

Towards Intelligent Machine Learning Models for Intrusion Detection System

¹Pullagura Indira priyadarsini*, ²P V R N S S V Sai Leela, ³Bankapalli Jyothi

¹Department of Information Technology, Vardhaman College of Engineering, Hyderabad, Telangana, India,

²Department of Computer science and Engineering, GITAM Deemed to be University, Hyderabad, Telangana,

³Assistant Professor, Department of Computer science and Engineering, GITAM Deemed to be University, Hyderabad, Telangana

¹indupullagura@gmail.com, ²saileelavenkata@gmail.com, ³jbankapa@gitam.edu

Article History: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

Abstract: The Internet has become an important resource for mankind. Explicitly information security is an interminable domain to the present world. Hence a more potent Intrusion Detection System (IDS) should be built. Machine Learning techniques are used in developing proficient models for IDS. Imbalanced Learning is a crucial task for many classification processes. Resampling training data towards a more balanced distribution is an effective way to combat this issue. There are most prevalent techniques like under sampling and oversampling. In this paper, the issues of imbalanced data distribution and high dimensionality are addressed using a novel oversampling technique and an innovative feature selection method respectively. Our work suggests a novel hybrid algorithm, HOK-SMOTE which considers an ordered weighted averaging (OWA) approach for choosing the best features from the KDD cup 99 data set and K-Means SMOTE for imbalanced learning. Here an ensemble model is compared against the hybrid algorithm. This ensemble integrates Support Vector Machine (SVM), K Nearest Neighbor (KNN), Gaussian Naïve Bayes (GNB) and Decision Tree (DT). Then weighted average voting is applied for prediction of outputs. In this work, much Experimentation was conducted on various oversampling techniques and traditional classifiers. The results indicate that the proposed work is the most accurate one among other ML techniques. The precision, recall, F-measure, and ROC curve show notable outcomes. Hence K-Means SMOTE in parallel with ensemble learning has given satisfactory results and a precise solution to the imbalanced learning in IDS. It is ascertained whether ensemble modeling or oversampling techniques are dominating for Intrusion data set.

Keywords: Accuracy, F-measure, Feature Selection, Intrusion Detection System, Machine Learning, OWA, KDD cup 99 data set, K-Means, SMOTE, Precision, Recall.

1. Introduction:

In recent days with the abnormal increase of commercial and public services that are accessible through the World Wide Web, data security is compromising. This is because of vulnerabilities or "attacks" caused to the system or application software [1] [2]. Computer networks can evade those attacks with the use of number of access constraint strategies which are treated as filters. An Intrusion Detection System (IDS) filters and potentially monitors the actions arising in a system, and act out whether those events are indications of an attack and organizes an approved usage of the system [3]. In this regard, rapid advancement in technology in association with the vast boom of data has led to numerous issues. Among them, most of the issues are related to high dimensionality and imbalanced data. Since analyzing massive data is very hard and the classification performance also deteriorates. Several new methodologies in machine learning were emerging to handle them best. These are used in medical diagnosis, Intrusion Detection System (IDS), fraud detection, text classification, and so on. The IDS is considered as a classification and modeling system. Given above-said challenges, customized technologies such as neural networks, Support vector machines (SVM), K-nearest neighbor (KNN), logistic regression (LR) have inherent shortcomings.

High Data dimensionality is the crucial confronting issue for performance. It has reshaped statistical thinking against new scientific complications. Feature selection and extraction have become pivotal aspects. Since high dimensionality is challenging traditional statistical theory? Several novel insights are required and many new phenomena are to be explored. Therefore Reduction of High dimensionality and analysis is evolving as the biggest research area in statistics in the last twodecades [4]. The need for reduction in dimensionality is to avoid irrelevant and redundant data to get less computational cost and improve data quality. It also prevents the crisis of overfitting data [5]. Since the IDS data set is enormous, it needs feature selection. By this, there is a possibility of shortening the computational cost, time for categorization and thus proficiency is enhanced [6][7]. Class imbalance is a different significant challenge in machine learning. Unevenness in a data set occurs when some of the classes are under-represented compared to all other classes. It may occur with one majority and one minority class in case of a two-class problem. It also could ensue with one majority and many minority classes in case of Multi-class problem. Observing multiclass instances with dissimilar misclassification costs of classes is tougher than the usage of two class categorizations [8] [9]. There are most prevalent techniques like under sampling and oversampling in dealt with the imbalanced issue. Oversampling techniques are employed for balancing the dataset by increasing the minority one. A SMOTE (Synthetic Minority Over-sampling Technique)

[25] is capable of handling this imbalanced problem successfully. It generates new instances of minority class by considering the k-nearest neighbors.

Ensemble learning is ascertained to augment the predictive ability by the unification of single classifiers and has been pragmatic to imbalanced data-sets [10]. Bagging [11] is one of the traditional ensemble methods employed for improvising classification techniques. Another popular approach Boosting [12] [13] is implemented as a sequential ensemble type. Tracking all the related issues on ensemble methods and single models on agribusiness time-series data [14] we have driven an ensemble framework for IDS. In this proposed work, the authors have ascertained whether ensemble modeling or oversampling techniques are dominating for Intrusion data set.

Our work suggests a novel hybrid algorithm HOK-SMOTE, which considers an Ordered weighted averaging (OWA) approach for choosing the best features from the KDD cup 99 data set and K-Means SMOTE for imbalanced learning. In contrast, Ensemble methodology is applied for classification of attacks with normal data. In this work, much Experimentation was conducted on various oversampling techniques and traditional classifiers. The succeeding fragment of this paper is planned as follows. Section 2 gives detailed and current methods in feature selection and ensemble learning. Section 3 featured the proposed practice. Section 4 presents empirical work and results inferred. It also aims in presenting the comparisons of the performance of proposed and other oversampling methods. Lastly in section 5, the conclusions are presented.

2. Related work

Machine learning has caught the attention of a lot of researchers to provide solutions, especially for wide-ranging big data problems. It operates eventually on high dimensional data in making prudent predictions and is gaining fresh momenta. On the confronts mentioned above, there are some previous works for building prudent IDS, for apt feature selection, oversampling, and hybrid techniques. Recently Salo et al. [15] associates the feature selection approaches of Information Gain and Principal Component Analysis (PCA) with an ensemble learner based on Instance-Based learning algorithms (IBK), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). Indira et al [16] has combined Canberra distance, City block distance, Euclidean distance, Chebyshev distance, and Minkowski distance and produced a fuzzy ensemble feature selection and produced remarkable results using ensemble learning. In Tama et al [17], improved IDS based on hybrid feature selection and two-level classifier ensembles on the NSL-KDD and UNSW-NB15 datasets. For machine learning community, the focuses are well known, so we are not describing in detail here.

2.1. On Feature Selection

Based on the strategy that some specific features have more tendencies in improvising classification accuracy, feature selection is considered as a key step in data mining. In a study of Wang et al [18], a conversion of original features with logarithms of the marginal density ratios has been done by procuring novel, transformed features and hence refined the performance of an SVM model. At recent times, Vajihajisalem et al. [19] illustrated a fusion of two methods such as artificial bee colony (ABC) and artificial fish swarm (AFS) along with the fuzzy C-means clustering (FCM) for dividing the training data set and correlation-based feature selection (CFS) techniques for feature selection. Zhang et al. [20] have specified a cost-based feature selection procedure with multi-objective particle swarm optimization (PSO). In that they have done a comparison of multi-objective PSO with several multi-objective feature selection approaches verified on five benchmark data sets.

2.2. On oversampling techniques

Recently there have been numerous approaches to handle the class imbalance problem. These approaches can be categorized into two Ways: data level approaches and algorithm level approaches. Data level approaches include oversampling, under-sampling, and SMOTE (Synthetic Minority Over-sampling Technique) techniques. Algorithm level approaches are Threshold method, one class learning, and Cost-sensitive learning [21]. Random over Sampling (ROS) is an algorithm for increasing the size of minority class instances to rebalance class distribution in a dataset. This scheme is arbitrarily replicating minority class samples. Thus, the learning rate of this technique is slow. The drawback of ROS may cause over fitting. On the other hand, it can duplicate the number of errors [22]. Random Under Sampling (RUS) is one of the methods to balance the imbalanced data set. This method is modifying the data before learning. RUS removes some majority class instances to rebalance the instances in classes in a particular dataset. This approach achieves faster learning because it has less data points than the original sample. This method has a major drawback of loss of valuable information while randomly eliminating the majority class instances. Thus, it may cause misclassification because of eliminating the important patterns in a given dataset [23].

Oversampling techniques are employed for balancing the dataset by increasing the minority one. There are varieties in oversampling techniques that are anticipated in recent times such as Random oversampling [24], SMOTE [25]. SMOTE is capable of handling this imbalanced problem successfully. It generates new instances of minority class by considering the k-nearest neighbors. Then, it takes the difference between the particular feature vector and its nearest neighbor under consideration. A random number between 0 and 1 multiplies this difference. Finally, this multiplication output adds to the particular feature vector to increase the instances of the

minority class. There are several applications based on the SMOTE technique [26], borderline-SMOTE [27], safe-level-SMOTE [28], ADASYN [29], SVM-SMOTE [37] and SMOTE-RSB [30].

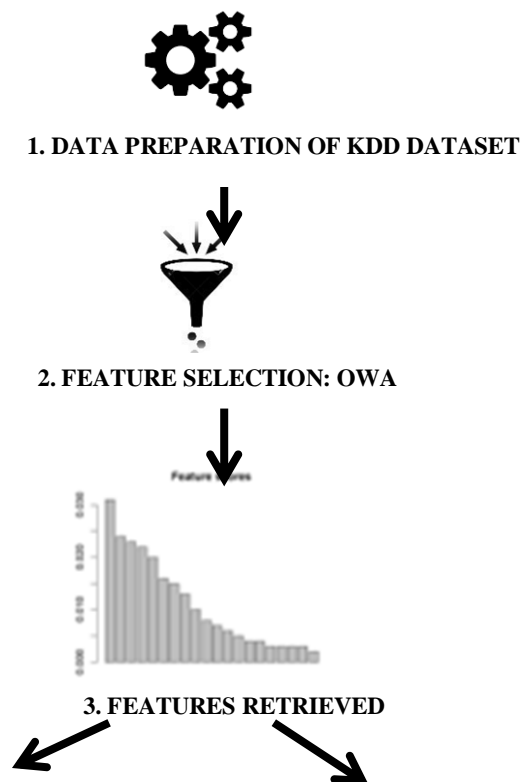
2.3. On ensemble and hybrid approaches

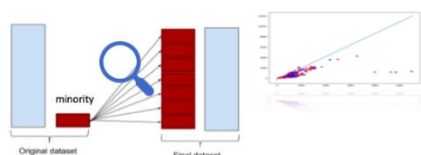
In a work done by Ren et al [31], they have given a data augmentation method to construct IDS, named DO_IDS. In this, they used data sampling, iForest for sampling data, and fusion of Genetic Algorithm (GA) and Random Forest (RF) for optimizing sampling ratio. In the procedure of feature selection, a grouping of GA and RF is done for choosing the optimum feature subset. Then DO_IDS is assessed by the intrusion detection dataset UNSWNB15. Recently, Indira et al has given a novel ensemble model for IDS based on two algorithms Fuzzy ensemble Feature selection and Fusion of multiple classifiers [32]. Many hybrid approaches using both feature selection and ensemble methods have been produced to improve the performance of IDSs. In the research made by Malik et al. [33], a combination approach of Particle Swarm Optimization (PSO) and Random Forest (RF) is used for dimensionality reduction and RF is for classification. It has enhanced percentage of the Accuracy. Pham et al. [34] built a hybrid model, which utilizes gain ratio technique as feature selection and bagging to combine tree-based base classifiers. Experimental results shown paramount performance is achieved by the bagging model that exploited J48 as the base classifier and done on a 35-feature subset of the NSL-KDD dataset. Abdullah et al. [35] also built IDS using IG based feature selection and ensemble learning algorithms. The experiments on the NSL-KDD dataset designate that the uppermost accuracy attained when using RF and PART as base classifiers with the product probability rule.

3. Proposed Methodology

In view of challenges in imbalanced data, intelligent techniques are obtained by taking the advances in machine learning for the implementation of effective IDS. The proposed work augments the concerns of high dimensionality and imbalanced data distribution with an innovative feature selection method and an oversampling technique respectively. Two different strategies are used in the proposed methodology. One suggests a novel hybrid algorithm HOK-SMOTE, which considers an Ordered weighted averaging (OWA) approach for choosing the best features from the KDD cup 99 data set and K-Means SMOTE for imbalanced learning, followed by individual classifiers. The other implements an Ensemble model with the OWA Feature selection without sampling techniques. It is built with four base classifiers. Ensemble classification is powerful than data sampling techniques for the up surging capability of classification for imbalanced data.

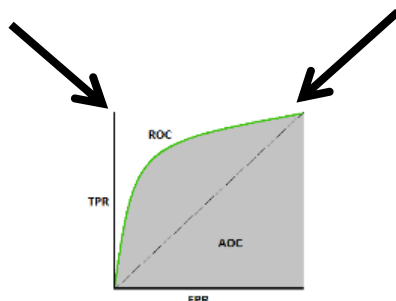
By modeling an intelligent algorithm HOK-SMOTE, oversampling of the minority class in the chosen dataset has been done. Figure 1 below depicts the pictorial framework of the proposed HOK-SMOTE, which has the following four components:





4. OVERSAMPLING: KMeans SMOTE

5. ENSEMBLE CLASSIFIER



6. RESULTS OBTAINED

Fig1. Proposed Flow of Work

- Data preparation: The first level is to convert raw data into a structure fit for analysis by applying various preprocessing to the original dataset.
- Feature Selection: To exceed the high-dimensionality problem, the feature selection approach based on ordered weighted averaging (OWA) is exploited to lessen the dimensionality of the data set.
- K-Means SMOTE oversampling: To focus on the imbalanced dataset problem, in this we utilized, K-Means Synthetic Minority Oversampling Technique for Over-sampling the minority (illegitimate) class.
- Ensemble Classification: Ensemble classifier is built based on weighted average voting on the base classifiers.

In search of a novel optimum feature set, filtering methods were used. The Intrusion data set has a majority of samples which are Quasi constant. Striving on this, we exploited aggregation operators for finding the best features. Ordered Weighted Average (OWA) [38] is the predominant one in information aggregation and it is given by Yager. Later it was used in multiple applications. In this paper, an ordered weighted average (OWA) methodology is used to obtain feature scores. For doing so, the critical part lies with identifying weights. OWA takes ‘AND’, ‘OR’ and averaging cases. Learning the weights is done by analyzing the data set.

Definition1: An Ordered weighted Averaging (OWA) operator of dimension n is a mapping $F: X^n \rightarrow X$, with associated weight vectors $W = \{w_1, w_2, \dots, w_n\}^T$. Such that F is defined as $F(a_1, a_2, \dots, a_n) = W * B^T = \sum_{j=1}^n w_j a_{id(j)}$. Where $B = \{b_1, b_2, \dots, b_n\}$ is an argument vector of ‘F’ in descending order, $a_{id(j)} = b_j$.

Then $a_{id(j)}$ is the largest element in the collection of the elements $\{a_1, a_2, \dots, a_n\}$. There are some conditions on which weights are to be noticed. By taking different weights, we can implement different OWA operator.

- 1) $W^* = W_1 = 1 \ \& \ W_j = 0$ for $j \neq 1$ which gives $F^*(a_1, a_2, \dots, a_n) = \text{Max}_i[a_i]$.
- 2) $W_n = 1 \ \& \ W_j = 0$ for $j \neq n$ which gives $F^*(a_1, a_2, \dots, a_n) = \text{Min}_i[a_i]$.
- 3) $W_n = W_j = \frac{1}{n}$ for all ‘j’ gives the simple average. Then $F^*(a_1, a_2, \dots, a_n) = \frac{1}{n} \sum_{i=1}^n a_i$

Therefore for each feature, OWA is calculated. It lies between ‘0’ to ‘1’. Here in this work, we applied averaging operator. For undertaking this, we have considered 41 different feature values. They should be scaled. Then feature1 values are ordered in descending. Then weights are taken as given in step3 above. Then applying OWA operation we get OWA score of feature1. Similarly the same procedure should be undergone for all remaining 40 features.

It is given in the HOK-SMOTE algorithm in figure 2 below; the feature selection part is done by considering all the features $\{F1, F2, \dots, F41\}$ from the data set. A threshold is selected. After obtaining the OWA scores, they are compared with the threshold. The output is returned and recorded. The Feature Selector ‘FS’ is the operation by which the optimized features are chosen based on a threshold. The algorithm is detailed as complies. The input is ‘KDD dataset’, ‘M’, ‘ x_i ’, ‘ d_i ’, ‘k’, ‘j’, ‘K’, ‘ C_j ’ and c_i . Where the KDD dataset is the dataset taken, M is the total number of features in the data set; C_j is the class label in the dataset, x_i is the individual feature, w_j is the weight vector, b_j is the argument vector, d_i is the OWA score, c_i is the number of clusters. The output of the HOK-SMOTE will be optimized Features (FS) and oversampled data samples (instances). Step 1 of the HOK-SMOTE algorithm is the data preparation techniques applied to the data set. Step 2 and 3 illustrates the Feature selection

approach followed in this proposed work. In step 4 of the algorithm, the resultant of OWA that is FS is passed. FS is the number of chosen features in the KDD dataset.

The process of KMEANS SMOTE [40] involves clustering, filtering, and oversampling. In step 5, the entire input space is clustered using KMEANS. In step 6, it finds 'k' clusters by reducing the within-cluster sum of squared error.

Reduce the within-cluster quantity of squared error is given by

$$\arg \min \sum_{i=1}^k \sum_{x_j \in C_i} \|x_j - c_i\|^2 \quad (1)$$

Then in step 7, repeat for each sample until reaching the closest mean and also calculate the new mean for each cluster. Now, filtering is done. In step 8; filter out clusters that have a high number of majority class samples. In step 9, assign more synthetic samples to clusters where minority class samples are sparsely distributed. In step 10, oversampling each filtered cluster is done by SMOTE.

SMOTE (f,n,k) finally gives oversampled data. The parameters of SMOTE () are 'f' is the number of filtered clusters, 'n' is the number of minority samples, and 'k' is the k number of nearest neighbors. The effect is balanced samples of both majority and minority classes. Then the dataset is fed to the classifier.

Proposed HOK-SMOTE Algorithm

Input: KDD data set, $M=\{F1,F2,F3,\dots,F41\}, w_j, b_j, x_i, d_i, k, j, K, C_j, c_i$

Output: FS, instances

Step 1: Normalize the KDD data set.

Labeling the features and classes

Step 2:

Apply ordered weighted averaging (OWA) as feature selector

(i) For each feature 'i' in the M

(ii) Do

$d_i = \text{Calculate} \sum_{j=1}^n w_j b_j$.

Return d_i

Step 3: $FS = \max \sum_{n=1}^M d_i$

Step 4: Pass the FS for imbalanced learning

Step 5: cluster the entire input space using KMEANS

Step 6: Distribute the number of samples to generate across clusters.

$$\arg \min \sum_{i=1}^k \sum_{x_j \in C_i} \|x_j - c_i\|^2$$

Step 7: Do

For each sample until the closest mean

Calculate the new mean for each cluster

Step 8: f = clusters which have a high number of majority class samples

Step 9: assign more synthetic samples to clusters where minority class samples are sparsely distributed.

Step 10: oversample each filtered cluster using SMOTE (f,n,k)

Step 11: return instances after oversampling

Step 12: End

SMOTE (f,n,k)

Do until the dataset is balanced

For each minority sample

Step 1: choose a minority class sample; find its k nearest neighbors

Step 2: Randomly select an instance among the k nearest neighbors

$dif = \|x_{origin} - x_k\|$,

Step 3: compute synthetic data in feature space.

$C_{syn} = x_{origin} + \|x_{origin} - x_k\| \times P_{uniform}$

Step 4: End

Fig2. Proposed Hybrid HOK- SMOTE Algorithm

The second part encompasses the Ensemble modeling. It is indulged to obtain more accurate and diverse classification predictions. Here Ensemble classifier is the one that combines predictions from four base learners. Firstly data is trained and model is built with K-nearest neighbor, and then followed by Gaussian Naïve Bayes, Decision Tree, and Support Vector Machine classifiers. Validation is done on the testing data obtaining four predictions. The ensemble algorithm is depicted below in figure 3. The resultant of four base classifiers is weighted average voting methodology through which the optimistic results are obtained. The prediction of class

labels is done based on the predicted probabilities of individual classifiers. The weighted average voting is given as the

$$\text{Final decision} = \text{Argmax} \sum_{j=1}^k w_j p_{ij} \quad (2)$$

Where w_j is the weight that can be assigned to the j^{th} classifier and ‘k’ is the total number of classifiers and $i = \{0,1\}$.

Ensemble Algorithm:

Input: KDD dataset with ‘FS’ features, n, Final decision, w_j, p_{ij}

Output: accuracy, precision, F1score, ROC,

Start:

Step 1: Take the KDD (n x FS) data set;

Step 2: Classify KDD dataset:

Step 3: model1=Training data is fit to KNearestNeighbor Algorithm;

Step 4: model2=Training data is fit to Gaussian Naïve Bayes Algorithm;

Step 5: model3=Training data is fit to Decision tree Algorithm;

Step 6: model4=Training data is fit to Support Vector Machine Algorithm;

Step 7: pred1=prediction is done on testing data for model1;

Step 8: pred2=prediction is done on testing data for model2;

Step 9: pred3=prediction is done on testing data for model3;

Step 10: pred4=prediction is done on testing data for model4;

Step 11: Final decision= $\text{argmax} \sum_{j=1}^m w_j p_{ij}$

Step 12: Evaluation Measures: accuracy, precision, F1score, ROC,

Step 13: End

4. Empirical work and results inferred

Empirical work is done on the KDD Cup 99 dataset [36]. It is the data set that was started from MIT’s Lincoln Lab. From this, a chunk of it is taken for experimentations and labeled as “KDD dataset”. It has records as a proportion to the records of KDD cup 99. It contains 10230 samples. It has 41 features and two class labels {attack, normal}. It has 9188 ‘attack’ samples and 1042 ‘normal’ samples.

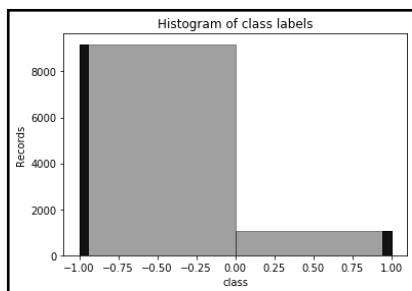


Fig 4. Class label distribution in the KDD dataset

Several experimentations were done on this data set. In this 50% data is taken for training and 50% is taken for testing. The histogram of records in the KDD data set is shown in figure 4.

The trials on the proposed scheme were conducted on R and Python interfaces in Anaconda 3.6 Environment [39]. Anaconda is an open-source platform for Python and R language. It holds about 100 of the commonly used Python packages for data science. The system type of 64-bit Operating System with a processor of Intel i5, Memory of 1TB, and 4GB RAM has chosen for doing several tests. It was commonly practiced in research areas like machine learning, artificial intelligence, data science, and so on. On the KDD data set, normalization is applied. The class labels are assigned as {1, 0} for ‘normal’ and ‘attack’ instances respectively. Features are also given labels as {F1, F2....F41}. Now to the data set, the OWA operator is applied. It has taken $F^* (a_1, a_2, \dots, a_n) = \frac{1}{n} \sum_{i=1}^n a_i$ (i.e., averaging case) where it is fixed weighting operator. The parameters for OWA are ‘x’ and ‘w’. Where ‘x’ is the individual feature values and ‘w’ is the weight vector (1/length(x)) executed in R platform. As a result, based on the score obtained the features are selected. Here the threshold is 0.45. The resultant features are

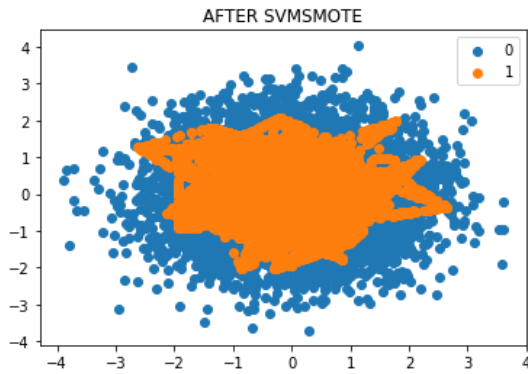


Fig9. After applying oversampling SVMSMOTE technique

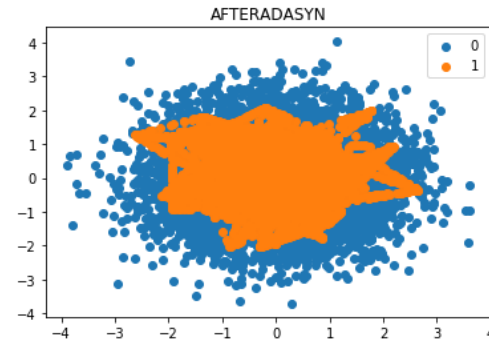


Fig 10.

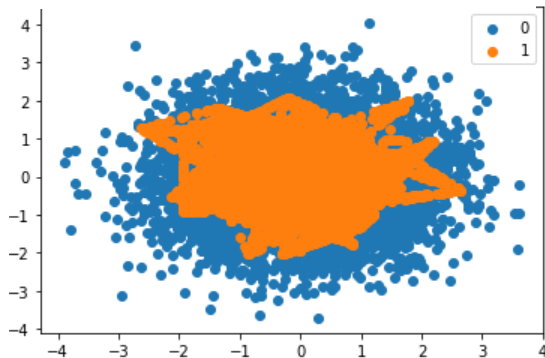


Fig 11.

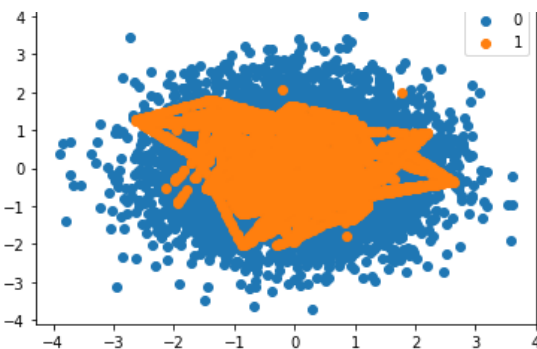
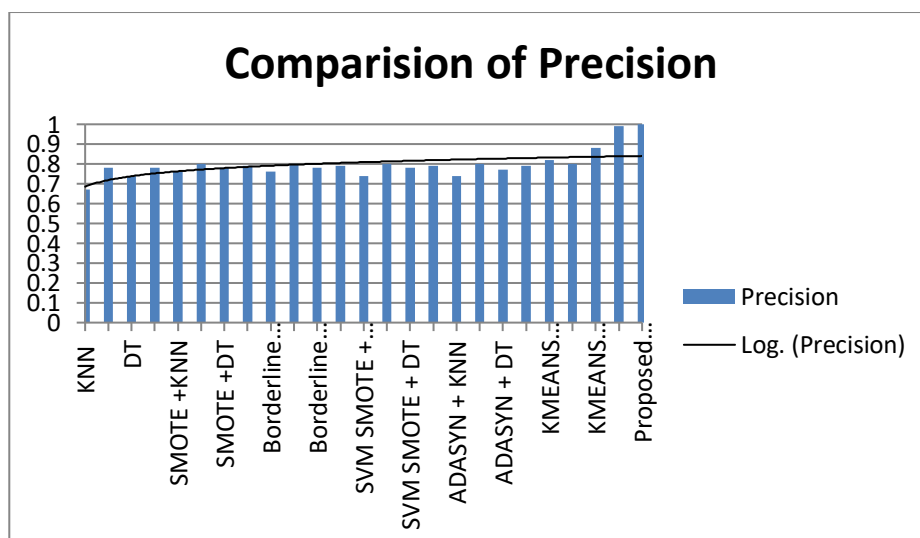
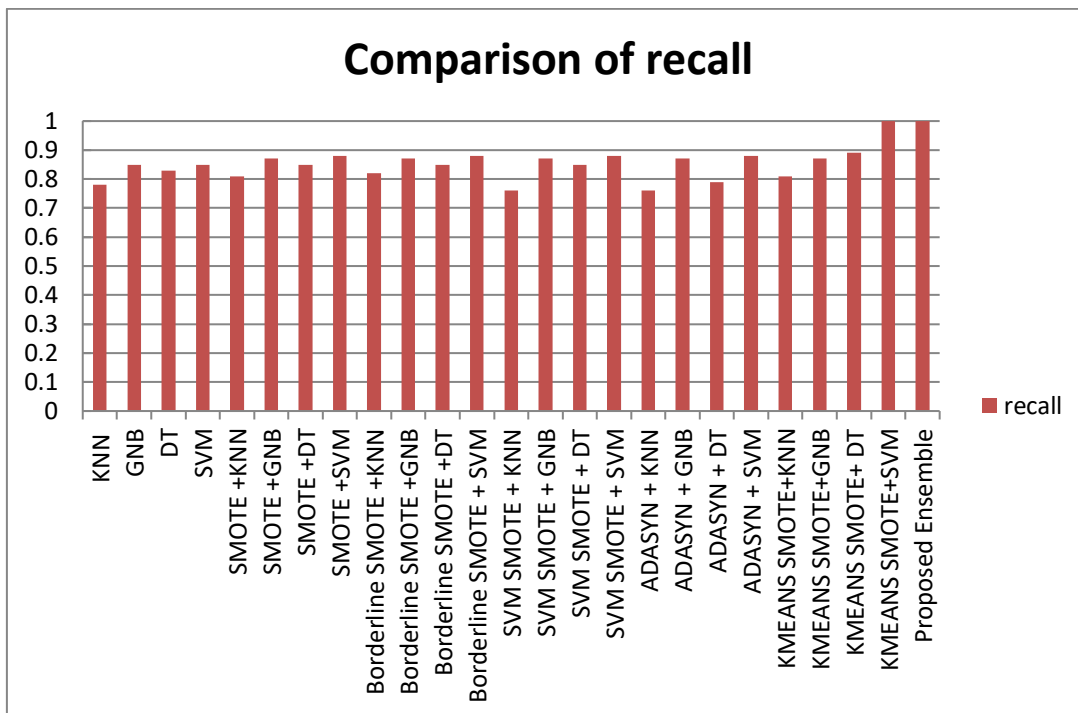
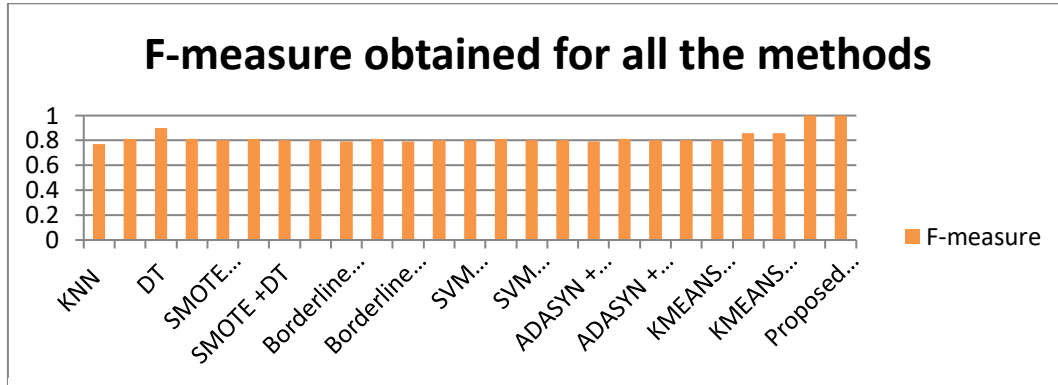


Fig 12.

Second initiative is, after applying OWA to the data set, 'FS' gives the chosen features. It is sent to the ensemble model. The one is an ensemble of four parallel classifiers namely KNN, GNB, DT, and SVM. The predicted output is obtained from the weighted average voting strategy. In the proposed work, a total of 25 different experimentations were done for evaluating the performance. The assessment standards employed for making the examinations are illustrated. 1) The F-measure is generalized as $2 * (\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})$. 2) The Area under the Curve is observed as the AUC. It insists on the quality of the classification methodology. It is evaluated as the excellence of the model's estimates regardless of what classification threshold is taken. 3) Precision is expressed as the success likelihood of generating a correct positive-class classification. 4) A Recall is designated as the model's proficiency in finding out all the data points of interest in a data set. 5) The accuracy of a classifier is professed as the division of the number of valid predictions to the complete amount of input samples. The precision, F-measure, recall, accuracy obtained for all the classifiers with and without oversampling techniques (Ensemble) are shown in figure 13, 14, 15 and 16 respectively.

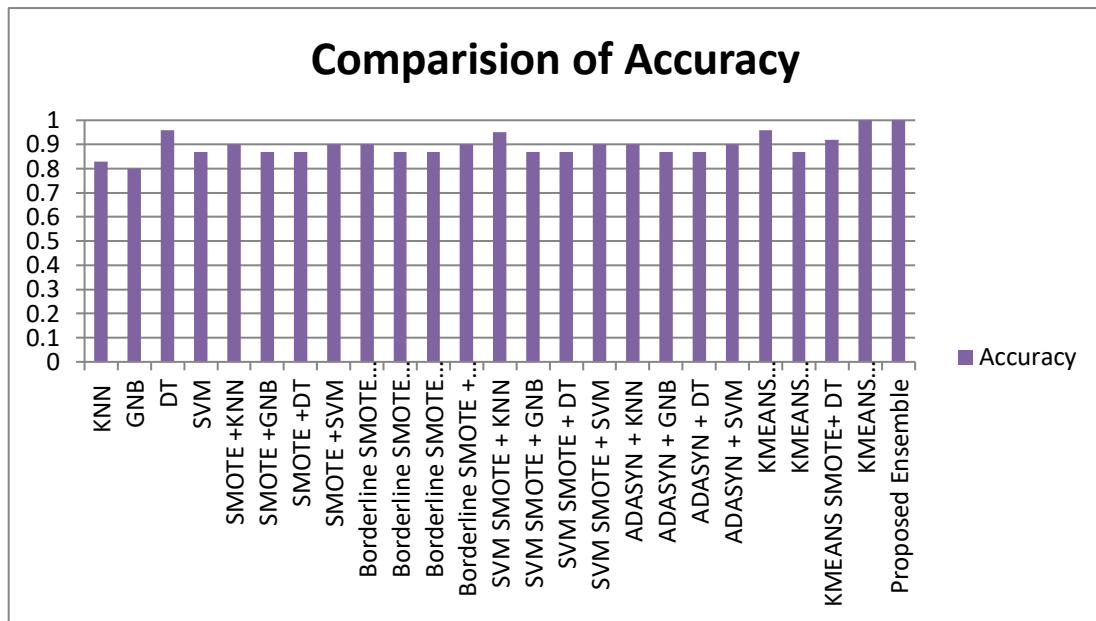




The model performance is evaluated and shown in the above figures. A large number of models were constructed and evaluated. Here the test data and train data are taken as 50%

The model performance is evaluated and shown in the above figures. A large number of models were constructed and evaluated. Here the test data and train data are taken as 50% of total data. The validations are completed on 50% of the data set. The training data is 5115 instances. So the oversampling is done on 4616 samples of class '0' and 499 samples of class '1'. Then after applying SMOTE, Borderline SMOTE, SVM SMOTE, ADASYN and the K-Means SMOTE, it yielded 4616 samples of class '0' and 4616 samples of class '1'.

From the observations made, it is evident that the proposed HOK-SMOTE on SVM and ensemble model predicts better accuracy. Proposed HOK-SMOTE on SVM has given 1.0 accuracy, F-measure 1.0 Precision, and Recall is 1.0. The Proposed Ensemble has given accuracy rate, F-measure, Precision, and recall of 1.0, Receiver Optimistic Characteristic (ROC) Curve achieved for the proposed ensemble is shown in figure 22 below with AUC 1.0. Since no single classifier gives less than 0.5 and where a perfect classifier gives 1.0. This can be taken as an optimistic measure of how the model is worthy. While in the Intrusion detection system both attack and normal data should be correctly predicted, so the accuracy rate of the model is more prioritized than anything else in evaluating the model, we are evident that the model is an optimistic one. And therefore we choose the ROC curve to determine the quality of the proposed model.



Several experimentations were done in this work; we show some of the ROC curves as evidence in proving our model is best apt for IDS. For analogies, ROC curves of the F-SMOTE-SVM model ('F' specifies Feature selection is based on OWA), F-SVM SMOTE-SVM model, HOK-KMEANS SMOTE-SVM model, and proposed ensemble are depicted below in figures 17, 18, 19 and 20 respectively. The proposed ensemble with OWA feature selection is compared with the Hybrid algorithm HOK-SMOTE on SVM. Hence it is obvious ensemble model with KNN, GNB, DT, SVM has given identical results with the oversampling techniques. Among all the oversampling techniques from the observations made, KMEANS SMOTE is paramount over others. The OWA scores achieved in R programming environment are shown in figure 21 below. The ROC curve of Proposed Ensemble in the Python environment is shown in figure 22.

Centered on these fallouts, we acclaim either KMEANS SMOTE or Ensemble modeling with the above said base classifiers for imbalanced high-dimensional datasets is good. Feature subset selection is also one of its inherent attainments for such sort of data sets. As such we recommend Ordered weighted averaging for opting optimal features.

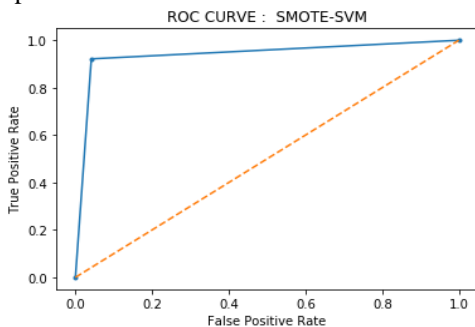


Fig17. ROC curve of F-SMOTE-SVM model

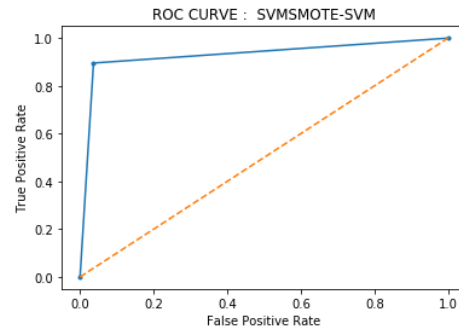


Fig18.

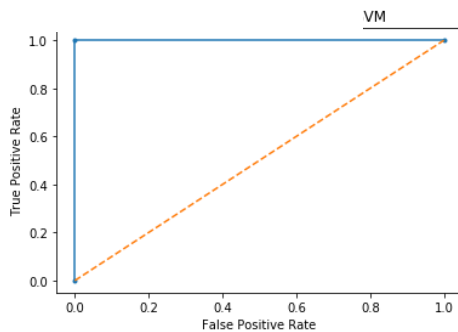


Fig 19. ROC curve of HOK-KMEANS SMOTE-SVM (Proposed Hybrid algorithm on SVM classifier)

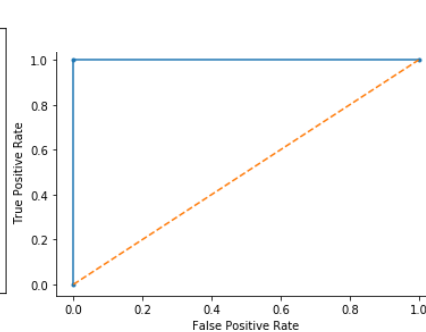


Fig 20. ROC curve of Proposed Ensemble (Without oversampling)

```

RGui (64-bit) - [R Console]
File Edit View Misc Packages Windows Help
[10143] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
[10165] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
[10187] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
[10209] 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
> y = owa(num1,w=rep(1/length(num1),length(num1)))
> y
[1] 0.1681623
> f<-paper14[2]
> vec=unlist(f, recursive = TRUE, use.names = TRUE)
> num1=as.numeric(vec)
> y = owa(num1,w=rep(1/length(num1),length(num1)))
> y
[1] 2.062854
> f<-paper14[3]
> vec=unlist(f, recursive = TRUE, use.names = TRUE)
> num1=as.numeric(vec)
> y = owa(num1,w=rep(1/length(num1),length(num1)))
> y
[1] 22.02727
> f<-paper14[4]
> vec=unlist(f, recursive = TRUE, use.names = TRUE)
> num1=as.numeric(vec)
> y = owa(num1,w=rep(1/length(num1),length(num1)))
> y
[1] 9.986022
> f<-paper14[5]
> vec=unlist(f, recursive = TRUE, use.names = TRUE)
> num1=as.numeric(vec)
> y = owa(num1,w=rep(1/length(num1),length(num1)))
> y
[1] 0.1656132
> f<-paper14[6]
> vec=unlist(f, recursive = TRUE, use.names = TRUE)
> num1=as.numeric(vec)
> y = owa(num1,w=rep(1/length(num1),length(num1)))
> y
[1] 0.7663295
|

```

Fig 21. The OWA scores (FS of the features calculated in R programming)

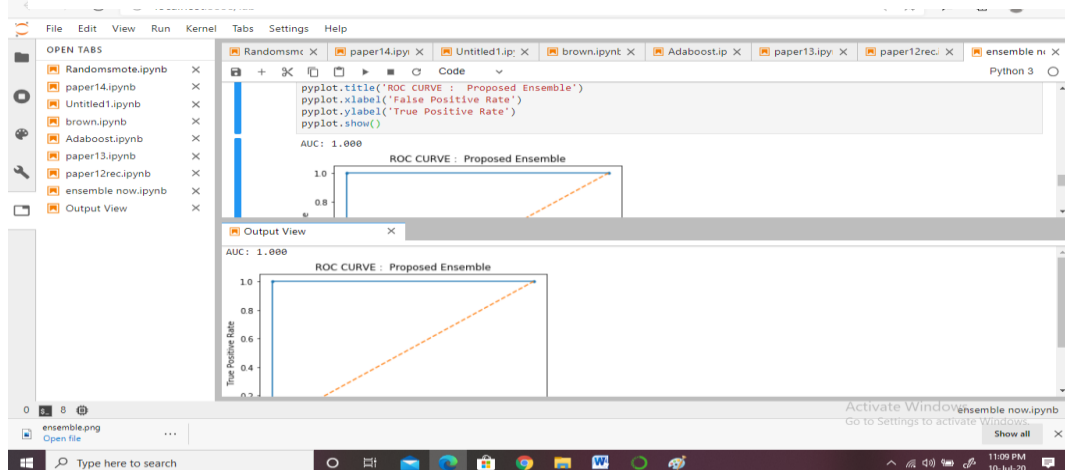


Fig 22.ROC curve of Proposed Ensemble (Without oversampling) and AUC obtained

5. Conclusions and discussions

Previously several machine learning techniques have proposed solutions to the challenges of imbalanced learning and increasing accuracy of the model but still, imbalanced data is not negotiating issue. Therefore, to combat this, data distribution challenge and high dimensionality, an oversampling technique, and an innovative feature selection method were emphasized in this paper. Our work suggested a novel hybrid algorithm that considers an ordered weighted averaging (OWA) approach for choosing the best features from the KDD cup 99 data set and K-MEANS SMOTE for imbalanced learning. Here an ensemble model is compared against the hybrid algorithm. This ensemble integrates SVM, KNN, Naïve Bayes, and DT. For predictions weighted average voting is applied. The results indicate that the proposed work is the most accurate one among other ML techniques. Hence K-Means SMOTE in parallel with ensemble learning has given remarkable results and a precise solution to the imbalanced learning in IDS. From the results shown the proposed ensemble with OWA feature selection is compared with the Hybrid algorithm HOK-SMOTE. Hence it is evident that the ensemble model with KNN, GNB, DT, and SVM has given identical results with the oversampling techniques. Among all the oversampling techniques from the observations made, KMEANS SMOTE is paramount over others. To our awareness, there are certainly no previous works that have reflected the properties of ensemble modeling methods against data sampling procedures for intrusion detection data set. As Upcoming work, we explore possibilities of novel models for ensemble learning and oversampling techniques. And as scope for feature selection we study other aggregation methods.

References:

1. J. McHugh, A. Christie, and J. Allen, "Defending Yourself: The Role of Intrusion Detection Systems", *IEEE Software*, Sept. Oct. (2000), pp. 42-51.
2. P.E. Proctor, "The Practical Intrusion Detection Handbook", Prentice Hall, (2001).
3. Ghosh A. K. (1999). Learning Program Behavior Profiles for Intrusion Detection. USENIX.
4. Donoho, D. L., High-dimensional data analysis: the curses and blessings of dimensionality. Aide-Memoire of the lecture in AMS conference "Math challenges of 21st Century(2000). Available at <http://www-stat.stanford.edu/~donoho/Lectures>.
5. Andrew Y N., "Preventing overfitting of crossvalidation data", In Proceedings of Fourteenth International Conference on Machine Learning, pages 245–253, (1997).
6. B. Sujitha, V. Kavitha, "Layered approach for intrusion detection using multi objective particle swarm optimization", *International Journal of Applied Engineering Research*, vol. 10, pp. 31999-32014, (2015).
7. E. Zorarpaco, S.A. Ozel, "A hybrid approach of differential evolution and artificial bee colony for feature selection", *Expert Systems with Applications*, vol. 62, pp. 91-103, (2016).
8. Wang, S.; Yao, X. Diversity analysis on imbalanced data sets by using ensemble models. In Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining, Nashville, TN, USA, 30 March–2 April (2009); pp. 324–331.
9. Krawczyk, B. Learning from imbalanced data: Open challenges and future directions. *Prog. Artif. Intell.* (2016), 5, 221–232. [CrossRef]
10. Fan, X.N.; Tang, K.; Weise, T. Margin-Based Over-Sampling Method for Learning from Imbalanced Datasets. In *Advances in Knowledge Discovery and Data Mining*; Springer: Berlin/Heidelberg, Germany, 2011; Volume 6635, pp. 309–320.
11. Breiman, L.: Bagging predictors. *Machine Learning*, 24, 123–140 (1996).
12. Schapire, R.E.: The strength of weak learnability. *Machine Learning* 5, 197–227, (1990).
13. Freund, Y. (1995): Boosting a weak learning algorithm by majority. *Information and Computation* 121, 256–285 (1995).
14. Ribeiro, M. H. D. M.; Coelho, L. S. Ensemble approach based on bagging, boosting and stacking for short-term prediction in agribusiness time series, *Appl. Soft Comput. J.*, p. 105837, (2019).
15. Salo, F., Nassif, A.B., Essex, A., (2019), Dimensionality reduction with ig-pca and ensemble classifier for network intrusion detection. *Computer Networks* 148, 164–175. doi:10.1016/j.comnet.2018.11.
16. Pullagura Indira priyadarsini, G Anuradha, "A novel ensemble modeling for Intrusion Detection system", in *International journal of electrical and computer engineering*, vol.10,no.2, (2020).
17. Tama, B.A., Comuzzi, M., Rhee, K.H., Tse-ids: A two-stage classifier ensemble for intelligent anomaly-based intrusion detection system. *IEEE Access* 7, 94497–94507(2019). doi:10.1109/ACCESS.2019.2928048.
18. H. Wang, J. Gu, and S. Wang, "An effective intrusion detection framework based on SVM with feature augmentation", *Knowledge-Based Systems*, vol. 136, pp. 130–139, (2017).
19. V. Hajisalem and S. Babaie, "A hybrid intrusion detection system based on ABC-AFS algorithm for misuse and anomaly detection", *Computer Networks*, vol. 136, pp. 37–50, (2018).
20. Zhang Y, Gong DW, Cheng J, "Multi-objective particle swarm optimization approach for cost-based feature selection in classification", *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 14(1), pp.64–75, (2017).
21. D. K. S. Kotsiantis and P. Pintelas, "Handling imbalanced datasets: A review," *GESTS International Transactions on Computer Science and Engineering*, vol. 30, no. 1, (2006).
22. M. K. H. Lee, S. Jung and S. Kim, "Synthetic minority over-sampling technique based on fuzzy c-means clustering for imbalanced data," *International Conference on Fuzzy Theory and Its Applications (iFUZZY)*, (2017).
23. J. V. G. L. R. A. C. C. Q. T. E. Y. David. S. Smith, X. Li and E. B. Welch, "Robustness of quantitative compressive sensing mri: The effect of random undersampling patterns on derived parameters for dce- and dsc-mri," *IEEE Transactions on Medical Imaging*, (2012).
24. Batista GE, Prati RC, Monard MC, A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explor Newslett.* (2004) 6(1):20–9.
25. Z. L. K. Sheng and D. Zhou, "An adaptive resampling algorithm based on cfsfdp," *International Conference on Computational Intelligence and Applications, ICCIA*, (2017).
26. L. O. H. N. V. Chawla, K. W. Bowyer and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, (2002).
27. W.-Y.W. H. Han and B.-H. Mao, "Borderline-smote: a new over-sampling method in imbalanced data sets learning," *Advances in intelligent computing*, (2005).

28. K. S. C. Bunkhumpornpat and C. Lursinsap, "Safe-level-smote: Safe-level synthetic minority over-sampling technique for handling the class imbalanced problem," *Advances in knowledge discovery and data mining*, (2009).
29. H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw., IEEE World Congr. Comput. Intell.*, Jun. (2008), pp. 1322_1328.
30. R. B. E. Ramentol, Y. Caballero and F. Herrera, "Smote-rsb: hybrid preprocessing approach based on oversampling and under sampling for high imbalanced datasets using smote and rough sets theory," *Knowledge and information systems*, vol. 33, no. 2, (2012).
31. Jiadong_Ren, Jiawei_Guo, Wang_Qian, Huang_Yuan, Xiaobing_Hao, and Hu_Jingjing, "Building an Effective Intrusion Detection System by Using Hybrid Data Optimization Based on Machine Learning Algorithms", *Security and Communication Networks Volume 2019*, Article ID 7130868, 11 pages <https://doi.org/10.1155/2019/7130868>.
32. Pullagura Indira priyadarsini, G Anuradha, "A novel ensemble modeling for Intrusion Detection system", in *International Journal of electrical and computer engineering*, vol.10,no.2, (2020).
33. Malik, A.J., Shahzad,W., Khan, F.A., Network intrusion detection using hybrid binary pso and random forests algorithm. *Security and Communication Networks* 8, 2646–2660. doi:10.1002/sec.508,(2015).
34. Pham, N.T., Foo, E., Suriadi, S., Jeffrey, H., Lahza, H.F.M., Improving performance of intrusion detection system using ensemble methods and feature selection, in: *Proceedings of the Australasian Computer Science Week Multiconference*, ACM. p. 2.doi:10.1145/3167918.3167951, (2018).
35. Abdullah, M., Balamash, A., Alshannaq, A., Almabdy, S., Enhanced Intrusion detection system using feature selection method and ensemble learning algorithms, *International Journal of Computer Science and Information Security (IJCSIS)* (2018).
36. MahbodTavallae, EbrahimBagheri, Wei Lu, and Ali A. Ghorbani "A Detailed Analysis of the KDD CUP 99 Data Set", *Proceedings of the (2009)*, IEEE Symposium on Computational Intelligence in Security and Defense Applications.
37. Sain, H., Purnami, S.W.: Combine sampling support vector machine for imbalanced data classification. *ProcediaComput. Sci.* 72(Complete), 59–66,(2015).
38. Yager R.R., On ordered weighted averaging aggregation operators in multicriteria decision making, *IEEE Transactions on Systems, Man, and Cybernetics* 18(1), (1988), pp. 183-190.
39. Anaconda software Available:<https://www.anaconda.com/download>.
40. Last, F., Douzas, G., & Bacao, F. (2017). Oversampling for Imbalanced Learning Based on K-Means and SMOTE.