Diverse Analysis of Data Mining and Machine Learning Algorithms to Secure Computer Network

Neeraj Kumar (phdcs100009.16@bitmesra.ac.in)
Birla Institute of Technology - Extension Centre Patna https://orcid.org/0000-0001-8922-4059
Upendra Kumar
Birla Institute of Technology - Extension Centre Patna

Research Article

Keywords: Intrusion Detection, Dimensionality Reduction, Clustering, Machine Learning, Data Mining

Posted Date: March 17th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-305354/v1

Version of Record: A version of this preprint was published at Wireless Personal Communications on December 2nd, 2021. See the published version at https://doi.org/10.1007/s11277-021-09393-0.

Diverse Analysis of Data Mining and Machine Learning Algorithms to Secure Computer Network

Neeraj Kumar[1], Dr. Upendra Kumar[2]
[1]Research Scholar, Computer Science & Engineering, Birla Institute of Technology, Mesra - India.
phdcs100009.16@bitmesra.ac.in
[2]Department Of Computer Science & Engineering, Birla Institute of Technology, Mesra - India.
upendrakr@bitmesra.ac.in

Abstract

Information and Communication Technologies, to a long extent, have a major influence on our social life, economy as well as on worldwide security. Holistically, computer networks embrace the Information Technology. Although the world is never free from people having malicious intents i.e. cyber criminals, network intruders etc. To counter this, Intrusion Detection System (IDS) plays a very significant role in identifying the network intrusions by performing various data analysis tasks. In order to develop robust IDS with accuracy in intrusion detection, various papers have been published over the years using different classification techniques of Data Mining (DM) and Machine Learning (ML) based hybrid approach.

The present paper is an in-depth analysis of two focal aspects of Network Intrusion Detection System that includes various pre-processing methods in the form of dimensionality reduction and an assortment of classification techniques. This paper also includes comparative algorithmic analysis of DM and ML techniques, which applied to design an intelligent IDS. An experimental comparative analysis has been carried out in support the verdicts of this work using 'Python' language on 'kddcup99' dataset as benchmark. Experimental analysis had been done in which we had found more impact on dimensionality reduction and MLP performed well in the true classification to establish secure network. The motive behind this effort is to detect different kinds of malware as early as possible with accuracy, to provide enhanced observant among various existing techniques that may help the fascinated researchers for future potential works.

Keywords: Intrusion Detection, Dimensionality Reduction, Clustering, Machine Learning, Data Mining.

1. Introduction

Due to the wide spread of Information Technology in every aspect of human life, providing protection to computers from hazards has become an essential need. The mass use of computer technologies has given rise to various vulnerabilities and threats such as Virus, Trojans, Denial of Services (DoS), ransomwares etc. Although there are many tools like antivirus software, gateway and firewalls to guard against these malicious threats. They are not sufficient enough to handle the large variety of intrusion attempts, especially when there are botnets with malicious intents. Therefore, the IT Communities urgently needs some intelligent and reliable tool or mechanism to cope with these vulnerabilities and threats. Unlike the antivirus software, firewalls and some access control schemes, IDS is an intelligent tool to detect the known attacks as well as unknown attacks. That's why IDS, using various classification techniques is the prime inspiration for this thoughtful of analysis.

Most of the researchers strive to make IDS more and more intelligent so that they could detect the new type of attacks to the maximum possible extent. Conventionally, these systems are classified into three categories namely Signature, Anomaly and a Hybrid detection system. In signature-based detection, depraved pattern of network traffic or application data has identified in order to find the malicious intents while in anomaly-based systems deviations from the good or normal behaviour is detected in order to detect the intrusions, which usually relies on Machine Learning (ML) techniques. Hybrid approaches use the benefits of both Signature and Anomaly types of detection techniques. Although each technology has its advantages and disadvantages, anomaly-based systems have some crucial benefits over other techniques. Since anomaly-based systems detect the deviation from the normal behaviour, it can easily detect the anomalous events as well as detect the attacks that are previously not known. There are lots of latest classification methods of Data Mining (DM) & ML tools available to trace out the intrusion.

Another aspect of analysis is dimensionality reduction because most of the time all the dimensions of dataset is not participated or not necessary for the desired goal. It can be achieved in two ways, one of them is feature selection method in which we select the relevant feature and feed it to the system as per the requirement and discard the remaining features. It is used for linear dataset. Feature selection can be done in three steps: 'Filter' to find out the information gain, 'Wrapping' for accuracy and 'Embedding' to determine the errors, which in turn assists in removing or adding the necessary feature. Another method of dimensionality reduction is feature extraction, where feature is selected as per inter dependencies among dimensions; it is used for non-linear dataset. It can be done in many ways. First extracting the row by determining the missing values as it decreases the accuracy. Another way is to eliminate the feature that has low variance i.e. minimum differences between features and has least impact on accuracy. Extraction may be achieve by calculating the high correlation of dimensions. If two dimensions having high correlation, in that case one of the dimensions can be extracted, which will have no impact on accuracy. In the same context, feature extraction may be attained by Principal Component Analysis (PCA) using orthogonal projection, generation of dataset with new dimensions, Backward feature elimination & Forward feature selection works on network training with least error and Random Forest (RF) to prepare the dimension as per given target. Hence, dimensionality reduction plays a significant role in sinking the space complexity.

The benefits of both the aspects mentioned above have inspired us to do collectively comprehensive analysis in a single paper. So that researchers can able to opt the best combination of classification and dimensionality reduction techniques. To achieve maximum attack detection accuracy rate along with keeping very less time and space complexity, discussions have been carried out in this paper about various classification techniques of DM and ML as well as dimensionality reduction along with all essentials associated with them are reviewed and analysed.

Structure of the paper has been organised in four sections. To get acquainted with our work a discussion has been done above in the Sect. 1, comprehensive survey on various methods used for

dimensionality reduction & classification using DM and ML techniques has been reviewed and analysed in the Sect 2. An experimental comparative analysis over various methods used for the evaluation of the performance of classifiers in Sect 3. Sect. 4 concludes the paper as an observant of comparative experimental and analytical work along with some proposals for the future work.

2. Literature Survey

2.1 Pre-Processing and Dimensionality Reduction

Pre-Processing is a process on dataset so that it can further be used for the training of different classifiers. This step involves various treatments of the dataset such as data cleaning, data integration, data transformation, data reduction, data discretization etc. Data cleaning involves filling in missing values, identify or remove outliers and resolve inconsistencies. Data integration consists of combining multiple datasets. The main purpose of pre-procession is the data reduction and is available in lightweight with the same significance. Dimensionality reduction is the process which produces the same or improved analytical results after reduced dataset in volume. The most challenging part of the dimensionality reduction is the Feature Selection or Feature Reduction. Selecting the most relevant features or reducing the dimensions improves the performance of algorithm causing the improvement in detection rate. Again these techniques can be classified as: Feature Selection (e.g. filter based, wrapper based etc.), Feature Projection (e.g. PCA, Linear Discriminant Analysis etc.) and Combinatorial Approaches, where researchers try to fuse the features from both former types. Different techniques used in pre-processing are discussed in this section.

2.1.1 Principal Component Analysis

Principal Component Analysis is a process derived from statistics that converts a set of observations of correlated variables into a set of values of linearly uncorrelated variables called principal components by using the orthogonal transformations. This is one of the most acceptable methods used for the purpose of dimensionality reduction in datasets.

Dataset – Generally dataset is the multivariate collection of data which is used for the training of model/classifiers. The most common 'kddcup99' and 'NSL-KDD' datasets are associated with the research of the Network Intrusion Detection System. Earlier the 'kddcup99' dataset was used extensively for the research but later 'NSL-KDD' dataset came into trend, which is the normalized version of the kddcup99.

Authors have been using PCA to perform their analytical experiment on kddcup99 dataset with all the 41 dimensions for dimensionality reduction by viewing 10 principal components in such a way that there is no any kind of dependencies among them. They obtained 99.7% accuracy which stayed nearly same as the original dataset due to noisy data with original 41 features using RF algorithm as classifier [1]. Researchers also employed PCA along with the Fisher Discriminant Ratio for the feature selection. For reducing the original features to a different set of reduced number of features and evaluating the performance of the classifier on those set of features in which Self-Organizing Maps acts as the main classifier which obtained around 90% accuracy with 20 number of features out of 41 kddcup99 dataset[2].

For dimensionality reduction, the authors have applied the approach of coupling the PCA with Information Gain (IG). The experimental results suggested that the ensemble classifiers were able to achieve an accuracy up to 98.24% [3]. One of the author has incorporated the PCA in a different way by extracting the hidden features that helps in exploring the fitness of encoding categorical features. It was subsequent probability of an attack trained on the feature in the context of IDS. KNN classifiers used to find the hidden features in numeric form and with 40 distinct classes of NSL-KDD dataset were able to achieve 98.05% detection rate accuracy and false alarm rate of 35%. [4].

It clears that PCA is the most reliable process for dimensionality reduction. It eliminates correlated features, enhances the performance of classifier, reduces the problem of overfitting and improves

visualization. Some precautions like normalization of data should be kept in mind before applying PCA. Otherwise it will not be able to find the optimal principal components due to which, the original features will turn into Principal Components and not be treated as original features.

2.1.2 Information and Rough Set Theory

Information-theoretic measures have also been tried and tested for the feature selection purpose. Various types of entropies such as Conditional Entropy, Relative Entropy, Correlation, Variance, Gini-Index etc., the concept of Information Gain and Correlation analysis are used to pick up the most appropriate features from a dataset having a large number of dimensions.

In association with other pre-processing technique, mutual information based feature selection algorithm was also applied to enhance the effectiveness of feature selection. These combinations were found more proficient in handling both linear as well as non-linear dataset features. A greedy feature selection method using mutual information and combination of both the feature mutual information and feature-class mutual information was used. Moreover, it was capable of framing 24 optimal subset of features which was found to be beneficial in minimizing the redundancy. It has been observed from the outcomes that the performance of the classifiers like SVM, KNN, RF, DT was improved with best computational complexity [5, 6]. Rough set theory came into existence to provide mathematical solution for imperfect knowledge about data. It has been using in wide variety of fields. Its importance in ML and AI is irreplaceable. It has been used for the purpose of feature selection from the datasets having large number of dimensions either in the stock version or with some sort of manipulation.

Authors made an effort for rapid feature selection using the rough set theory; supervised feature selection coupled with improved harmony search was used to reduce the dimensions of the dataset. The classifiers gained an accuracy of 90% after the feature selection. The time taken for the classification also got reduced [7]. An IDS using Rough Set Theory along with SVM was discussed for data pre-processing and dimensionality reduction. They trimmed down the number of dimensions from 41 to 29 after applying the rough set theory. Finally the dataset with fewer dimensions was passed to the SVM classifier. The results showed improvement in false positive rate and with 92.44% accuracy [8].

A lot of researches have been contributed so far but they were still lacking in development of adaptive model of IDS. This problem was also taken into consideration by authors in their paper and they developed an adaptive model of IDS, which is capable in detecting new attacks also.
On applying fuzzy rough set theory an optimal attribute was achieved using the information gain ratio criterion prior to the feature selection. Further global optimal Gaussian mixture model had been applied for clustering. They proposed such combination to extract the inherent structure of network instances to attain highly discernible, stable and normal intrusion pattern. NSL-KDD full dataset has 125937 records for training and 22544 records for testing. By this proposal records had found reduced to 20% i.e. 25192 records as subset of training dataset. It is more applicable for addressing the dynamic nature of network environment [9].

Through this study, we came to know about the rough set theory and its capability to reduce the original dataset into least number of sets, having similar information as the original dataset. This algorithm is more effective in discovery hidden data without having addition information about the data. Hence it is broadly adaptable.

2.1.3 Genetic Algorithms

Genetic algorithm (GA), in the field of computer science and scientific research, refers to heuristic approach, which is encouraged by the process of expected selection out of the larger class of evolutionary algorithms. They rely on bio-inspired operations such as mutation, crossover and selection to generate high-quality solutions to optimize the search problems. Recently, these algorithms were found very important place in the field of AI and ML. Researchers also used the same algorithms for the purpose of feature selection. It was used to select the feature then classification was performed with the help of DT and improvements were found in the classification attributes of the DT for developing robust IDS [10]. An improved Non-Dominated Sorting GA-III

was purpose for feature selection. This approach helped in overcoming the imbalance problem as well as removing the redundant features. Higher classification accuracy as well as the lower computational complexity was the key results of this approach [11].

Later researcher proposed Hyper-Graph-Genetic-Algorithm in support of SVM for optimization of the parameter and feature selection. Parameters play a crucial role in the performance of classifiers like SVM. Hyper-clique property was used to swift search for optimal solution and to deal with traps in the local minima. With the help of GA the value of parameter and feature subset for every chromosome were extracted by calculating their fitness and decision is taken whether that will be the part of training & testing dataset or not. This approach maximized the detection rate and minimized the false alarm rate along with the optimal number of features [12].

The researcher also made an effort to reduce misclassification by using the benefits of GA for developing a new model IDS. A paper based on the DT and GA was proposed to generate decision tree C4.5 and GA was used to overcome the problem of small disjunction in the C4.5 on kddcup99 dataset [13].

2.1.4 Other approaches

Apart from the above methods, various experiments were also carried out to find even better optimization techniques for the feature selection. In continuation to optimize feature selection method, Cuttlefish optimization algorithm was proposed which in turn found the enhancement in the performance of classifiers used for detection. Cuttlefish algorithm basically changes the pattern into different colours. It works on the light reflected and visibility mechanism for the matching pattern as subset. DT evaluated survival of this subset's feature with the help of their fitness. Experimental results of kddcup99 dataset were found that out of 41 features less than 20 features were suggested and obviously as the number of features decreased, the Detection Rate increased. The results also showed that with selected feature subsets classifiers gave higher detection accuracy rate along with a decrease in the false alarm rate [14].

The redundant and irrelevant attribute of data causes negative impact on classification performance. A paper focused on filter based feature selection method which is also one of the tried and accepted approach for selecting a subset of features from a dataset having higher dimensions. Mutual information based feature selection algorithm is used to increase the effectiveness of feature selection. This algorithm was capable of handling both the linear and non-linear data features. Further this algorithm was tested on three different datasets i.e. kddcup99, NSL-KDD and Kyoto 2006+, which accomplished classifiers with better accuracy and lower computational cost [15].

To deal with uncertainty of features, lots of papers have been contributed using Fuzzy, GA-Fuzzy, and Nero-Fuzzy as Soft Computing approaches. An IDS was designed with wrapper-based feature selection algorithm for feature selection and Neuro-Tree model as the classification engine. It gained 98.4% detection rate which was superior to the algorithms from the DT classifiers to which it compared with. Limitation of this approach was that it is effective for feature selection not for classification [20]. An intrusion detection system using genetic fuzzy rule mining approach evaluated the feature selection techniques. This was a multi-objective optimization technique. The proposed scheme was also able to use as genetic feature selection wrapper to search for an optimum feature issue. Using minimum number of features, the performance of the classifiers has been improved. The combined classification method, ant colony algorithm and SVM were used for the design of an efficient and reliable classifier for IDS. They were able to choose 19 most relevant features out of 41 features. They achieved the accuracy of 98.6249% in 10-fold cross validation [16]. SVM, Decision Tree (DT) and Simulated Annealing (SA) algorithms were also introduced in the same context like SVM and SA. This combination was used to find the most well-selected features. DT and SA were used to obtain rules for detection class identification. In addition SA adjusted the optimal parameters for DT and SVM. The proposed algorithm performed more efficiently in comparison to the compared algorithms [17].

To identify the important features out of the total available features from the dataset, another hybrid approach like Feature Vitality Based Reduction Method was proposed. The process involves the deletion of one input feature from the dataset at a time and training & testing the classifier. The

same process was continued until the performance of the classifier was improved. They used probability based Naive Bayes as the engine for the classifier. The proposed feature selection method performed better than Correlation-based: Information Gain and Gain Ratio feature selection methods. A hybrid intrusion detection system consisting of K-means, K-Nearest Neighbour (KNN), and Naive Bayes was used. Relevant features were selected using entropy-based feature selection method. K-means was used for classification, followed by the hybrid combination of KNN and Naive Bayes as classifiers. The main goal was to reduce the false alarm rate [18, 19].

A combinatory approach of Binary Gravitational Search Algorithm (BGSA) and Mutual Information (MI) was proposed termed as MI-BGSA. BSGA method was used for global search as a wrapper-based feature selection method. Finally MI was integrated to improve the BGSA and choose the most relevant features resulting in higher accuracy and detection rate in comparison to other standard wrapper-based and filter-based feature selection methods. [20]. The proposed algorithm outperformed the regular wrapper based and filter-based feature selection methods. They combined the affinity propagation with the feature clustering. 14 UCI dataset was used for the experiments to evaluate the classification accuracy and computational time. Later unsupervised feature selection method was introduced using the technique of deeming features as irrelevant, which exhibits low dependencies with the rest of the features. Experiments were carried out on several datasets. Outcomes exhibited that the proposed algorithm was able to detect insignificant features without limiting the performance of the clustering algorithm [21, 22].

Several combinational research proposals have employed in feature optimization to achieve maximum detection rate of accuracy with minimum false detection rate (FDR). It was observed that the variables that proved useful in a tree-based algorithm and other classification like SVM and DT, models are more reliable. Therefore, pre-processing of the dataset and dimensionality reduction finds a very important place in the process of intrusion detection. It not only helps classifier algorithms to perform better but also lowers the computational cost by reducing the complexities. In subsequent section, study has been done on the other important aspect of intrusion detection i.e. true classification.

2.2 Data Mining Approaches

Data Mining is a process, which has widely been used by many organizations to turn raw data into meaningful information. It is a computational process of determining pattern in big dataset, including approaches at the intersection of ML, Artificial Intelligence (AI), statistic and database system. The objective of DM is extraction of knowledge and patterns from large amount of data, not the extraction of data itself. Extracting previously unknown, fascinating pattern such as clusters of data, unusual records i.e. anomaly detection and dependencies (association rule mining) using automatic or semi-automatic analysis on big-data is the honest job of DM. The extracted data in the form of information is used for further predictive analysis, ML etc. It exists as a pool of techniques like Classification, Clustering, Anomaly detection, Association rule learning and Regression. In collaboration with DM's techniques, we can be able to reduce the dimensions of large datasets as well as attain the accuracy in classification for intrusion detection.

2.2.1 Classification Techniques

Classification is a technique where we classify data sample into given classes. The main objective of a classification issue is to identify the class under which a new data will fall. It works on the basis of supervised learning, which includes training as well as testing data samples for accuracy in predicting the correct class. It may also predict continuous data values using regression analysis. These techniques are used to forecast cluster association for data occurrence. It is a task of considering each attribute of a record and assigning that record to a particular class, which is also known as target. Several techniques have been used to find anomalies. When broadly classified they fall into three categories i.e. DM, ML and Hybrid techniques. DM techniques are used various clustering methods like Naive Bayes, Fuzzy Logic and Support Vector Machines etc. ML includes Neural Network (NN) and its variants while Hybrid methods tends to take the advantages of both DM and ML techniques.

Thus an effort has been made to classify the incoming connection as 'normal' or 'attack'. Studies about some popular works accomplished with various classifier algorithms like DT, RF, Naive Bayes, K-Nearest Neighbours etc. have been presented in the forthcoming sections.

Decision Tree - The decision tree is a decision support tool as tree-like model used to create a rule-base, which uses further to take decisions and possible outcomes. It is most applicable for classification as well as regression. It makes decision based on previously seen data and classifies the new data to a particular class. Decision is taking on each node of the tree and final class of the instance is decided at the leaf node. Numerous decision tree algorithms are available like CART (Classification And Regression Tree), ID3 (Iterative Dichotomiser), C4.5 or J48 etc. have been used to create multi-way tree, in deciding and choosing node sequence in greedy fashion to generate the rule.

A hybrid approach was proposed to deal with both types of attack detection (misuse and discrepancy) with the help of a rule-based decision support system. Where misuse-types of attacks were taking care by J48 DT and through Self-Organizing Map anomaly-type of attack has handled. Overall results of the experiment were found astonishing. The overall detection rate was 99.90% and missed rate was only 0.1 % [23]. A proposed Markov blanket model for the feature selection came into picture using Bayesian Network, CART and Ensemble methods for the classification task. They were able to choose 17 of the 41 features in the kddcup99 dataset for classification using the CART algorithm. They got the higher accuracy in CART in comparison to the Bayesian Network. Overall 93.64 % accuracy was achieved with DT after feature reduction [24]. DT works on Divide and Conquer methods therefore the major challenge is to choose the best attribute for each node during splitting step which offers the maximum information gain. GA also came into existence to solve the same for constructing the DT. During construction the features used were assigned '1' and the features which were not used were marked with '0'. Through calculating the fitness of feature attributes were selected by using the concept of genes having a certain threshold frequency selection and DT for the purpose of classification to increase the detection rate and decrease the false alarm rate [39].

There was also a proposal to improve the performance of SVM with the help of DT for feature selection. The main proposal was node information delivered by the DT to increase the performance of the SVM. The idea was simply proving node information as an additional feature along with the original features of the dataset to the SVM for further outcome. In this paper, the experimental results of DT and SVM were compared separately and a hybrid combination of DT and SVM applied as the base classifier for further processing. The results were found to be more fruitful in order to achieve the goals with enhanced accuracy in comparison to any other classifier algorithms and thus it was declared the winner of the kddcup99 contest. A feature selection method as well as intrusion detection was proposed using SVM, DT and simulated angling. Features were chosen through derived decision rules applied on training dataset. As a result of this approach misclassification rates were found very low [10] [19] [25]. Many combinatory approaches comprising of DT, SVM, Genetic and Neuro were proposed by the researchers to improve the detection accuracy and dimensional reduction. A Wrapper based algorithm for feature selection using Neuro-Tree had been proposed in order to achieve better accuracy. Although evaluation of the proposed approach with the family of other six DT (Decision Stump, C4.5, Naive Baye's Tree, RF, Random Tree and Representative Tree) classifiers were done, the proposed algorithm topped the chart of accuracy comparing various members of the DT algorithms. The proposed algorithm gained the detection rate of 98.38% and error of only 1.62%. IDS with integrated anomaly detection and misuse detection was proposed by an author, in which C4.5 was suggested to generate tree as DT to develop misuse kinds of intrusion detection model and one class SVM was used for anomaly detection. NSL KDD dataset was used for the experiment. The result showed clearly in the improvement in detection performance and detection speed in the paper [26, 27].

Yet there was an issue with accuracy of attack detection, especially for class-type attacks which have very few samples or we can say that entities are less in number in these datasets like Probe, U2R and R2L in case of kddcup99. The DT algorithm using binary split and ID3 algorithm with

quad split was used to improve the detection rate of Probe, U2R and R2L attacks as a remedy of the above issue [28].

Recently it was considered to look at a privacy property while implementing IDS in a network using DT. To achieve the same, researchers suggested to incorporate a pruning model into DT. So that privacy related concern of IDS, particularly IP addresses should not be ignored. Since IP address comes under the personal property, all sensitive information from IDS will be "pruned" and hidden. The methodology behind modified DT was that "If at the time of tree generation an IP address is chosen as the splitting attribute then it will examine and identify for similarity or dissimilarity", by comparing it with predetermined sensitive IP addresses. This paper focused on sorting sensitive IP addresses using the sorting method in combination with DT [29].

Most of the researchers aimed in reducing the features using DT, CART and C4.5 causing improvement in accuracy due to which DT is useful in real-time forecasting even for big data. It is notable from the results that the DT classifier outperformed after the feature selection. It is not dependent on any pre-processed data. The advantage is that the missing values within dataset had no effect in developing the tree but its prediction is unstable. Any minor change in the dataset may lead to inaccurate predictions. It is more expensive and prone to overfitting.

Random Forest – RF is the most popular and powerful supervised learning ML algorithm. This algorithm can be viewed as an extension of DT based on bagging algorithm. It produces the output by creating many sub data trees, after combining them, to reduce the problem of overfitting. Many researchers for the task of Intrusion Detection System have also used RF. RF can be castoff in par with algorithms like SVM. It also shows RF's ability to signify the importance of diverse nature of features, by showing different detection rates at different number of selected features. A paper used NN, SVM and RF to predict the importance of feature using rank as the basis. RF discards that irrelevant feature which has low rank, makes RF different from others existing algorithms [30, 31, 32].

Rule based IDS was restricted to the detection of known attack but was not able to detect novel attack. As a solution of the same researchers suggested the use of RF along with K-means hybrid approach to build the IDS which was able to detect both the anomaly and misuse detection. RF was used to construct a model for misuse detection whereas K-means was used to handle novel anomaly intrusion. This showed the importance of each feature of the kddcup99 dataset for implementation of IDS. Their results exhibit that the proposed approach can achieve high detection rate as well as can detect new intrusions. An IDS framework based on RF and weighted K-means, according to the results of the work, was able to achieve high detection rates for anomaly detection with 12.6% false positive rate. In contrast to the misuse detection rates which achieved lower detection rates (92.73%) but good false positive rates (0.54%) [33]. To deal with the concern of 'Class Imbalance', scholars introduced Map Reduce technique for imbalanced data using RF. Although this paper was not relevant for the case of IDS but it sheds some light on the real potential and the capabilities of the RF. Therefore this paper can be considered as an inspiration for accomplishing IDS tasks using RFs. [34]. Along with these papers many researchers have also carried out IDS either using RF alone or making it hybrid with other algorithms. Many researchers have used RFs so that they can compare the results of their own proposed algorithm with that obtained using RF.

It perceives that RF is capable of performing classification tasks for big data but it performs poorly in regression, handling missing values and maintaining accuracy. Apart from this, it was found capable to handle large dataset with higher dimensionality. It is also capable of predicting high accuracy even when a forest has a number of DTs. It can't predict beyond the range of the data in case of regression in which the problem of overfitting occurs due to noisy data. This algorithm takes much more training time in comparison to DT.

Naive Bayes – Naive Bayes is a conditional probability based model of classifier. In this model the assumption of independent forecast is almost correct. This algorithm has been used as a classifier in DM. Naive Bayes has been studied extensively since 1950's. This classifier has also been widely

used for the task of intrusion detection. Some researchers have used it alone while some have used it as hybrid model with some other classifier.

Pseudo-Bayes estimator's technique was used to improve the anomaly detection system's capability to detect new attacks and reduce the false alarm rate as much as possible. DT was compared with Naive Bayes using the same kddcup99 dataset. Although the quality of classification was almost the same in both classifications, it was found that Naive Bayes was about 8 times faster than DT [35]. A hybrid approach towards IDS using K-means Clustering and Naive Bayes was used to carry out the experiment using kddcup99 dataset. Initially, using K-means, attacks and normal instances were separated into different classes. Later, Naive Bayes was used to classify the attacks into further more categories. This approach achieved very low false alarm rate and higher accuracy than carrying out the whole experiment on K-means or Naive Bayes alone. Vitality Based Reduction method choose the most relevant features out of the dataset and put the reduced dataset to Naive Bayes for the classification [21]. In addition a combination of Entropy Minimization Discretization (EMD) and Proportional K-Interval Discretization (PKID) were used for the feature selection and Hidden Naive Bayes for the classification task. It was found that Hidden Naive Bayes performed very well having accuracy of 99.96% as compared to other six classifiers used in the experiment. Moreover like other described classifiers this classifier also has been used extensively, either for classification using itself or for the purpose of comparison with other classifiers [36]. An advanced Naive Bayesian classifier as Relief algorithm was introduced, in which, assign weights to each attribute of the dataset to maintain a relationship between attributes intended for better classification results. The proposed classifier showed better true positive rates and lower false alarm rates during detection [37].

From above Naive Bayes algorithm, it has been learnt that it improves the accuracy in less amount of training time by introducing only sample set of predictor. On the other hand, in this algorithm it is not so easy to obtain a completely independent predictor set that is not mutually independent between attributes.

2.2.2 Clustering Techniques - Clustering in DM is the task of grouping the similar objects into clusters, so that the object in one cluster are more similar to each other than those objects which lies in another clusters. There are several clustering techniques present under DM. One of the most widely used clustering technique, in the task of IDS, is K-means. In order to build an IDS initial the dataset was normalized and then the single-linkage clustering technique used for the clustering process. Further labelling of clusters had been prepared. This system was able to detect large number of intrusions while keeping the false positive rate reasonably low [38]. A clustering heuristic for the intrusion detection called it 'Y-means' came into existence. It was based on 'K-means' to overcome the problem of number of clusters dependency and degeneracy. Evaluation of the work was done on the kddcup99 dataset. Ultimately 82.32% detection rate and 2.5% false alarm rate were attained [39].

A genetic clustering method for the intrusion detection was proposed, which was able to create clusters that classifies the intruders as 'normal' or 'abnormal' automatically. In the first stage, they did the clustering and in the second stage, they applied genetic optimization to obtain the nearest optimal detection result. The overall detection rate including all the classes of attacks was about 61%. However, the experiment was feasible and effective for the intrusion detection [40]. Several contributions having combined approach using K-means and NN were proposed for intrusion detection. Where K-means has been used to automatically select an optimal set of samples and then the outcome passes to the NN. The experiment was shown in a paper, they found improvements in terms of time complexity and performance of NNs, as the addition of clustering before feeding the dataset to NNs. Hierarchical clustering was used to speed up the training process of the SVM, in which the experiment was carried out on kddcup99 dataset [41]. The experiment also revealed improvements in training time as well as accuracy of classification and detection rate.

A hybrid scheme was proposed to drop the false alarm rate of the proposed discrepancy IDS by pooling the K-means with KNN and Naive Bayes in which feature selection was performed using an entropy-based feature selection method to select relevant features and discard irrelevant ones. Firstly, the feature was selected using the foretold feature selection algorithm. Secondly, the clustering was

done using K-means and finally they performed classification using hybrid approach of k-neighbour and Naive Bayes. Results witnessed the improvement in the detection rate as well as reduction in the false positive rate. Detection rate increased to 99.35% using this method [42].

Later, another K-means and C4.5 algorithm based hybrid approach was proposed for network anomaly detection. In this hybrid approach K-means was used to partition the dataset into the cluster for training and testing using Euclidian distance then rules were generated using C4.5 DT for classification. In the performance measure chart of the experiment, it was intelligently surpass majority of the classifiers with precision of 95.6% and accuracy of 95.8 %. In another proposal a classifier based on Cluster Centre And Nearest Neighbour (CANN) was proposed. In this proposal two distances of each data sample from its cluster centre and neighbouring data sample were measured and added. This CANN was applied by the researcher in KNN classifier. The results obtained using kddcup99 dataset clear the effective improvement in the training and testing time as well as accuracy in comparison to KNN and SVM. Another researcher proposed a new ensemble method of construction using PSO generated weights. Local unimodal sampling (LUS) used as meta-optimizer. The experiments performed on kddcup99 dataset showed that it could generate better ensembles, which outperform normal ensemble methods [43, 44].

K-means simplifies clusters of different sizes, shapes and is easy to implement for clustering. To determine the initial value 'k' in the K-means algorithm, it may run several times but the problem arises for outlier cases. It does not perform well with large dimension, so it would be preferable to have dimensionality reduction using PCA or any other modified clustering algorithms.

2.3 Machine Learning Approaches

ML is one of the subset of AI which uses the statistical techniques and gives the computer's ability to learn from data itself instead of being explicitly programmed. ML was extensively used for IDS tasks. Generally, the classification task in IDS is accomplish using the ML techniques. Sometimes, they used ML along with other DM classifiers with an aim to improve the performance or efficiency of the existing classifiers [45]. ANN (Artificial Neural Network) is one of the most widely used algorithms for the task of classification and has been used heavily in IDS. Tons of experiments were carried out using ML techniques since ML is also one of the hottest and most emerging topics of the last decade and has witnessed many developments.

Support Vector Machine (SVM) – SVM splits the dataset into different classes by determining the centre point decision boundary, which is closer to the opponent class also known as hyperplane. For linear separation via SVM, the hyperplane that has the maximum marginal width of the decision boundary is chosen to avoid misclassification. Therefore, MMH (Maximal Margin Hyperplane) is the key factor for classification with maximum accuracy. For Non Linear data classification, Kernel is the key factor which accepts the LD (Low Dimensional) feature space and gives an output with HD (High Dimensional) feature space. It is supervised learning model, so labelled dataset is used for training and has been used heavily in last decade for the misuse type of intrusion detection. SVM is used for classification as well as in regression too.

As far as SVM classification is concerned, researchers have contributed plenty of papers. Initially papers were just to get acquainted with the principle of the SVM as it is capable to classify the data-sample into two or multi-classes. Later, SVM was utilized to develop intrusion detection system using DARPA 1998 dataset with NN. It was found that the training time for SVM was less than NN. It was also found that SVM has a slightly higher rate of correct detection as compared to NN when used for the same task. However, only the binary classification was possible using SVM, which was a disadvantage for that time. A classification task was performed using SVM and detection rate was used as evaluation criteria. Overall result was better than the winner of the contest of kddcup99. Detection accuracy was found in case of 'Normal' 99.3%, 'DoS' 91.6%, but lacking in 'Probe' 36.65, 'U2R' 12%, 'R2L' 22%. Poor performance in 'Probe', 'U2R' and 'R2L' was due to less no. of training sample [46, 47, 48].

The proposed Rough Set and SVM combined approaches were not only restricted to feature selection but suitable for classification also. It was found that more accurate results were obtained after feature selection in terms of training and testing time. With the help of the Rough Set, only 29 out of 41 features of kddcup99 were selected, trained and classified using SVM. As a result, they achieved accuracy ranging from 86.79% to 89.13% with 29 features compared to 41 features [12]. Some other combinatory approaches for feature removal method called 'gradually feature removal method was suggested. In this paper they used Ant Colony algorithm and SVM. Evaluation were done onto 10-fold cross validation to train the network. After reducing feature into each round, dimensions reduced from 41 to 4 and accuracy achieved was 98.6249% [18].

A Multi-Levelled-Hybrid intrusion detection model was suggested and tested on 10% kdd dataset. For pre-processing, initially symbolic attributes 'protocol_type', 'service'and'flag' were converted into numeric form. A modified version of K-means algorithm was also presented so that a high-quality training dataset can be used. Subsequently they applied modified K-means on each sub-category (Labelled attack: Normal, DoS, Probe, R2L, U2R) and generated new different number of attributes, which were chosen for different targets of every category. Number of instances of before and after applying modified K-means on 10% KDD dataset in this manner was Normal (97278 to 639), DoS(391458 to 140), Probe(4107 to 134), R2L(1126 to 51), U2R(52 to 25). Further classification had been performed with this newly generated dataset using SVM and ELM (Extreme Learning Machine). Thereafter testing was done with multi-level model on corrected dataset of kddcup99. The comparative result was found using basic K-means with Modified K-means, with these metric of evaluation: Accuracy (91.88% to 95.75%), DR (92.13to 95.17) and FAR (9.16 to 1.87) [49].

A proposal came into existence in the direction of impulsive protection in the wireless sensor networks to provide secure communication between two sensor nodes. In which clustering had been performed by sampling on the node's weight. Then cluster head execute adaptive chicken swarm optimization algorithm. Through this adaptive nature, it reduces the time taken for the best cluster head selection. The higher degree of representation in comparison to other sampling techniques makes it superior to other prevailing techniques. In another method, Rotated Random Forest (RRF) had been applied to reduce the features within dataset. For classification, SVM & ML two-phase classification approach was applied to reduce the dataset. In the first phase, sensor node predicts about the intrusion in the binary classification and in the second phase malicious sensor nodes were found capable in predicting about their types. In this paper, researchers concluded that the RRF performed feature selection with higher detection accuracy and comparatively less time in comparison to normal RF. After that, by taking the advantage of this reduced feature the SVM was used as classifier resulting in improved accuracy of above 90% for DoS, Normal and Probe types of attack for detection of kddcup99 dataset with minimal FDR. At the same time, computational cost using these approaches were observed to be higher and it also performed poor in detecting classes like R2L and U2R of kdcup99 dataset as less number of samples was there to train. An author presented a combinatory approach of SVM-GA-ANN paper to deal with this problem [50].

The ML based approach was offered to overcome the problem of less no. of sample data of any feature to train any system. This paper describes a new hybrid approach for feature selection and attack detection. Wrapper method, the GA with multi-parent crossover and multi-parent mutation (MGA) with SVM, namely MGA-SVM, were used for feature selection. A hybrid gravity search (HGS), a particle swarm optimization (PSO) and ANN approaches in conjunction with MGA-SVM-SVM-HGS-PSO-ANN were used to train the classifier. Performance was compared with other standard methods like Chi-SVM, gradient descent (GD-ANN) and DT, GA-ANN and PSO (GSPSO-ANN) using only 4 features out of 43 features of NSL-KDD dataset and maximum detection accuracy of 99.3% was attained [51].

From this analysis clears that most of the researchers adopted SVM to improve detection efficiency. At the same time disadvantage associated with SVM was that training & testing time and detection accuracy performance decreases for big-data. Hence it is recommendable for high dimensional space but would not be suggested for big-data without dimensionality reduction.

Neural Network– Neural Network (NN) is parallel computing device which works as human brain and can take decision with fast computation. In order to better identify intrusions, it can be effectively applied to deal with the upstairs issues. NN has just been utilized to take care of numerous issues related to pattern recognition, DM and complete AI.

The challenge is so far an identification of new types of intrusion for which there is no prior knowledge. An NN modelled IDS was introduced to overcome this challenge along with very less FDR. Through the experiment, they obtained 96% DR with 7% FDR. However, this paper proved to be a source of inspiration for many future work in the field of IDS, which identifies the problem of high false alarm rates and low detection rates of new attacks using NN [52].To deal with the new attack, they had suggested a combination of discriminatory training and general keywords approach to minimize such problems. New keywords were added, which detect actions that were common to many attacks and used simple NN discriminant training to produce output. The improved system received a detection rate of approximately 80% per day at a low false alarm rate of approximately one false alarm [53].

It was noticed that the sample, which was small in number, had a lower intrusion detection rate over any supervised learning classifier. They had less participation in training so clearly perform poorer during testing. Many researchers proposed their issues when a small number of input feature were served for training. A contribution to deal with this issue, they had proposed a model of SVM-NN approach on kdcup99 dataset. They were able to make 23 class feature identification using NN and SVM for the cost effective and real time Intrusion Detection System. They applied the techniques of deleting one feature at a time and perform experiment then rank the importance of that input feature. This process was repeated for all features individually and a set of features was obtained based on their rank. As an effect of the above dimensional reduction procedure, the results were mostly notable in terms of training time [54, 55]. A paper on intrusion detection using hierarchical NN-based IDs was proposed to identify misuse attacks as well as anomaly attacks correctly & adaptively. Also parallel hierarchical IDS was introduced to enhance the performance of serial hierarchical IDS. Through experiments, they achieved 89% detection rate and 1.6% false positive rate. The paper also concludes how parallel hierarchical IDS was superior to a serial hierarchical NN that uses the hierarchical identity model PCA-NN. They attained PCA-NN based IDS suitable for adaptive online computing for misuse detection and anomaly detection. The paper shows ameliorated results over existing similar works of that time using its proposed method. However, in this paper they did not consider a class whose samples were less in numbers and dealt with anomaly type of attacks only [56]. In sight of above concerns, a researcher proposed a hybrid system called Artificial Immune System. In that paper Kohen Self Organizing Map (SOM) method adopted for the network intrusion detection. This approach found capable of handling both anomaly as well as misuse detection attack. Anomalous network connections were initially detected using an artificial immune system. Initially it was suggested to flag anomalous connections for categorisation using SOMs. Later those features were removed which have higher-level information in the form of cluster. These experiments were carried out on the well-known kddcup99 dataset. Hence, it is observed that their experimental results had improved in comparison to other results obtained on similar tasks at the time [57].

Later on hybrid soft computing Fuzzy-ANN approach was recommended to deal with mainly two issues, one of them was lower detection precision for low frequent attacks and other was lack in detection accuracy. This hybrid model uses the three-stages classification. Firstly, they generated different training subsets using fuzzy classification technique. Then different ANNs were trained in the second stage and subsequently in the last stage a meta-learner and fuzzy aggregation module was introduced to learn again and combine the different ANN's results. The concept behind this divide and conquer approach was adopted. The results were in fact productive and appeared in the context of the proposal. The system also improved the detection rate of less frequent attacks such as R2L and U2R of the kddcup99 dataset [58]. In this context another proposal was also made to simplify the same target using mutual information based feature selection method, coupled with multilayer NN. In experimental assessment they had compared it with Multi-Layer-Perceptron (MLP) and Radial

Basis Function networks. The accuracy of the proposed model was observed to be better as compared to the existing proposals [59, 60].

Through the ML based analysis we perceived that Anomaly-based search is an identification technique whereby IDS looks for vulnerabilities based on user-defined rules, not based on signatures already stored in IDS. This type of identification typically uses AI to distinguish between normal traffic and anomalous traffic.

2.4 Hybrid Approaches

These approaches comprise of combined algorithms of DM or ML techniques to create or to implement a new algorithm from an existing algorithm. The goal is to improve either the performance or the efficiency of existing algorithms or to create a new algorithm to complete a task. In their quest to create the most efficient ID, researchers performed many tasks by combining different algorithms of DM or ML. Some of the works were described in the above sections of this chapter. During the period 2000 to 2007 the study focused on single and hybrid classifiers. After analyzing many papers, it was suggested to adopt ML to develop robust and computationally profitable IDS [61].

A research team proposed a novel intelligent hybrid approach to classify useful and useless features meeting the feature reduction through the first ranking method. This was achieved by combining both ranks which were derived from information gain and correlation. These reduced features were then fed to forward NNs, where training as well as testing was performed on the kddcup99 dataset. Later they tested on five different test datasets and compared the suggested approach to the lack of with and without feature selection using different evaluation metrics [62]. Furthermore, this review approach was extended using C4.5, Naive Bayes and RF classification algorithms. The review researcher focused on cases of oversampling and undersampling features like U2R, R2L of the kddcup99 dataset. Because one class on oversampling can weaken the performance of another class. They concluded that sampling the class was more effective in comparison to monitor the appropriate events for the U2R and R2L range. For solution U2R and probe orbit attacks, they found the Naive Bayes classification to be comparatively appropriate [63].

Hybrid approaches were found to be more suitable for detecting both misuse and mismatch attacks. Misuse Intrusion Detection (MID) is a static type of approach that typically deals with well-known attacks using a set of rules. MID discovers a known attack with false positives. MID related issues are not useful in detecting new attacks. Whereas Anomaly Intrusion Detection (AID) is a dynamic approach which can detect unusual activity on the network. For such attacks first we need to know about the normal traffic to detect any discrepancy. It is capable of responding the new attacks due to dynamical change in network traffic other than normal traffic but at the same time caution should be taken as all abnormal traffic is not malicious. Therefore in case of AID we have to handle more alarms that are positive. While blocking attempts that matches the rule we can be more specific with MID and it is better to use alerts in AID. Therefore researchers can decode the right approach that should be taken as per their requirements. Most of the researchers preferred hybrid approach to develop an IDS model because the amalgamation of these techniques can do a superior job by countering their own flaws.

3. Comparative Discussion and Experimental Analysis

The aforesaid survey draws a lot of conclusions. Fig.1 shows the year-wise distribution of review articles since 1992. It also reflects the trends in approaches used by researchers for single or hybrid solutions to implement robust IDS for intrusion detection. From Fig.1, we have seen that hybrid approaches were becoming popular with researchers proposing work for the past decade. This journey of algorithmic analysis yielded better results for handling both types of attack.

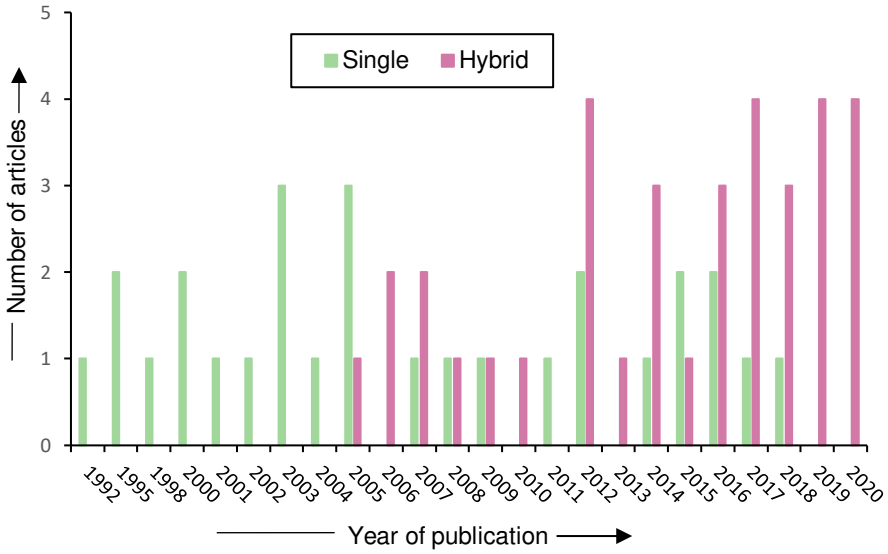

Fig. 1. Year Wise Distribution of Articles for the types of Classifier Design.

Observations:

The above survey on dimensional reduction for designing efficient IDS using various techniques for classification concludes that dimensional reduction plays a significant role in designing any IDS. It was found that reducing the dimensions, especially in case of huge dimensions, the dataset is complex and time consuming. Following are the major observations that were found as a research gap :

(i) For different classifier, dissimilar set of features was chosen for the same dataset, which infer that dimension reduction is classifier dependent.

(ii) Different features were used for the detection of different attacks within a classifier.

Evaluation Metrics:

- **Classification accuracy** – Classification accuracy is the ratio of number of correct predictions made by the classifier to the total number of samples given as input to the classifier. Mathematically, it is given by

$$\text{Accuracy} = \frac{\text{Number of correct predictions made by classifier}}{\text{Total number of samples as input to the classifier}}$$

- **Confusion Matrix:** It is a matrix which gives the complete performance of the model. One can infer basically the several important terms from the confusion matrix:
- **True Positive:** True positive is used to indicate that the classifier has detected an attack precisely.
- **True Negative:** True negative is used to indicate that the classifier has not made a mistake in detecting a normal condition.
- **False Positive:** False positive is used to indicate those attacks, which have not actually occurred but detected by the classifier.
- **False Negative:** False negative is used to indicate that even a particular attack has occurred; the classifier was not able to detect the intrusion.
- **Precision:** Precision is the ratio of number of true positives (TP) to the sum of number of true positives & the number of false positives (FP). Precision expresses

the relevancy of the data points which were actually relevant. Mathematically, it is given as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall:** Recall is the number of true-positives divided by the sum of number of true-positives (TP) and number of false-negatives (FN). In simple words, it is the capability of the classifier to find all the positive samples. Mathematically, it is given as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F-score:** F-score is simply the weighted harmonic mean of the precision and recall. Value of 1 is considered as best and value of 0 as worst for F-beta score. Mathematically, it is given by

$$\text{F} - \text{score} = \frac{2 * Recall * Precision}{Recall + Precision}$$

Moreover, the overall accuracy has also been calculated. The ratio of the total number of samples detected correctly to the total number of samples that was present in dataset had been considered as the overall accuracy of the classifier.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

It is simply a fraction having the sum of true-positives with true-negatives as numerator and total number of samples as denominator.

Comparative Experiments Performance:

To rationalize the knowledge acquire through survey, we performed an experiment using evaluation metrics on various DM and ML based popular classifiers like DT, MLP, RF, K-means, KNN and SVM. Experiment is carried out using 'Python' language. Dataset kddcup99 was opted as benchmark, where out of 41 features only 5 feature (duration, protocol type, flag, diff_srv_rate and dst_host_rerror_rate) were selected. Table 1 shows the 'detection accuracy' obtained through different classifiers, before and after feature reduction.

Table 1. Comparison of Accuracy Detection (%) before and after Feature Selection

Accuracy on Classifiers	Decision Tree	MLP	Random Forest	K-Means	KNN	SVM
Before feature selection	85.009	94.93	96.805	1.472	91.425	0
After feature selection	95.53	93.277	92.868	96.7	96.9	92.62

Graphical representation of the above result has been exhibited in Fig. 2 which shows the comparison of Accuracy Detection Ratio (%) before and after feature selection among the classifiers.

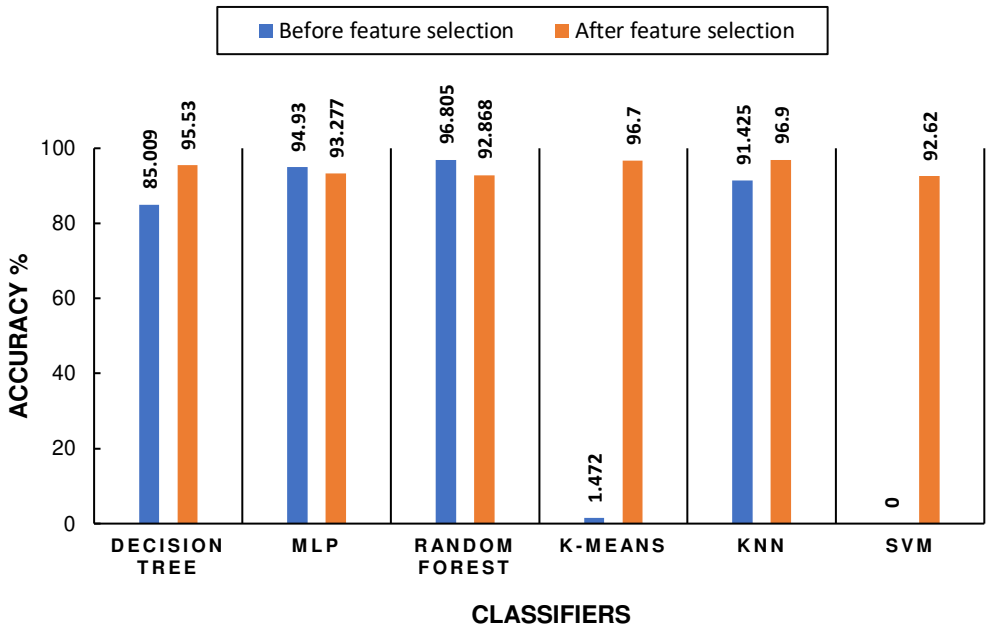

Fig. 2. Comparison of Accuracy Detection Ratio (%) before and after Feature Selection

Results:

The above experimental results shown in Table 1 and Fig. 2, following information inferred:

- The effect of dimensionality reduction on DT & KNN and for true classification enhanced MLP results were obtained. At the same time in case of RF and MLP the result was at par.
- In the case of classifiers like K-Means and SVM, rate of intrusion detection was found to be significantly improved. Furthermore, we observed that SVM do not performed well due to the large dataset prior to feature selection.

It is notable that the AI-based hybrid approach is more efficient in feature selection as well as in classification, which increased the detection accuracy. Therefore this analytical work has been recommended as an intelligent intrusion computing technique.

4. Conclusion and Future Work

The paper provides a comprehensive comparative analysis of diverse algorithmic for the IDS and feature reduction using assorted DM and ML hybrid techniques. Unlike other analysis, this work is not constrained to any fussy aspect of IDS such as algorithms used for classifier or dimensionality reduction techniques rather had been compared to other research papers and meticulous discussions have been done on various methods of pre-processing of dataset, dimensionality reduction as well as different algorithms which is used as classifiers appraisal. Experimental Results revealed that how the trend in the usage of various algorithms for the design of IDS has shifted from the pure DT to ML and other hybrid techniques. It has been also observed that dimensionality reduction is directly proportional to time complexity and further accuracy in true classification. Hybrid solution had been found efficient to detect known and new kinds of attacks. Through the comparative analysis and simulation's outcome concludes that ML as well as hybrid approaches play a vital role in design of robust IDS to secure computer network.

However, ML approaches have proposed sensitive IDS which regularly decreases the false alarm rate. Still numerous challenges are in the view of enhancing the performance of classifying

algorithms. It has been observed that a set of different features was chosen for different classifier and to detect different attacks within classifier. Hence, choosing suitable features for different classifier affect the computational cost. Further work may be extended to develop a method in which the same feature set could be selected not only for different classifiers but for different types of attacks also within a classifier for the same dataset. Ancillary it had been perceived that real time intrusion detection with stumpy cost would be a desire for IDS to secure computer network.

References

[1] K. K. Vasan and B. Surendiran, "Dimensionality reduction using Principal Component Analysis for network intrusion detection," *Perspectives in Science*, vol. 8, pp. 510-512, 2016.

[2] F. Kuang, S. Z. Jin and W. Xu, "A novel SVM by combining kernel principal component analysis and mproved chaotic particle swarm optimization for intrusion detection," *Soft Computing*, vol. 19, pp. 1187-1199, 2015.

[3] F. Salo, A. B. Nassif and A. Essex, "Dimensionality reduction with IG-PCA and ensemble classifier for network intrusion detection," *Computer Networks*, 148, pp. 164-175, 2019.

[4] S. Nerella and M. Shashi, "Encoding Approach for Intrusion Detection Using PCA and KNN Classifier," *Proceedings of the Third International Conference on Computational Intelligence and Informatics*, Springer, Singapore, pp. 187-199, 2020.

[5] F. Amiri, M. R. Yousefi, C. Lucas, A. Shakery and N. Yazdani, "Mutual information-based feature selection for intrusion detection systems," *Journal of Network and Computer Applications*, vol. 34, pp. 1184-1199, 2011.

[6] N. Hoque, D. K. Bhattacharyya and J. K. Kalita, "MIFS-ND: A mutual information-based feature selection method," *Expert Systems with Applications*, vol. 41, no. 14, pp. 6371-6385, 2014.

[7] H. H. Inbarani, M. Bagyamathi and A. T. Azar, "A novel hybrid feature selection method based on rough set and improved harmony search," *Neural Computing and Applications*, vol. 26, no. 8, pp. 1859-1880, 2015.

[8] R. C. Chen, K. F. Che, Y. H. Chen and C. F. Hsieh, "Using rough set and support vector machine for network intrusion detection system," *First Asian Conference on Intelligent Information and Database Systems*, IEEE, pp. 465-470, 2009.

[9] L. Jinping, W. Zhang, Z. Tang, Y. Xie, T. Ma, J. Zhang, G. Zhang and J. P. Niyoyita, "Adaptive intrusion detection via GA-GOGMM-based pattern learning with fuzzy rough set-based attribute selection," *Expert Systems with Applications* 139, p.112845, 2020.

[10] G. Stein, B. Chen, A. S. Wu and K. A. Hua, "Decision tree classifier for network intrusion detection with GA-based feature selection," *Proceedings of the 43rd annual southeast regional conference*, ACM, vol. 2, pp. 136-141, 2005.

[11] Y. Zhu, J. Liang, J. Chen and Z. Ming, "An improved NSGA-III algorithm for feature selection used in intrusion detection*," Knowledge-Based Systems*, vol. 116, pp. 74-85, 2017.

[12] M. G Raman, N. Somu, K. Kirthivasan, R. Liscano and V. S. Sriram, "An efficient intrusion detection system based on hypergraph-Genetic algorithm for parameter optimization and feature selection in support vector machine," *Knowledge-Based Systems*, vol. 134, pp. 1-12, 2017.

[13] C. Azad and V. K. Jha, "Decision Tree and Genetic Algorithm Based Intrusion Detection System," *In Proceeding of the Second International Conference on Microelectronics, Computing & Communication Systems (MCCS 2017)*, Springer, Singapore, pp. 141-152, 2019.

[14] A. S. Eesa, Z. Orman and A. M. A. Brifcani, "A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems," *Expert Systems with Applications*, vol. 42, no. 5, pp. 2670-2679, 2015.

[15] A. Ambusaidi, X. He, P. Nanda and Z. Tan, "Building an intrusion detection system using a filter-based feature selection algorithm," IEEE transactions on computers, vol. 65, no. 10, pp. 2986-2998, 2016.

[16] J. Xia, S. Zhang, J. Yan, X. Ai and K. Dia, "An efficient intrusion detection system based on support vector machines and gradually feature removal method," *Expert Systems with Applications*, vol. 39, no. 1, pp. 424-430, 2012.

[17] S. W. Lin, K. C. Ying, C. Y. Lee and Z. J. Lee, "An intelligent algorithm with feature selection and decision rules applied to anomaly intrusion detection," *Applied Soft Computing*, vol.12(10), pp. 3285-

3290, 2012.
[18] S. Mukherjee and N. Sharma, "Intrusion detection using naive Bayes classifier with feature reduction," *Procedia Technology*, vol. 4, pp. 119-128, 2012.
[19] K. Siddique, Z. Akhtar, M. A. Khan, Y.H. Jung and Y. Kim, "Developing an Intrusion Detection Framework for High-Speed Big Data Networks: A Comprehensive Approach," *KSII Transactions on Internet & Information Systems*, vol.12(8), pp. 4021-4037, 2018.
[20] H. Bostani and M. Sheikhan, "Hybrid of binary gravitational search algorithm and mutual information for feature selection in intrusion detection systems," *Soft computing*, vol. 21, no. 9, pp. 2307-2324, 2017.
[21] Y. Zhang Y. Li, T. Zhang, P. K. Gadosey and Z. Liu, "Feature clustering dimensionality reduction based on affinity propagation," *Intelligent Data Analysis*, vol. 22, no. 2, pp. 309-323, 2018.
[22] L. Talavera, "Dependency-based feature selection for clustering symbolic data," *Intelligent Data Analysis*, vol. 4, no. 1, pp. 19-28, 2000.
[23] O. Depren, M. Topallar, E. Anarim and M. K. Ciliz, "An intelligent intrusion detection system (IDS) for anomaly and misuse detection in computer networks," *Expert systems with Applications*, vol. 29, no. 4, pp. 713-722, 2005.
[24] S. Chebrolu, A. Abraham and J. P. Thomas, "Feature deduction and ensemble design of intrusion detection systems," *Computers & Security*, vol. 24, no. 4, pp. 295-307, 2005.
[25] S. Peddabachigari, A. Abraham, C. Grosan and J. Thomas, "Modeling intrusion detection system using hybrid intelligent systems," *Journal of Network and Computer Applications*, vol. 30, no. 1, pp. 114-132, 2007.
[26] S. S. S. Sindhu, S. K. Geetha and A. Kannan, "Decision tree based light weight intrusion detection using a wrapper approach," *Expert Systems with Applications*, vol. 39, no. 1, pp. 129-141, 2012.
[27] G. Kim, S. Lee and S. Kim, "A novel hybrid intrusion detection method integrating anomaly detection with misuse detection," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1690-1700, 2014.
[28] S. Puthran and K. Shah, "Intrusion detection using improved decision tree algorithm with binary and quad split," *In International Symposium on Security in Computing and Communication, Springer, Singapore*, pp. 427-438, September 2016.
[29] Y. J. Chew, S. Y. Ooi, K. S. Wong and Y. H. Pang, "Decision Tree with Sensitive Pruning in Network-based Intrusion Detection System," *In Computational Science and Technology,* pp.1-10, Springer, Singapore, 2020.
[30] Ho Tin Kam, "Random decision forests," Document analysis and recognition, 1995, *proceedings of the third international conference*, IEEE, vol. 1, pp. 278-282, 1995.
[31] Kim Dong Seong, Sang Min Lee and Jong Sou Park, "Building lightweight intrusion detection system based on random forest," *International Symposium on Neural Networks*, Springer, Berlin, Heidelberg, pp. 224-230, 2006.
[32] Zhang Jiong and Mohammad Zulkernine, "A hybrid network intrusion detection technique using random forests," *In First International Conference on Availability, Reliability and Security (ARES'06)*, IEEE, pp. 8-pp, 2006.
[33] R. M. Elbasiony, E. A. Sallam, T. E. Eltobely, and M. M. Fahmy, "A hybrid network intrusion detection framework based on random forests and weighted k-means," *Ain Shams Engineering Journal*, vol. 4, no. 4, pp. 753-762, 2013.
[34] S. Del Río, V. López, J. M. Benítez and F. Herrera, "On the use of MapReduce for imbalanced big data using Random Forest," *Information Sciences*, vol. 285, pp. 112-137, 2014.
[35] D. Barbara, W. Ningning and J. Sushil, "Detecting novel network intrusions using bayes estimators," *Proceedings of the 2001 SIAM International Conference on Data Mining*, Society for Industrial and Applied Mathematics, pp. 1-17, 2001.
[36] Koc, Levent, Thomas A. Mazzuchi and ShahramSarkani, "A network intrusion detection system based on a Hidden Naive Bayes multiclass classifier," *Expert Systems with Applications*, vol. 39, no. 18, pp. 13492-13500, 2012.
[37] Y. Wang, Li Yuzhou, T. Daxin, C. Wang, W. Wenyng, H. Rong, G. Peng and Z. Haijun, "A novel intrusion detection system based on advanced naive Bayesian classification," *In International Conference on 5G for Future Wireless Networks*, Springer, Cham, pp. 581-588, April 2017.
[38] Altman, S. Naomi, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175-185, 1992.
[39] Y. Guan, A. Ali, Ghorbani and B. Nabil, "Y-means: A clustering method for intrusion detection." *Electrical and Computer Engineering*, IEEE CCECE 2003, Canadian Conference, vol. 2. pp. 1083-1086, 2003.
[40] Y. Liu, K. Chen, X. Liao and W. Zhang, "A genetic clustering method for intrusion detection," *Pattern*

Recognition, vol. 37, no. 5, pp. 927-942, 2004.
[41] L. Khan, A. Mamoun and T. Bhavani, "A new intrusion detection system using support vector machines and hierarchical clustering," *The VLDB journal*, vol. 16, no. 4, pp. 507-521, 2007.
[42] H. Om and A. Kundu, "A hybrid system for reducing the false alarm rate of anomaly intrusion detection system," *Recent Advances in Information Technology (RAIT)*, 1st International Conference on IEEE, pp. 131-136, 2012.
[43] Muniyandi, A. Prabakar, R. Rajeswari and R. Rajaram, "Network anomaly detection by cascading k-Means clustering and C4. 5 decision tree algorithm," *Procedia Engineering*, vol. 30, pp. 174-182, 2012.
[44] W. C. Lin, S. W. Ke and C. F. Tsai, "CANN: An intrusion detection system based on combining cluster centers and nearest neighbors," *Knowledge-based systems*, vol. 78, pp. 13-21, 2015.
[45] Nguyen, T. T. Thuy and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," *IEEE Communications Surveys & Tutorials*, vol. 10, no. 4, pp. 56-76, 2008.
[46] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.
[47] S. Mukkamala, G. Janoski and A. H. Sung, "Intrusion detection using neural networks and support vector machines," *IEEE International Joint Conference on Neural Networks*, IEEE Computer Society Press, pp. 1702-1707, 2002.
[48] Kim, Dong Seong, and Jong Sou Park, "Network-based intrusion detection with support vector machines," *International Conference on Information Networking*, Springer, Berlin, Heidelberg, pp. 747-756, 2003.
[49] W. A. A. Yaseen, Z. A. Othmanand and M. Z. A. Nazri, "Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system," *Expert Systems with Applications*, vol. 67, pp. 296-303, 2017.
[50] G. M. Borkar, L. H. Patil, Dilip. Dalgade and A. Hutke, "A novel clustering approach and adaptive SVM classifier for intrusion detection in WSN: a data mining concept," *Sustainable Computing: Informatics and Systems*, pp.120-135, 2019.
[51] S. Hosseini and B. M. H. Zade, "New Hybrid Method for Attack Detection Using Combination of Evolutionary Algorithms, SVM, and ANN," *Computer Networks*: 107168, Elsevier, 2020.
[52] J. Ryan, M. J. Lin and R. Miikkulainen, "Intrusion detection with neural networks," *Advances in neural information processing systems*, pp. 943-949, 1998.
[53] R. P. Lippmann and R. K. Cunningham, "Improving intrusion detection performance using keyword selection and neural networks," *Computer Networks*, vol. 34, no. 4, pp. 597-603, 2000.
[54] Andrew H. Sung, and Srinivas Mukkamala, "Identifying important features for intrusion detection using support vector machines and neural networks," *Applications and the Internet, 2003. Proceedings. 2003 Symposium on IEEE*, pp. 209-216, 2003.
[55] Chunlin Zhang, Ju Jiang and Mohamed Kamel, "Intrusion detection using hierarchical neural networks," *Pattern Recognition Letters*, vol. 26, no. 6, pp. 779-791, 2005.
[56] Liu Guisong, Yi Zhang and Shangming Yang, "A hierarchical intrusion detection model based on the PCA neural networks," *Neurocomputing*, vol. 70, pp. 1561-1568, 2007.
[57] Simon. T. Powers and Jun He., "A hybrid artificial immune system and Self Organising Map for network intrusion detection," *Information Sciences*, vol. 178, no. 15, pp. 3024-3042, 2008.
[58] G. Wang, J. Hao, J. Ma, and L. Huang, "A new approach to intrusion detection using Artificial Neural Networks and fuzzy clustering," *Expert Systems with Applications*, vol. 37, no. 9, pp. 6225-6232, 2010.
[59] V. M. Reddy, I. R. P. Reddy and K. A. N. Reddy, "Mutual Information-Based Intrusion Detection System Using Multilayer Neural Network," *In First International Conference on Artificial Intelligence and Cognitive Computing Springer*, Singapore, pp. 529-537, 2019.
[60] Sireesha Rodda, "Network Intrusion Detection Systems Using Neural Networks," *Information Systems Design and Intelligent Applications*, Springer, Singapore, pp. 903-908, 2018.
[61] C. F. Tsai, Y. F. Hsu, C.Y. Lin and W.Y. Lin, "Intrusion detection by machine learning: A review," *Expert Systems with Applications*, vol. 36, no. 10, pp. 11994-12000, 2009.
[62] M. Ishfaq, N. Kumar, "A feature reduced intrusion detection system using ANN classifier," *Expert Systems with Applications*, vol. 88, pp. 249-257, 2017.
[63] C. Trupti, S. Shukla and R. Wadhvani, "An analysis of "A feature reduced intrusion detection system using ANN classifier" by Akashdeep et al. expert systems with applications (2017)." *Expert Systems with Applications*, 130, pp. 79-83, 2019.

Figures

Figure 1

Year Wise Distribution of Articles for the types of Classifier Design.

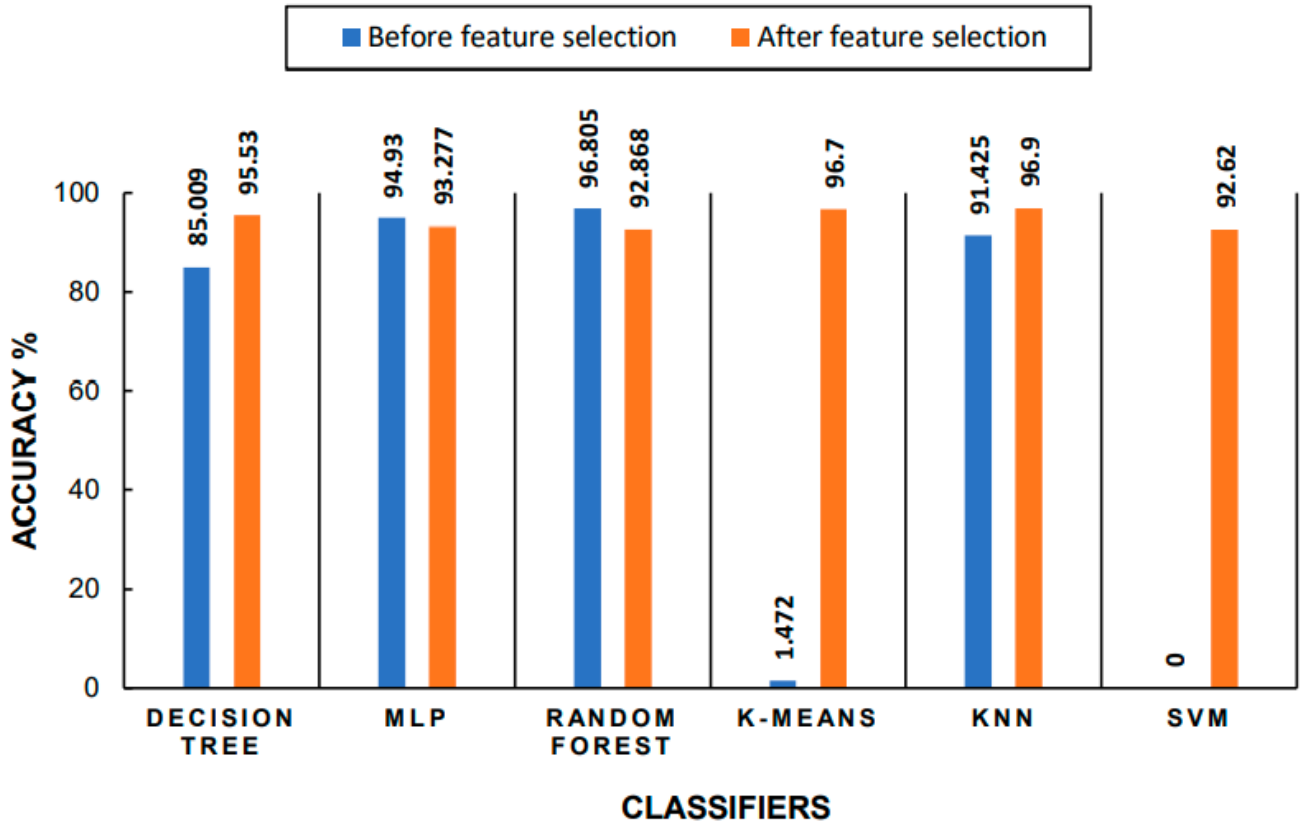

Figure 2

Comparison of Accuracy Detection Ratio (%) before and after Feature Selection