage of cryptography is one of the most important aspects of computer network defense. The primary target of an effective intrusion detection system is to find intrusion and use it to stop other several attacks [1].

Confidentiality, Integrity and Availability forms the core of the Security for a System or a Network. Confidentiality ensures that the sensitive Data or Resources must be kept away from unauthorized access. It is also known as Secrecy or Privacy. Integrity ensures that the Data must be kept consistent and correct and any unauthorized user will not be allowed to modify or fabricate it. Availability is achieved using Authentication, Authorization and Access Control of the system or resource. Authentication is defined as the process of validating the Identity of the User and if the User is Valid, he will be allowed to access the System. Authorization ensures that the Authenticated User will be allowed to Data and Resources for which he is authorized or privileged and other Resources must be kept away from his access. Access Control provides a set of Rules and Security Policies that must be followed for ensuring the protection or security of the System and avoiding any Security threat, attack, breach and vulnerabilities [2].

Detection mechanism, target systems (where they are located), analysis time, intrusion response, and architecture; these are the ways of IDS classification.

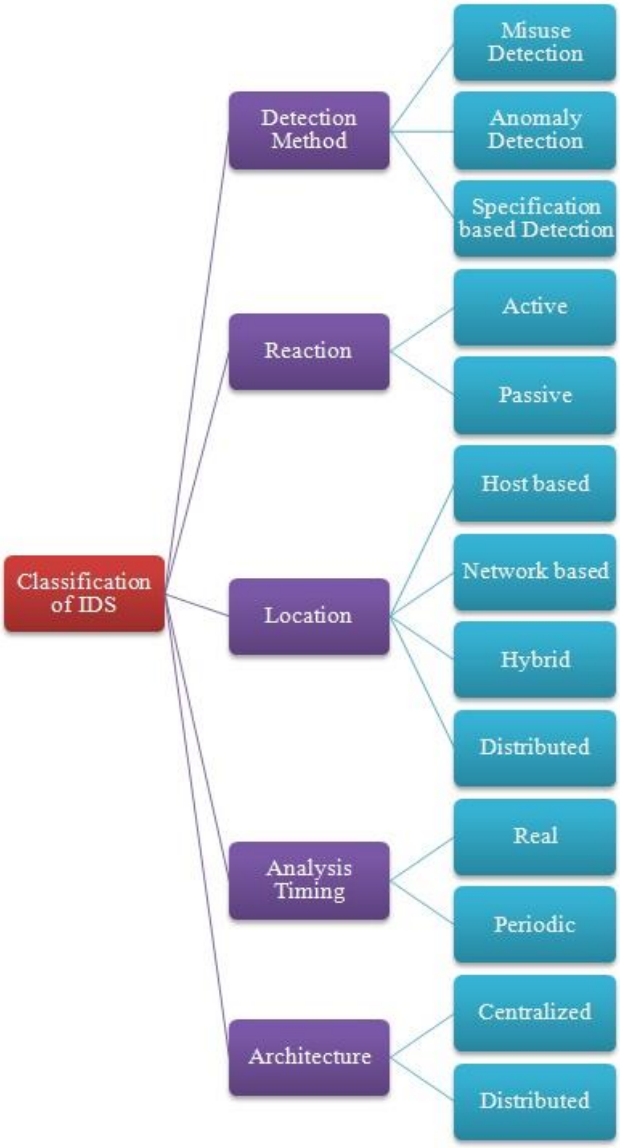

Figure 1: Classification of IDS

The stored or existing database is used to detect signature-based or misuse detection i.e. patterns match within the raising concerns, existing database about prospective misuse or attack. Anomaly detection investigates actions and raises an alert for odd behavior (deviation from normal) trends based on the possible behavior breach. Definite rules and strategies are specified in a specification-based intrusion detection skeleton, and deviations from such rules and strategies result in alarm generation [3].

An active IDS (now more commonly known as the IPS) is a system that is tuned to prevent alleged/malicious activities in progress without the need for an operator to act.

A passive intrusion detection system is one that is solely programmed to track, evaluate, and alert an operator of possible flaws and attacks [4].

A HIDS only monitors incoming device packets and outgoing device packets and alerts the system manager if it founds disbelieving or malevolent activities are detected. Network based IDS observe the traffic between the Sender Process and Receiver Process through the Protocol Packets shared between them. In a hybrid detection design, data from the host representative or computer is assorted with network understanding to provide an entire representation of the network-based system.

The hybrid IDS is more powerful than the majority of intrusion detection systems. In Distribution based IDS, multiple Sensors are collaboratively working together and generates an Intrusion Report, which will be sent to the central authority called Distributed IDS for further processing and necessary actions, in case, an Intrusion or Malicious Attack is detected [5].

A dynamic or real-time intrusion-detection tool gathers information about behavior taken on the environment in real time and conducts a continuous, real-time analysis. A static or periodic intrusion-detection tool takes a snapshot of the environment on a regular basis and analyses it for compromised applications, configuration bugs, and other issues. In centralized IDS takes a watchful eye on the network from a single point of centralized point of access like the gateway, servers, workstations, or PCs and focuses on activities spread across the entire organization while in distributed architecture multiple locations are involved, the data are collected from various remote locations and send to the master controller for decision making or action taken. Different groups of systems can be set up to serve different portions of the network.

RELATED WORK

Feature Engineering or Dimension Reduction are common terms used to dropping the size of obtainable Dataset in the resulting reduced Dataset that is indistinguishable to the original Dataset and all of the insights in the original Dataset are assured for the resultant Dataset occurs. Feature Extraction is a method of generating a fresh feature set from obtainable features, whereas Feature Selection is the method by which a portion of features from an existing Feature Set in such a way that the resulting Feature subset accurately reflects the original Dataset and the entire insights [6]. The performance of the algorithm or model being created is significantly influenced by the reduced dataset. A smaller dataset takes less memory than a larger dataset; the model will also take less time to learn over the smaller dataset. This finally increases the model's efficiency, performance, and classification accuracy. In comparison to Feature Extraction, Feature Selection is the most commonly used Dimension Reduction method. Feature Selection methods are further classified into two types, as Wrappers and Filters. Filters incorporate some estimation technique for evaluation of features. Estimation techniques include distance measures like Hamming, Manhattan, Euclidean, etc., consistency measures and correlation measures [7].

Significant Features are used in the resulting Dataset, whereas unnecessary features are detached, resultant in a smaller Dataset applied to train the model. Metrics like Mutual Information (MI), Correlation, Information Gain (IG), and others may quickly identify a feature's redundancy and relevance [8].

Imbalance Classes is a very common problem in numerous real-world scenarios. Due to this imbalance in classes, it becomes typical for Machine Learning models to predict all possible Classes with same efficiency.

This is because the model is more biased towards class having large number of instances, compared to classes with less number of instances. This paper proposed a Feature Selection technique that can solve the Bias-to-Majority problem. The authors suggested using a modified extension of GINI Index, termed Weighted GINI Index for Feature Selection. Their aim is to improve the ROC AUC and F Measure Performance Metrics. To judge the effectiveness of the proposed technique, the authors compared it with Chi Square, F Statistic and simple GINI Index based Feature Selection techniques. For all experiments, XGBoost Classifier is used along with the specific Feature Selection technique. Results prove that the Weighted GINI Index has significantly improved the F Measure, if 60 % of the Features are selected [9].

Most of the datasets in real-world suffers from the problem of imbalance classes. Due to this imbalance in classes, the performance of the Machine Learning models degenerates, as the minority classes are not predicted with same efficiency, compared to majority classes. This imbalance class problem can be solved to an extent using a suitable yet effective Feature Selection technique. The authors discovered a Weighted GINI Index based Feature Selection technique for Feature Selection. Decision Tree Classifier (CART) is used for testing the performance of the different Feature Sets for handling imbalance classes. The performance of method is compared with Chi Square, F Statistic and GINI Index based Feature Selection techniques to justify its effectiveness. They carried out their experiments on 2 bench-mark imbalanced datasets, namely STATLOG and LETTERS with imbalance ratio of above 9 and 24 respectively. The results and Statistical Analysis using Friedman Test and Wilcoxon Test justify that the proposed technique is efficient to overcome the Bias-to-Majority problem [10].

The performance of the Random Forest based Intrusion Detection System technique is tested over the bench-mark NSL KDD Dataset [11]. The proposed IDS works on the principle of Anomaly Detection. It is efficient in identifying the novel attack patterns or signatures; however they also exhibit the high False Alarm Rate. Machine Learning plays a crucial role in reducing the possibility of False Alarms in Anomaly Detection, in this manner making Anomaly Detection more successful with high Accuracy and low False Alarms. GINI Index is used for building the optimal Feature Subset on which the model is trained. Reducing the number of Features will reduce the computational complexity and processing time. However, only those Features are removed that are “redundant”; no “relevant” feature must be removed, otherwise the performance will degrade significantly. Based on GINI Impurity, 10 different Feature subsets are created and the optimal one is selected for training the Random Forest model. Results show that the model exhibits the Accuracy of 99.88 % and error rate of 0.12 %. The number of Features in optimal Feature subset is 25 [12].

The authors proposed an Intelligent Tree based IDS model for Cyber Security. The model is trained on reduced dataset, resulting in less processing time and computational complexity. GINI Index of Features is used to identify the features with high “relevancy” and low “redundancy”; collection of such features will collectively form the optimal feature subset, used for training the model. The resultant feature subset contains 19 features that will represent the original feature set. Decision Tree Classifier is trained using the reduced feature subset. UNSW-NB 15 bench-mark dataset is used for validating the performance of the proposed Feature Selection technique.

The proposed model achieves Accuracy, Precision, Recall, F Score and AUC of approx 97 %. The comparative analysis of Decision Tree using GINI based Feature Selection outperforms the other state-of-the-art techniques like K Nearest Neighbor, Support Vector Machine, Logistic Regression and Naïve Bayesian Classifier, trained over same reduced feature subset [13].

Mutual Information (MI) and Correlation supported feature reduction technique named Dense_FR with 20 features and Sparse_FR with 7 features are used for producing optimal features set of KDD-99 [14] datasets. These methods are also compared with Naïve Bayesian Classifiers, Logistic Regression, and Multi-Layer Perceptron and the recommended methods are showing prospective results over traditional methods [15]. Feature Reduction Back Propagation (FRBP) is based on Information Gain and correlation. This model is trained and tested on an Error back propagation algorithm with a forward phase and backward phase. This approach is separated into three modules depending on how they function, such as data pre-processing, information gain estimation, correlation results, constructing compatible sets to reduced features, back propagation neural network (BPNN) training, and forecast verification, among other things [16].

A two tier Feature Selection method where Feature Ranking is first tier and Additional Feature is second tier; here information gain based feature ranking is used. In the reduced subset of features, the four best features are chosen depending on their ranking. In the second tier of this feature subset, three more features are included. In the second tier, the correlation between features is employed as a criterion for identifying further features [17]. Correlation and symmetric uncertainty based feature selection method hold two steps. The correlation value is utilized in the first to find out the fitting subset of relevant attributes. The second step employs symmetric uncertainty to remove repeated features from the newly created feature subset, resulting in the finest set of reduced features [18]. Salih, and Abdulrazaq advised a Feature Selection method based on voting. The most relevant properties are identified using three estimates: Gain Ratio, Correlation and Information Gain. All estimation will calculate the set of most relevant features, and the top two rated features from each estimate will be combined to form the feature subset, which will represent the original dataset [19].

PROPOSED METHODOLOGY

This section presents the proposed method and related terminology.

3.1 Gini Index

The Gini index (Gini Impurity) calculates the probability that a particular variable will be incorrectly classified when selected at random. It only uses binary splitting and operates on categorical data, providing outcomes of "success" or "failure." After splitting along a particular attribute, the GINI index determines the purity of a given class. The optimal split raises the sets' purity as a result of the split. A statistic called Gini Impurity is used while creating decision trees to ascertain how a dataset's features are distributed. The Gini Index value lies between 0 and 0.5. If dataset is denoted as D with samples in k classes than the probability of samples that belongs to class i at a given node can be denoted as pi. Than the Gini impurity of dataset D is defined as:

$$Gini\ (D) = 1 - \sum_{i=1}^{k} {p_i}^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots 1$$

The Gini index calculates the variation from a perfectly equal distribution of income (or, in some situation, consumer expenditures) across individuals or families within an economy. The lower value of Gini index indicates higher relation among quality or relevancy and vice versa. The largest impurity is found in the node with a uniform class distribution. When all records fall under the same class, the least amount of impurity is obtained.

3.2 Algorithm

The purpose of feature selection is to reduce redundancy and increase relevancy. To archive the efficient reduced feature set Gini Index based method is presented.

Algorithm:

Step-1: Initialize KDD-99 Dataset.

Step-2: Calculate Gini Index of all attributes using the formula:

$$Gini\ (D) = 1 - \sum_{i=1}^{k} {p_i}^2 \ \ldots.\ (1)$$

Where pi denotes the probability of each unique value corresponding to particular attribute

Step-3: Select the top K% Attributes having lowest GINI Impurity where k in {20, 30, 50, 70, 80}

Step-4: Reframe the Dataset with selected Attributes corresponding to the value of k and 1 Class

Step-5: Use this Feature Reduced Dataset for Training and Performance Evaluation of the Machine Learning based IDS using Logistic Regression, Naïve Bayesian Classifier, Decision Tree and Random Forest.

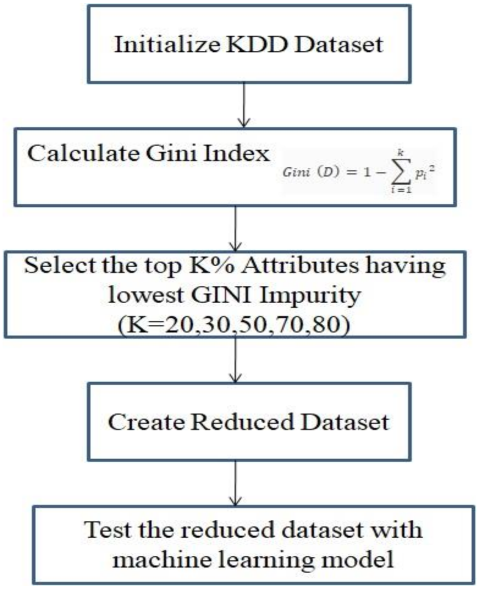
Initialize KDD Dataset
Calculate Gini Index
$Gini\ (D) = 1 - \sum_{i=1}^{k} p_i^{\,2}$
Select the top K% Attributes having lowest GINI Impurity (K=20,30,50,70,80)
Create Reduced Dataset
Test the reduced dataset with machine learning model

Figure 2: Proposed Method

In this paper five methods were proposed Gini20, Gini30, Gini50, Gini70 and Gini80 with 9, 13, 19, 25 and 32 attributes respectively based on the different index values. The value of K is used to select the reduced feature set accordingly.

Result And Discussion

Gini index based feature reduction data is compared with logistic regression, naïve bayes, decision tree and random forest techniques and results are compared with some existing methods as binary classification and multiclass classification.

4.1 Performance Parameter

The performance of the classifier is evaluated based on certain metrics like precision, recall, Accuracy, F1 Score for binary classification and F1 Micro, F1 macro, Jaccard Score, and Hamming Loss for Multiclass classification.

Precision is the measure of exactness of an IDS i.e., percentage of samples classified as attacks are actually attacks.

$$\text{Precision} = \frac{TP}{(TP+FP)} \qquad \text{…. (2)}$$

Recall is the measure of completeness i.e., percentage of samples classified as attacks are actually attacks.

$$\text{Recall} = \frac{TP}{(TP+FN)} \qquad \text{….(3)}$$

F1 measure is an alternative way for evaluating the performance of a classifier. It combines Precision and Recall measures, as it is the harmonic mean of both. Generally, F1 score is computed that assigns equal wattage to both the Precision and Recall.

$$\text{F1-Score} = \frac{2*TP}{(2*TP+FP+FN)} \qquad \text{….(4)}$$

Accuracy of an IDS is the percentage of correctly classified samples by the IDS.

$$\text{Accuracy} = \frac{(TP+TN)}{(P+N)} = \frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad \text{….(5)}$$

Hamming Loss represents the number of misclassified samples.

Accuracy = 1 - Hamming Loss

F1-score is evaluated as:

$$\text{F1-Score} = \frac{(2*\sum_{i=1}^{r} TP_i)}{(\sum_{i=1}^{r}(2*TP_i + FP_i + FN_i))} \qquad \text{….(6)}$$

Jaccard Score is an import metric for evaluating the performance of multiclass and multi-label classifiers. It is also useful to evaluate the performance of classifiers dealing with datasets having class imbalance among samples.

$$\text{Jaccard-Score} = \frac{\sum_{i=1}^{r} TP_i}{(\sum_{i=1}^{r}(TP_i + FP_i + FN_i))} \qquad \text{….(7)}$$

4.2 Dataset

The KDD-99 dataset is used for evaluation of model and it was developed by DARPA in 1999. It comprises 41 features and one target. Target or class can be binary means normal sample or attack sample; class can be multiclass means normal or any type of attack like DoS, Probe, R2L and U2R. KDD 10 percent contains 97, 278 normal samples and 3, 96, 743 attack instances comprising a total of 4, 94, 021 samples.

The attack instances are further classified in to four types with 4, 107 probe samples, 1, 126 R2L samples, 52 U2R samples and remaining 3, 91, 458 instances as DoS [20].

4.3 Binary classification

For the evaluation of model reduced datasets are used. Here KDD represents completed dataset without feature reduction [14], IGC3 is based on the information Gain and correlation [1], Dense_FR and Sparse_FR based on Mutual information and Kendall correlation [15, 16], Gini20 means Gini Index with K=20 and similarly Gini30, Gini50, Gini70 and Gini80 are compared. As far as number of attributes is concerned; the Gini20 is having 9 attributes, Gini30 with 13, Gini50 with19, Gini70 with 25 and Gini80 with 32 attributes.

Table 1:
Evaluation of Linear Regression

Data Set	Precision	Recall	F1 Score	Accuracy
KDD [14]	98.39	98.25	98.3	97.86
IG C 3 [1]	98.21	97.94	98.05	97.44
DenseFR [15]	99.55	97.57	98.54	98.34
SparseFR [15]	95.99	98.34	97.12	96.33
GINI 20	**99.74**	*97.1*	**98.38**	**98.07**
GINI 30	**99.68**	*97.33*	**98.47**	**98.21**
GINI 50	**99.54**	*97.42*	**98.45**	**98.24**
GINI 70	**99.58**	*97.31*	**98.42**	**98.16**
GINI 80	**99.55**	*97.56*	**98.53**	**98.33**

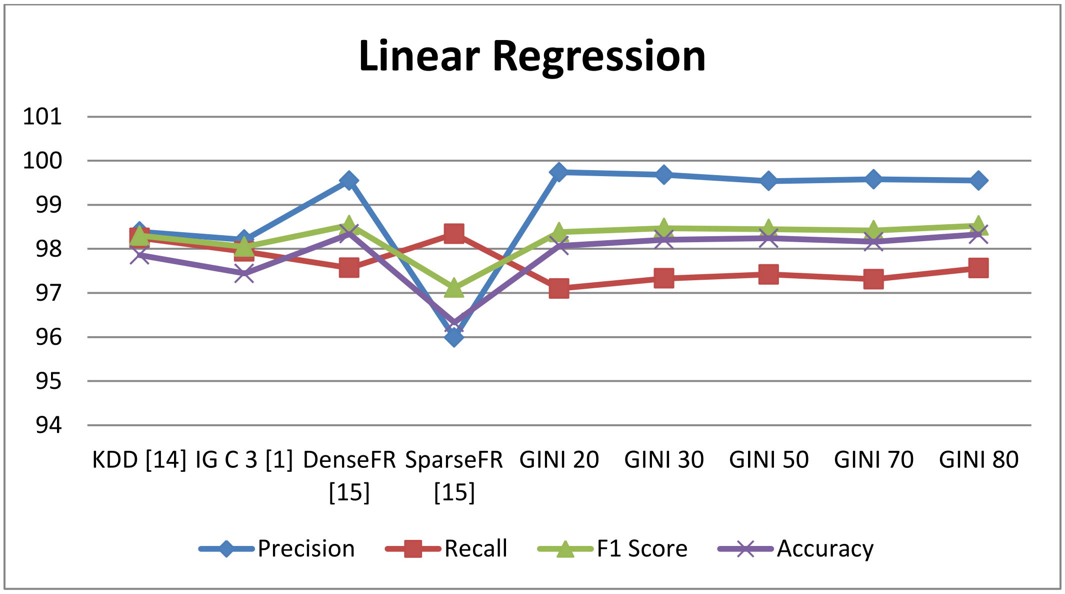

Figure 3: Results of Linear Regression

According to the table 1 and Figure 3; Logistic regression offers good precision with Gini20 and Gini30 with 99.74 and 99.68 values. SparseFR is good in recall value, Gini80 come-up with highest F1-score as 98.53.

DenseFR is having highest accuracy value of 98.34 and Gini80 with 98.33. Proposed Gini methods are good in precision and F1-score and competitive in accuracy and recall.

Table 2:
Evaluation of Naïve Bayesian

Data Set	Precision	Recall	F1 Score	Accuracy
KDD [14]	94.03	98.2	96.04	94.59
IG C 3 [1]	94.76	98.2	96.41	94.96
DenseFR [15]	96.59	92.53	94.31	92.42
SparseFR [15]	69.9	54.33	56.37	52.09
GINI 20	**98.95**	88.78	93.44	90.51
GINI 30	**98.85**	89.78	93.96	91.29
GINI 50	**97.53**	92.3	*94.7*	*92.81*
GINI 70	**96.84**	92.51	*94.43*	*92.55*
GINI 80	*96.59*	*92.53*	*94.31*	*92.42*

As per table 2; Naïve Bayesian gives good precision with Gini20 and Gini30 with 98.95 and 99.85 respectively. IGC3 and KDD original good in recall value.

F1-score and accuracy need to improve for proposed Gini methods. The proposed methods are failed to offer prominent results with Naïve Bayesian.

Table 3:
Evaluation of Decision Tree

Data Set	Precision	Recall	F1 Score	Accuracy
KDD [14]	99.93	99.67	99.8	99.71
IG C 3 [1]	99.91	86.32	89.83	89
DenseFR [15]	99.75	99.46	99.61	99.46
SparseFR [15]	99.55	98.46	99	98.88
GINI 20	*99.93*	99.66	*99.8*	*99.71*
GINI 30	**99.94**	99.65	99.79	99.7
GINI 50	**99.94**	**99.69**	**99.82**	**99.73**
GINI 70	*99.93*	*99.67*	*99.8*	99.7
GINI 80	**99.95**	99.66	**99.87**	*99.71*

According to the table 3; Decision Tree offers good precision with Gini80 as 99.95 and Gini50 as highest recall with 99.69.

Gini50 also come-up with highest F1-score as 99.82 and accuracy 99.73. Rest Gini methods are also showing compromising results precision, recall, F1-score and accuracy.

Table 4:
Evaluation of Random Forest

Data Set	Precision	Recall	F1 Score	Accuracy
KDD [14]	99.99	99.65	99.82	99.73
IG C 3 [1]	99.97	85.89	89.63	88.84
DenseFR [15]	99.79	98.78	99.28	99.19
SparseFR [15]	99.55	98.48	99.01	98.9
GINI 20	*99.97*	*99.62*	*99.79*	*99.71*
GINI 30	*99.98*	*98.9*	*99.43*	*99.4*
GINI 50	*99.99*	*98.86*	*99.41*	*99.39*
GINI 70	*99.99*	*99.64*	*99.82*	*99.73*
GINI 80	*99.99*	*99.57*	*99.78*	*99.7*

Table 4 represents results of random forest methods. According to the results all dataset offers similar types of results in terms of precision, recall, F1-score and accuracy.

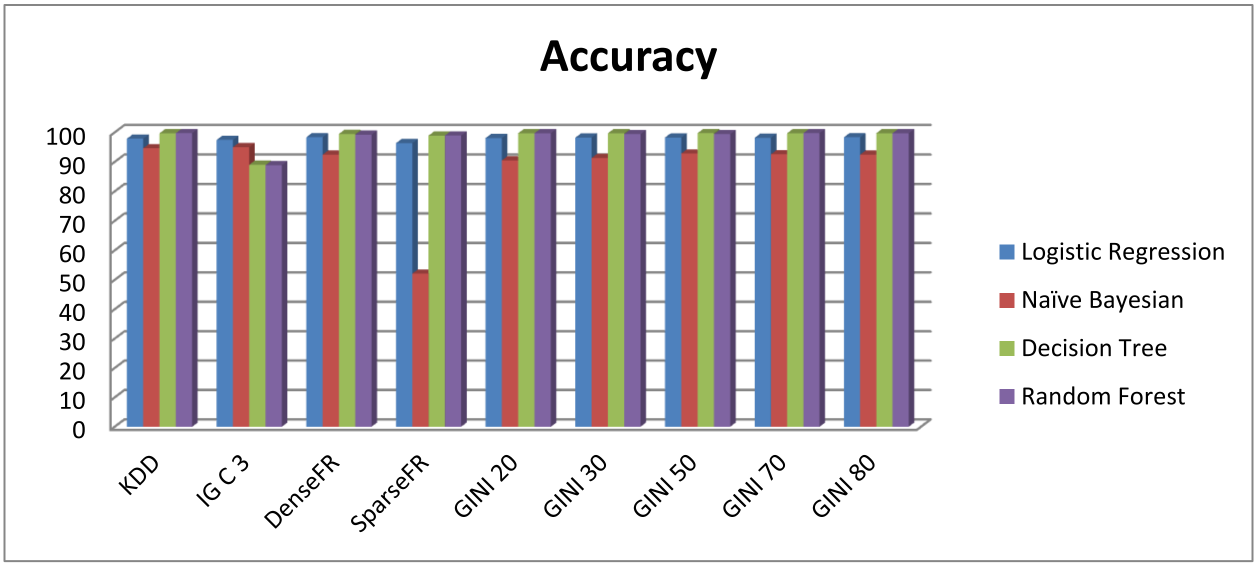

Figure 4: Accuracy Comparison for Binary classification.

Figure 4 shows the accuracy comparison of existing methods like KDD-99, IGC3, DenseFR and SparseFR along with proposed methods Gini20, Gini30, Gini 50, Gini70 and Gini80 with respect to Logistic regression, Naise Bayesian, Decision Tree and Random Forest. Proposed methods are showing the slightly better results as compared to existing.

4.4 Multiclass Classification

The performance of multi-class Classification can be evaluated through hamming loss, Jaccard Macro, Jaccard Micro, F1-Macro and F1-Micro.

Table 5:
One v/s Rest (Logistic Regression)

Data Set	H Loss	F1 MAC	JAC MAC	JAC MIC	ACC / F1 MIC
KDD [14]	2.06	53.4	52.27	96.02	97.94
IG C 3 [1]	4.29	52.34	49.08	91.88	95.71
DenseFR [15]	2.21	**59.3**	55.13	95.71	97.79
SparseFR [15]	2.45	55.97	52.77	95.26	97.55
GINI 20	*2.28*	*53.43*	*51.63*	*95.59*	*97.72*
GINI 30	*2.98*	50.91	*49.19*	*94.3*	*97.02*
GINI 50	*2.48*	*52.86*	*51.37*	*95.21*	*97.52*
GINI 70	*2.26*	*53.25*	*52*	*95.63*	*97.74*
GINI 80	***2.16***	*53.19*	*51.9*	*95.83*	*97.84*

Table 5 shows the performance analysis of Logistic Regression for multi-class.

Here Gini80 having the lowest hamming loss i.e. 2.16, DenseFR with highest F-1 macro of 59.3, Gini20 and Gini80 offers higher Jaccard micro 97.72 and 97.84 respectively.

Table 6:
One v/s Rest (Naïve Bayesian)

Data Set	H Loss	F1 MAC	JAC MAC	JAC MIC	ACC / F1 MIC
KDD [14]	6.81	48.99	45.6	87.7	93.19
IG C 3 [1]	10.69	44.99	40.94	81.77	89.31
DenseFR [15]	6.48	**52.76**	**47.46**	88.06	93.52
SparseFR [15]	**5.92**	50.77	46.46	89.01	**94.08**
GINI 20	*6.11*	*49.46*	*47.41*	***89.57***	***93.89***
GINI 30	*7.99*	*48.54*	*45.99*	*86.7*	*92.01*
GINI 50	*6.63*	*49.36*	*45.91*	*87.98*	*93.37*
GINI 70	*6.62*	*49.39*	*45.93*	*88*	*93.38*
GINI 80	*6.82*	*48.99*	*45.6*	*87.7*	*93.18*

Table 6 shows the performance analysis of Naïve Bayesian for multi-class. According to the values in table SparseFR having highest Accuracy/ F-1 Mico and lowest value of Hamming loss, 94.08 and 5.92 respectively. DenseFR offers higher result in F-1 macro and jaccard macro. Gini20 offers highest result in jaccard micro.

Table 7:
One v/s Rest (Decision Tree)

Data Set	H Loss	F1 MAC	JAC MAC	JAC MIC	ACC / F1 MIC
KDD [14]	4.51	66.85	64.24	92.42	95.49
IG C 3 [1]	15.25	58.23	54.97	82.14	84.75
DenseFR [15]	1.17	63.89	59.96	97.69	98.83
SparseFR [15]	1.8	60.83	57.66	96.49	98.2
GINI 20	**0.48**	**67.31**	**64.83**	**99.04**	**99.52**
GINI 30	**0.65**	**68.97**	**66.64**	**98.71**	**99.35**
GINI 50	**0.37**	**69.29**	**67.14**	**99.26**	**99.63**
GINI 70	*4.5*	*67.78*	*65.49*	*92.45*	*95.5*
GINI 80	*4.47*	*67.1*	*64.34*	*92.5*	*95.53*

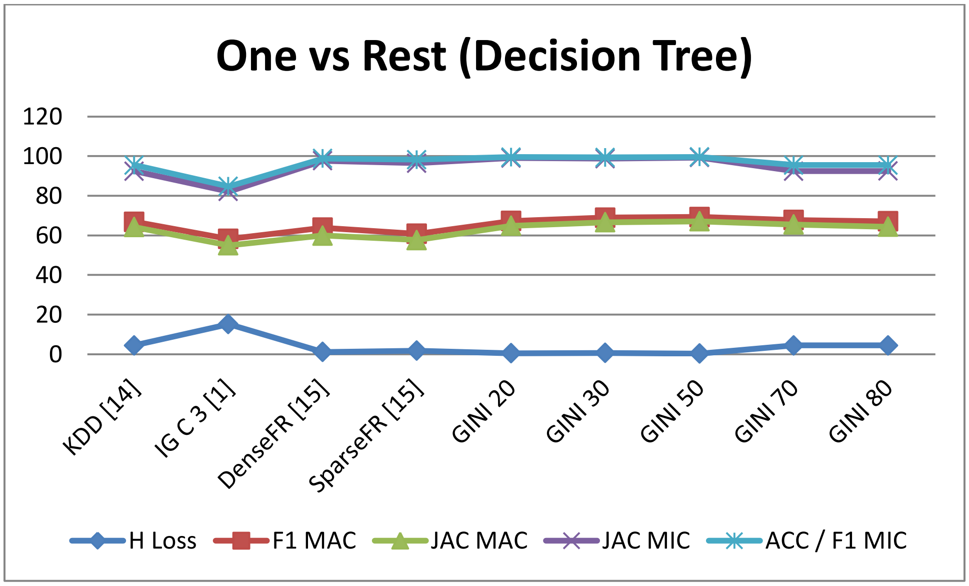

Figure 5: Results of One vs Rest (Decision Tree)

According to the table7 and figure 5; Decision Tree in multiclass classification offers good results in hamming loss, jaccard and F-1 scores with Gini20, Gini30 and Gini50.

Table 8:
One v/s Rest (Random Forest)

Data Set	H Loss	F1 MAC	JAC MAC	JAC MIC	ACC / F1 MIC
KDD [14]	4.45	79.63	75.29	92.52	95.55
IG C 3 [1]	15.24	68.22	63.97	82.16	84.76
DenseFR [15]	1.29	71.88	67.08	97.47	98.71
SparseFR [15]	1.69	60.97	57.84	96.72	98.31
GINI 20	**0.36**	**77.71**	**73.7**	**99.3**	**99.64**
GINI 30	**0.65**	**75.46**	**71.74**	**98.73**	**99.35**
GINI 50	**0.84**	**71.98**	**68.33**	**98.34**	**99.16**
GINI 70	*4.68*	*70.75*	*66.65*	*92.08*	*95.32*
GINI 80	*4.46*	*75.97*	*71.63*	*92.5*	*95.54*

Multiclass classification with random forest is shown in table 8; as per table Gini20, Gini30 and Gini50 offers good results in hamming loss, jaccard and F-1 score.

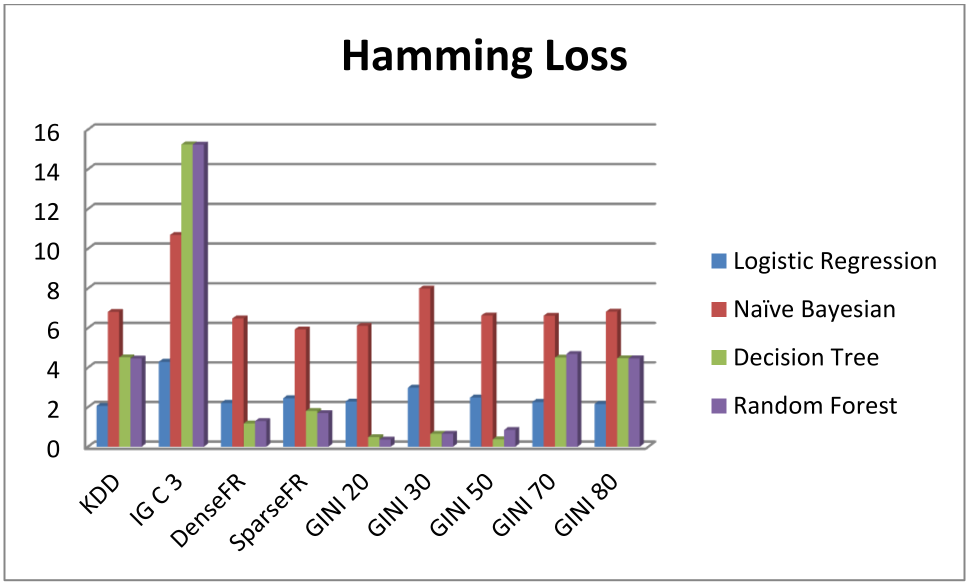

Figure 6: Hamming Loss comparison for multiclass classification.

Figure 6 shows the Hamming Loss comparison of existing methods like KDD-99, IGC3, DenseFR and SparseFR along with proposed methods Gini20, Gini30, Gini 50, Gini70 and Gini80 with respect to Logistic regression, Naive Bayesian, Decision Tree and Random Forest. Proposed methods are showing the lesser hamming loss as compared to existing.

CONCLUSION

The bigger dimensionality of dataset is a problem; they are collection of vast amounts of data i.e. scattered in several classes and features. Extremely higher execution or computation power is required to process these datasets.

To reduce the training time and processing overhead processing of reduce features are suggested. There are various methods recommended by the various researchers for the same. In this paper Gini Index based feature reduction based techniques were proposed. Gini Index is used to find the probability of irrelevant attributes. Higher value indicates fewer relevancies. Based on the concept five methods were proposed Gini20, Gini30, Gini50, Gini70 and Gini80 with 9, 13, 19, 25 and 32 attributes respectively based on the different index values. According to the result obtained proposed methods showing betterment in the results for both binary classification and multiclass classification.

REFERENCES

[1] Manzoor, I., & Kumar, N. (2017). A feature reduced intrusion detection system using ANN classifier. Expert Systems with Applications, 88, 249-257.

[2] Mehmood, T., &Rais, H. B. M. (2016, August). Machine learning algorithms in context of intrusion detection. In 2016 3rd International Conference on Computer and Information Sciences (ICCOINS) (pp. 369-373). IEEE.

[3] Al-Yaseen, W. L., Othman, Z. A., &Nazri, M. Z. A. (2017). Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system. Expert Systems with Applications, 67, 296-303.

[4] Yue, W. T., &Çakanyıldırım, M. (2010). A cost-based analysis of intrusion detection system configuration under active or passive response. Decision Support Systems, 50(1), 21-31.

[5] Sahu, A., Maravi, Y. P., Sharma, S., & Mishra, N. (2020, February). Error back propagation based Recurrent Neural Networks for Intrusion Detection system. In 2nd International Conference on Data, Engineering and Applications (IDEA) (pp. 1-6). IEEE

[6] Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. Neurocomputing, 300, 70-79.

[7] Dua, M. (2020). Attribute selection and ensemble classifier based novel approach to intrusion detection system. Procedia Computer Science, 167, 2191-2199.

[8] Shen, P., Ding, X., Ren W. and Liu, S., 2021. “A Stable Feature Selection Method based on Relevancy and Redundancy”, Proceedings of SCSET 2020, published under Journal of Physics: Conference Series, Volume 1732, IOP Publishing.

[9] Liu, Haoyue, MengChu Zhou, Xiaoyu Sean Lu, and Cynthia Yao. "Weighted Gini index feature selection method for imbalanced data." In 2018 IEEE 15th international conference on networking, sensing and control (ICNSC), pp. 1-6. IEEE, 2018.

[10] Liu, Haoyue, MengChu Zhou, and Qing Liu. "An embedded feature selection method for imbalanced data classification." IEEE/CAA Journal of Automatica Sinica 6, no. 3 (2019): 703-715.

[11] NSL-KDD dataset. Available on: https://www.unb.ca/cic/datasets/nsl.html, Accessed on Jan 2022.

[12] Negandhi, Prashil, Yash Trivedi, and Ramchandra Mangrulkar. "Intrusion detection system using random forest on the NSL-KDD dataset." In Emerging Research in Computing, Information, Communication and Applications, pp. 519-531. Springer, Singapore, 2019.

[13] Al-Omari, Mohammad, Majdi Rawashdeh, Fadi Qutaishat, Mohammad Alshira’H, and Nedal Ababneh. "An intelligent tree-based intrusion detection model for cyber security." Journal of Network and Systems Management 29, no. 2 (2021): 1-18.

[14] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, accessed on June 2022.

[15] Dubey, G.P. and Bhujade, R.K., 2021. “Optimal feature selection for machine learning based intrusion detection system by exploiting attribute dependence”. Materials Today: Proceedings, 47, pp.6325-6331.

[16] Dubey, Ghanshyam Prasad, and Bhujade, Rakesh Kumar, 2021. "Investigating the Impact of Feature Reduction Through Information Gain and Correlation on the Performance of Error Back Propagation Based IDS." International Journal of Electrical and Electronics Research (IJEER), Volume 9, Issue 3, e-ISSN: 2347-470X, pp: 27-34.

[17] Alhaj, T.A., Siraj, M.M., Zainal, A., Elshoush, H.T. and Elhaj, F., 2016. "Feature selection using information gain for improved structural-based alert correlation." PloS one 11, no. 11, e0166017. Online available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0166 01 7.

[18] Shahbaz, M.B., Wang, X., Behnad, A. and Samarabandu, J., 2016. "On efficiency enhancement of the correlation-based feature selection for intrusion detection systems." In 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), pp. 1-7.

[19] Salih, A.A. and Abdulrazaq, M.B., 2019. "Combining best features selection using three classifiers in intrusion detection system." In 2019 IEEE International Conference on Advanced Science and Engineering (ICOASE), pp. 94-99.

[20] KDD-Cup-99 task description. Available at https://kdd.ics.uci.edu/databases/kddcup99/task.html accessed on Nov-2022.

[21] Dr. Rakesh Kumar Bhujade, Dr. Stuti Asthana, “An Extensive Comparative Analysis on Various Efficient Techniques for Image Super-Resolution”, The International Journal of Emerging Technology and Advanced Engineering (ISSN 2250–2459(Online), Volume12, Issue 11, November 2022, 153-158.

[22] Dr. Rakesh Kumar Bhujade, Dr. Stuti Asthana, “An Extensive Comparative Analysis on Various Efficient Techniques for Image Super-Resolution”, The International Journal of Emerging Technology and Advanced Engineering (ISSN 2250–2459(Online), Volume12, Issue 11, November 2022, 153-158.

[23] Nandini K. Bhandari, Manish Jain, “EEG Emotion Realization through Hybrid Network”, International Journal of Emerging Technology and Advanced Engineering (ISSN 2250–2459(Online), Volume 12, Issue 8, August 2022, 187-195.