# Intrusion-Miner: A Hybrid Classifier for Intrusion Detection using Data Mining

Samra Zafar[1]
School of Electronic Information and
Electrical Engineering
Dalian University of Technology
Dalian 116024, China

Muhammad Kamran[2]
College of Computer Science and
Engineering
University of Jeddah, Jeddah,
Saudi Arabia

Xiaopeng Hu[3]
School of Electronic Information and
Electrical Engineering
Dalian University of Technology
Dalian 116024, China

***Abstract*—With the rapid growth and usage of internet, number of network attacks have increase dramatically within the past few years. The problem facing in nowadays is to observe these attacks efficiently for security concerns because of the value of data. Consequently, it is important to monitor and handle these attacks and intrusion detection system (IDS) has potentially diagnostic ability to handle these attacks to secure the network. Numerous intrusion detection approaches are presented but the main hindrance is their performance which can be improved by increasing detection rate as well as decreasing false positive rates. Optimizing the performance of IDS is very serious issue and challenging fact that gets more attention from the research community. In this paper, we proposed a hybrid classification approach 'Intrusion-Miner' with the help of two classifier algorithm for network anomaly detection to get optimum result and make it possible to detect network attacks. Thus, principal component analysis (PCA) and Fisher Discriminant Ratio (FDR) have been implemented for the feature selection and noise removal. This hybrid approach is compared with J48, Bayesnet, JRip, SMO, IBK and evaluate the performance using KDD99 dataset. Experimental result revealed that the precision of the proposed approach is measured as 96.1 % with low false positive and high false negative rate as compare to other state-of-the-art algorithm. The simulation result evaluation shows that perceptible progress and real-time intrusion detection can be attained as we apply the suggested models to identify diverse kinds of network attacks.**

***Keywords*—Intrusion detection system; principal component analysis; intrusion-minor; fisher discriminant ratio***

## I. Introduction

The networked computer systems are playing a progressive vital role in our society with hastily increasing adoption of internet. Although, internet brings enormous advantages, it also has increased threats of computer systems connected to the internet becoming target of intrusions by cyber criminals [1, 2]. However, it is impossible to have a safety mechanism without susceptibility. Consequently, they are inadequate to make the infrastructure absolutely secure due to careless layout, malicious attacks and implementation flaws continuously try to escapade system's weaknesses. It is important to monitor and identify these attacks so, it will become traditional to invent a security mechanism. This security mechanism is known as intrusion detection and it is considered as a crucial part of the present security approaches.

An IDS helps in keeping a track of malicious attacks or breaking of the policy of a system or a network and reports to control room by generating alarms. Fig. 1.clearly illustrates the whole process. Intrusion detection are basically divided into two design approaches, misuse and anomaly detection system based on detection philosophy [1, 3, 4]. For a misuse IDS approach, information gathered from traffic analyzed by the IDS to compare it to large database having signatures of already known attacks. These signature attacks are documented by human experts. It is not effective for unknown and novel attacks for which the signature are not yet available. On the other hand in anomaly detection approach system administrator defined the baseline, breakdown, normal state of traffic load, typical packet size and protocol in advance. Network segments are monitored by a detector to stack up against the normal baseline state and inspect for deviations. It can detect potentially a wide range of novel attacks [5].

An IDS monitors all ingoing and outgoing activities of network. They manage this by collecting information from a number of systems and network resources. It identifies attacks, probes, exploits and other vulnerabilities of the network analyzing the heaped information. An IDS respond to malicious attacks in one of the various ways, for example by generating alarm, creating the event or paging an administrator. An IDS may comprise of software and hardware equipment and sensor devices. These devices can be implemented anywhere in a network. These IDS can be implemented using data target, response and data mining techniques based on IDS.

These are four types of system attacks on network:

- Denial of Service attack (DoS): In these attacks, the attacker prevents a valid user to get access by blocking him. For this, the attacker tries to occupy the resources of the computer system in such a way that they become busy.
- Users-to-Root attack (U2R): In such attacks, the attacker tries to exploit the system weaknesses by locking up a legitimate user and accessing root component of the system. Few examples of U2R attacks are 'buffer overflow', 'load-module', 'perl', and 'rootkit'.
- Remote-to-Local attack (R2L): In these attacks, vulnerabilities of a machine allow an attacker to get an

access locally a legitimate user account without having his (or her) own account. A few examples of R2L attacks include ‘phf’, ‘ftp write’, ‘warezmaster’, ‘warezclient’, ‘spy’, ‘imap’, ‘multihop’, and ‘guess passwd’.

- Probing attack (PROBE): These attacks involve the bypassing of security by the attacker and collecting the data from the nodes in the network. Few example like ‘portsweep’, ‘ipsweep’, and ‘nmap’ are a few examples of PROBE attacks

Over the deployment of data mining methodologies, systematic IDS are developed to detect intrusion excellently and perform generalizations. Therefore, the installation and implementation of such kind of systems can be obviously complicated. The systems’ integral problems can be categorized into discrete problem sets based on proficiency, precision, and availability parameters [5, 7, 8]. Though, data mining techniques IDS generally designed for anomaly detection methodologies that have higher false positive occurrence as compared to other detection techniques that only focus on handcrafted signature. Therefore, previous techniques face difficulty during processing of data, online intrusions detection and require huge amount of data as compare to current methodologies [6, 27, 30].

Hence, constructing the proficient intrusion detection system is dynamic defense in the network system’s and it make it possible to detect network attacks. So, a hybrid classification approach Intrusion-Miner proposed to get optimum results. Then, to find best performance yielding classifiers, we will evaluate our proposed classifier intrusion miner on KDD99 dataset. We also evaluate the time taken by the algorithm for training of all the classifiers. Finally, a proposed hybrid approach compared with previous classifiers in-term of TP, FP and average accuracy.

In this paper the work has been organized as follows. In Section II, we discuss about the related work of current study and Section III contains the overview of the proposed methodology. This section provides proposed scheme with detail of its phases and general form of proposed model. In Section IV, we provide the detailed analysis including the result and discussion of relative performance. Finally, in Section V, we conclude the paper and show possible future work.

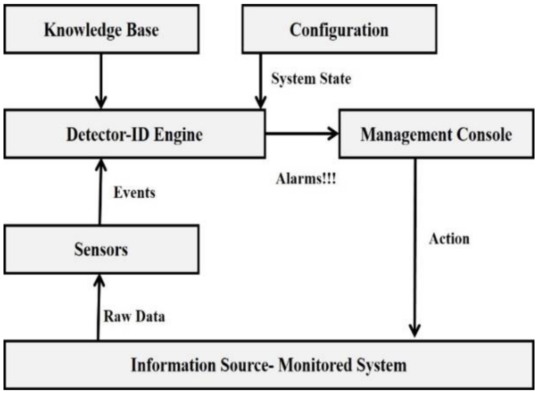

Fig. 1. Intrusion Detection Model.

## II. Related Work

The most primitive study regarding to intrusion detection system was first proposed in (1980) [9]. The approach was based on statistic method to analyze and monitor those attackers who get into the system in illegitimate manner. Consequently, the work is led by Dorothy E. Denning [10] in (1987) who build the first prototype of intrusion detection expert system (IDES). In this work, they executed a dual approach that utilized a rule based expert system as well as statistical variance components that has its basis on host system. Subsequently, in (1995) they lead their work and build a new version namely next generation intrusion detection expert system [11]. The description of inclusion of host IDS in to the network IDS namely a hybrid IDS model was proposed in [12]. This new hybrid model contain both misuse and anomaly modes. Then data mining techniques are applied on these features to learn rules that precisely define the actions of intrusions and normal activities. Both known and unknown intrusions can be detected efficiently by using this hybrid IDS [13]. The hybrid approach integrates self-organizing maps and statistical methods to detect the network anomalies proposed in [14]. Feature are selected and noise removal is achieved on the basis of FDR and PCA. In 2015, the progress of this hybrid methodology by combining two data mining approaches implemented in [7]. A novel K-means clustering algorithm is employed to reduce the number of features related with each a data point. Sunil Nilkanth Pawar [15] suggested a genetic algorithm technique that is based on chromosomes having variable length to build network IDS. For the generation of rules chromosome having relevant features were utilized. Fitness of each function is defined by an effective fitness function. Every single chromosome represents one or more than one rules for efficient anomalies detection. The efficiency of the proposed technique is proved by testing it on DARAPA 1998 dataset.

Consequently, to deal with the IDS efficiently a hybrid model namely SVM model that interconnect kernel PCA (KPCA) with upgraded chaotic intelligence scheme namely particle swarm optimization (PSO) proposed in [16]. Preprocessing on support vector machine (SVM) is functionalized by KPCA scheme to increase the SVM performance and decrease the training time. The extracted results shows higher accuracy and precision and shorten the training time. Iftikhar Ahmad [17] exploit PCA to select feature subset in his proposed approach. PCA is conducted on the basis of highest eigenvalues. Rather than using a conventional methodology to select the structures with their highest eigenvalues such as PCA, A genetic principal mechanism are employed to select subsets of features and for sorting purpose SVM is used. The obtained results indicates that it increase the detection rates and decreases the number of features. Chun Guo [18] presented a novel and tractable hybrid learning method for building effective IDS. The suggested model is known as Distance Sum-Based SVM (DSSVM). In this method, the distance sum and the cluster centers are defined along with interconnection among each data samples. Experimental results obtained from this model clearly showed that this hybrid approach shows better results in terms of intrusion detection and computational cost.

Saurabh Mukherjee [19] introduced a technique which utilizes a method based on feature vitality such as correlation based as well as gain ratio and information gain to recognize the anomalies in the selection system. However, the effective classifier namely naïve Bayes also implemented on intrusion detection system. The evaluated results showed batter performance of IDS. One of the most common and powerful data mining algorithms called K-Means clustering with the conjunction of Naïve Bayes classification for IDS recommended in [20]. As compare to the separate Naïve Bayes classification, this advance application offers high-reaching detection rate. However, the limitation of this approach is that more false positives are generated.

Zhi [21], came up with a newly proposed model for intrusion detection model which combines two classifiers C4.5 and hybrid neural network. As network attacks are classified into four categories neural network perform well in detecting Probing and DOS attacks whereas, R2L and U2R attacks are detected more accurately with the help of C4.5 classifier. Muniyandi [22] presented a novel hybrid approach which combines C4.5 and k-Means classifiers. The presented hybrid technique provide anomaly detection by cascading the C4.5 decision tree and k-Mean clustering methods. Simulation results show that the proposed technique gives impressive detection rate. Mrutyunjaya Panda [23] implemented Naïve Bayes grouping technique in his work to solve the issue of IDS. He worked for anomaly based network intrusion detection using KDD99 dataset. He also performed a comparison of back propagation neural network based approach with the adopted technique and results clearly showed that the suggested technique accomplishes better in terms of TP rate, TT and cost. Yang Li and Li Guo [24] proposed a network intrusion detection technique dependent on Transductive Confidence Machines for K-Nearest Neighbors (TCM-KNN) , by adopting this technique the anomalies can be identify efficiently with high detection rate, less false positive conditions by utilizing fewer nominated data and its features. The results of average TP and FP have in good agreement with values 99.6% and 0.1 % respectively. A method in which SVM used to categorize different types of attack proposed in [25]. This proposed method shows higher accuracy result with RBF and the accuracy value is 98.57%, for the NSL-KDD data set. Dhanabal [26] Analyze NSL-KDD data set and applied on SVM, J48, and Naïve Bayes for classifying attacks. In regarding to this some of experimental result demonstrates that CFS can be used for dimensionally reduction and in this case J48 classifier classifies the data with better accuracy From the literature review it is perceived that some algorithms performs well for a certain attack category while fails for others.

Therefore, we can expect much performance improvement from a multiple classifier selection model instead of using a single classifier in solitary. We take the advantage of information gain to address the data handling issue. Moreover, we used the hybrid of probabilistic algorithm to design the final architecture of our classifier. We evaluate classifiers on KDD99 data set to find optimum performance.

## III. Proposed Approach

This section provides the details about the proposed scheme. The proposed scheme mainly consists of three phase namely: (i) Feature Selection; (ii) Fisher Discriminant Ratio for Eigenvectors; and (iii) Classification. The main architecture of the proposed scheme has been shown as a 2-step engineering approach in Fig. 2 and top level architecture Fig. 3 respectively.

### A. *Feature Selection*

The first step of the proposed approach is to select the features from the input dataset. This step is important because it involves to identifying those features of the data that may trigger an alarm when an intrusion is suspected. Moreover, it also involves excluding those features from the classification step that do not play any vital role in the classification process. Furthermore, the redundant and irrelevant features are also filtered out; as a result, overall computation time of the algorithm is reduced along with the improvement in the classification results in terms of classification accuracy and generalization. The generalization is an important property of classification as it helps the algorithm to avoid over-fitting on a particular data.

The feature selection method is divided into three types that are wrapper, filter and hybrid methods. In filter method, a preprocessing step is performed to select a subset of features on the bases of selected criteria. In this preprocessing step, features are selected without considering their performance of the classifier. In this way, filter method are considered as less time consuming as compare to the wrapper method because; wrapper method evaluate the feature selection method on the bases of the outcomes of the classifiers. Even though wrapper method perform well as compare to filter method in terms of classifier accuracy, but when the classifier is changed the results obtained become not applicable in the same situation. To overcome these limitations of above two mentioned methods, a new approach was proposed that is called hybrid method. This method combines both filter method and wrapper method to support the classifier. This hybrid approach is used in our work with the filter method. In many applications PCA has been used to extract the most relevant information from dataset. It has been already successfully used in applications based on face recognition techniques [28]. In our proposed technique, a unique class of uncorrelated features is derived from a class of correlated features. Hence, PCA reproduce a class of orthogonal basis vectors to express the data as a linear combination of that basis. This method involves some classification task problems when new data is added because it takes more time in processing; similarly, it also lacks the desired property of invariant under a transformation of data.

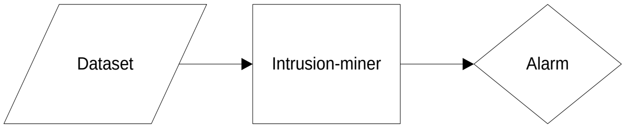

Fig. 2. A Two-Step Engineering Approach of Intrusion-Miner.

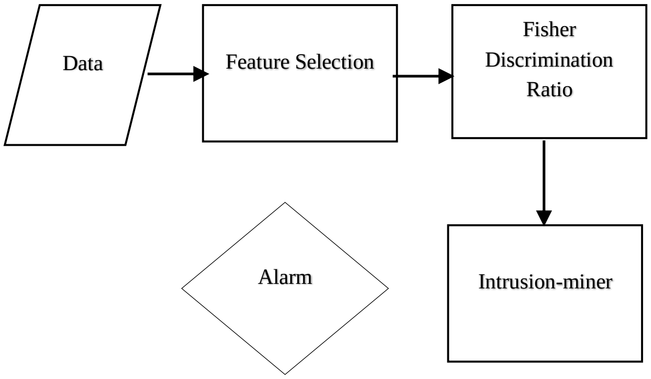

Fig. 3. Top Level Architecture of Intrusion-Miner.

To formally understand this process, consider the following modeling.

Suppose we have $X = \{x_{1,}x_{2,}x_{3,} \dots, x_N\}$ samples (or records) in the training data. By subtracting it from the mean $\overline{X}$, we obtain the shifted data manifold as $Y = \bar{X} - X$ for $y_i \in R^n$ with $y_i = (y_{i^1,} y_{i^2,\dots,} y_{i^n})^T$ and $i = 1,2,\dots,N_i$. The job of PCA is to search for orthonormal vectors $u_k = (u_{k^1}, u_{k^2}, \dots, u_{k^n})$ for $k = 1,2,\dots,N_t$ such that

$$\lambda_k = \sum_{r=1}^{N_t} (u_k^T \, y_r)^2 \qquad (1)$$

The goal is to maximize $\lambda_k$. The vectors $u_k$ verify that $u_l^T \, u_k = \delta_{lk}$, where $\delta_{lk}$ represents the kronecker data. Moreover, the scalar $\lambda_k$ and the vectors $u_k$ denote the eigenvalues and eigenvectors respectively and they are used to compute the covariance matrix as $C = YY^T$.

In the proposed Intrusion-miner, the role of the eigenvectors is to form a new feature space for removing the noise and also to reduce the features set. In this way, we project the samples in the training data by utilizing the space as defined by the eigenvectors. As a result, uncorrelated features are generated in the form of a set that can describe the data manifold. The classifiers are then run on these features to generate final result. Before a classifying a test sample $v$ (a data sample from the testing data), the sample needs to be projected onto the space spanned by the eigenvectors. This into the corresponding feature vector represented as:

$$\omega_k = (v - \bar{X}) * u_k \qquad (2)$$

In order to regenerate the original data from the principal components, $\bar{X}$ is used as:

$$v_k^{rec} = \bar{X} + \omega_k + u_k^T \qquad (3)$$

The above equation shows how that how the eigenvector $k$ is used to follow the reconstruction process $u_k^{rec}$ for the sample $v$. for ensuring the maximum use discriminative power of the projections, we propose to compute and use the FDR value as follows.

### B. Fisher Discriminant Ratio (FDR)

The data preprocessing step is of utmost importance in the classification of real-world datasets because they usually have some noise, missing values, invalid values, redundant values, and irrelevant features. To solve the problems faced by PCA regarding the selection of eigenvectors with higher values and yet missing the most discriminative ones, we propose to use FDR as:

$$FDR = \sum_{i=1}^{M} \sum_{j\neq 1}^{M} \frac{(\mu_i - \mu_j)}{(\sigma_i^2 + \sigma_j^2)} \qquad (4)$$

Where, $\mu_i$ and $\sigma_i$ represent the mean and variance for the class i respectively.

**Algorithm 1.** The main FDR algorithm follows.

- Compute $Y = X - \bar{X}$
- Compute the projection s of $X$ to corresponding eigenvectors
- Sort the eigenvectors based on their discriminative power computed using Eq. (4)
- For eigenvectors with lower values FDR values, subtract the projection of training samples from the corresponding eigenvectors.
- The main advantage of using FDR with PCA is that it allows using k most discriminative eigenvectors to make the classification task more efficient.

### C. Classification Module

Due to imbalanced nature of the intrusion datasets, a classifier or classification algorithm needs to exploit the local data distribution for making the decisions during classification. The instance based leaner k-NN is one such algorithm, so we propose to use this algorithm along with some desired characteristics of using the global data model for classification. Consequently, both the local and the global properties of the data can be used to classify the data efficiently. We used the hybrid of probabilistic algorithm BayesNet with Instance based learner (IBk) to design the final architecture of our classifier–IntrusionMiner. Moreover, the performance of various other classifiers like J48, BayesNet, JRip, SMO, and IBk are also the part of our study for performance comparison.

### D. General Form of Proposed Model

In order to find optimum performance yielding classifiers, we evaluate six classifiers on KDD99 dataset. Parameters selected for the performance comparison are FP and TP rate. These parameters could be considered the best point of comparison for classifying the algorithms. Moreover, it is important to record the overall average accuracy (AA).Similarly, the average Training Time (TT) of each algorithm also plays an important role for real world problems. An algorithm should be selected for building the final model if it performs well on all the attack categories. On the basis of performance, the proposed model will use one best classifier to detect network anomaly of each attack category.

## IV. Experiment and Results

This section describes the experiments and their results using the proposed model for effective intrusion detection. This section is further divided into sub-sections as follows.

### A. *Data Set Description*

The dataset used for this experiment was taken form [29] that was also used in KDD-99 dataset. The original dataset contains about 4,900,000 unique connections. Every connection vector contains 41 features. From these 41 features 7 features are discrete by nature and remaining 34 are continuous. The network activities are labeled as not normal or 'attacks' considering the normal network behavior.In our experiments, the following four types of attacks were simulated DoS, U2R, R2L and Probe. In our experiments on the KDD-99 dataset, we take the protocols like TCP, UDP, and ICMP into account. The actual training dataset used in our research work is made up of 494,021 records. Among which 97,277 (19.69%) were normal, 4,107 (0.83%) Probe, 391,458 (79.24%) DoS, 52 (0.01%) are U2R and 1,126 (0.23%) R2L.In the dataset there are 41 attributes associated with each connection and each attribute describes variant features of the connection. Each connection is differentiated by a label (attack type or normal) that is allocated to it.

The imbalance nature of the dataset is presented using Fig. 4. It is quite evident from Fig. 4 that the data is highly imbalanced as there is a huge difference in the number of records for each class. For brevity, in our experiments, we selected the 10% of the samples present in the KDD training dataset. Thus, we selected 9841 records from the 'Normal' class, 39072 records from the 'DoS' class, 437 records from 'Probe' class, 13 records having class of 'U2R', and 213 records form 'R2L' class making a total of 49,596 (that is 10% of total records).

### B. *Evaluation Setup*

The experiments were performed on a system having a processor of Intel(R) Core(TM) i3 with a 4GB RAM running Microsoft Windows 8.1 Pro. An open source machine learning package Weka was used for simulation. The version of Weka tool used is 3.6.0 which is the Windows version. We used Weka is in our research work because it provides numerous data mining and machine learning algorithms. It provides the facility of data preprocessing, clustering, classification, visualization, regression and association rules. However, only a subset of classifiers algorithms is exploited in our proposed work. The classification techniques mentioned in our classification module in Section 3.3 were used through Weka so that all the results can be compared to the performance of the proposed Intrusion-miner.

### C. *Data Pre-Processing*

The actual KDD99 dataset contains large number of records as mentioned above. In our experiments, we divided the dataset into two subsets. The first subset which is training set contains 49,596 instances in total and consists of 9,841 normal instances, 13 U2R instances, 39,092 DoS instances, 213 R2L instances, and 437 Probe instances. In the second subset, we separated 15,437 instances the act as an absolute testing set. For preprocessing the selected datasets, we followed the approach mentioned in Section 3.1 and 3.2 respectively. After the preprocessing step we can effectively evaluate the performance of selected classifiers by running them on these subsets of the dataset.

### D. *Classification and Performance Comparison*

Evaluation of data mining classifiers having best performing instances mentioned in Section 3.3 was done on KDD99 dataset. Fig. 5(a-d) shows the simulation results of these classifiers. Each classifier was compared on the basis of the parameters like TP-Rate (correctly identified positive cases), FP-Rate (negative cases that have been incorrectly classified as positive), and AA (total correctly classified instances divided by the total number of instances) as shown in Eq. (5), (6) and Eq. (7) and the time taken by the algorithm for training.

$$\text{TP}=\frac{d}{c+d} \quad (5)$$

$$\text{FP}=\frac{b}{a+b} \quad (6)$$

$$\text{AA}=\frac{a+d}{a+b+c+d} \quad (7)$$

Where a is correct predictions when precedent is negative, b is the number of incorrect prediction when precedent is positive, c is number of incorrect predictions when precedent is negative and while d is the incorrect predictions when precedent is positive.

Table I compares the results of the classifiers using the parameters AA, TP-Rate and FP-Rate for various attacks (classes). It is quite evident from Table I that the proposed Intrusion-Miner has better results as compared to other state-of-the-art classifiers. Also note that in most of the cases the Bayesian classifier BayesNet and the Instance based learner IBk performed better than J48, JRip, and SMO.

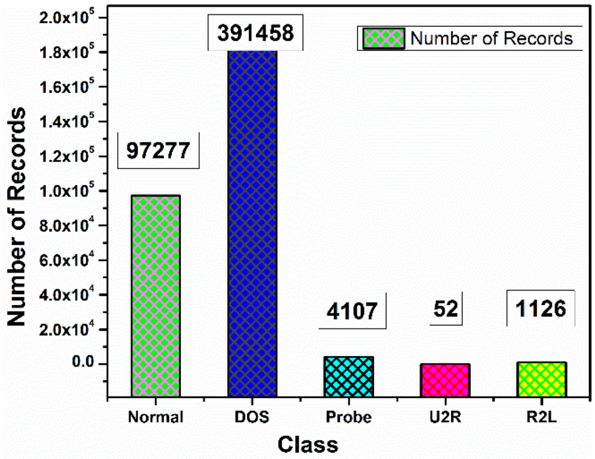

Fig. 4. The Number of Records for Each Class in the Dataset Depicting the Imbalanced Nature of the Dataset.

This is because of the nature of the dataset being imbalanced. Furthermore, the Intrusion-miner has even better results than BayesNet and IBk alone. We believe this is due to the fact that it exploits the properties of probabilistic nature of Bayesian classifier and learning capabilities of instance based classifiers and then combines both of these desired characteristics of the classifiers to achieve better classification results in terms of TP-Rate, FP-Rate, and AA. We also give the comparison of these algorithms in terms of time taken to build the classification model using the training set in Fig. 6. It is clear that the proposed Intrusion-Miner has also better speed than some of its counterpart. Although, some of the algorithms like BayesNet, J48, and IBk built the model even faster; however, the better accuracy of Intrusion-Miner as evident from Table I is enough to neglect this minor difference of time.

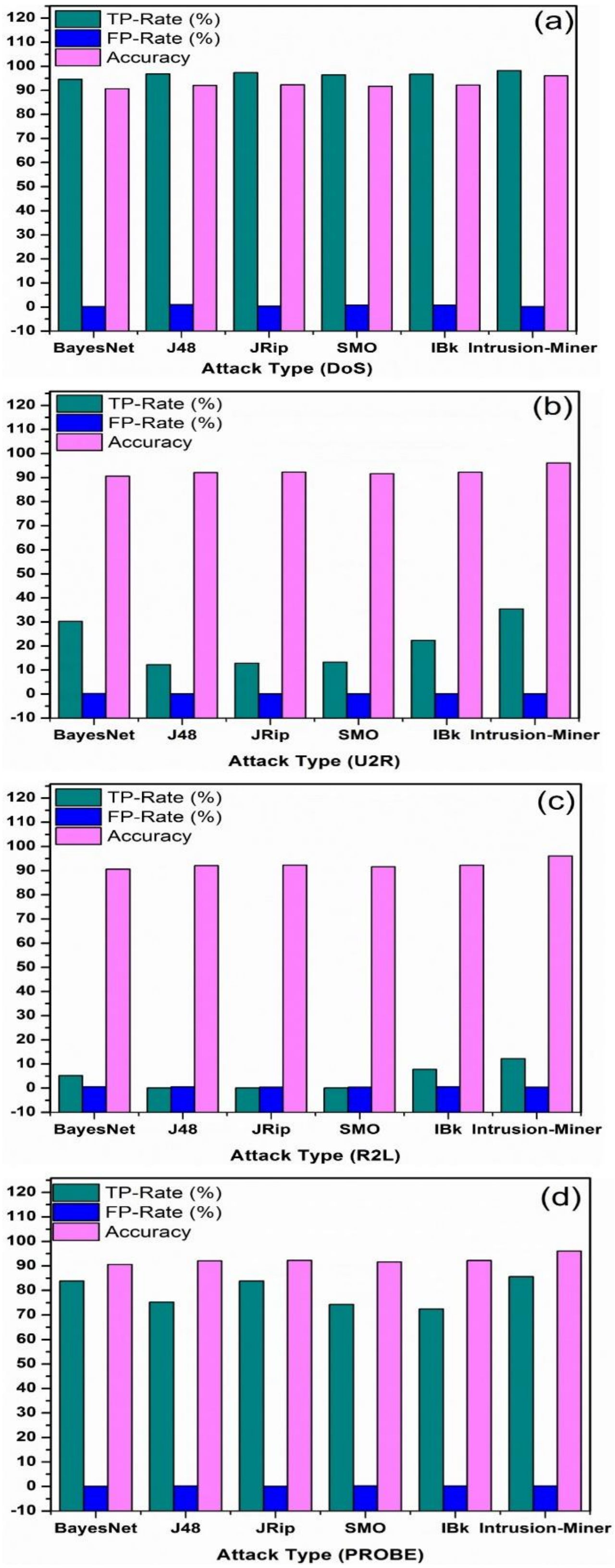

Fig. 5. Classification of Results as a Comparison of State-of-the-Art Classifiers with the Proposed Intrusion-Miner in Term of FP, TP and AA are Shown in (a) DoS, (b) U2R, (c) R2L, (d) Probe Respectively.

TABLE I. A COMPARISON OF STATE-OF-THE-ART CLASSIFIERS WITH THE PROPOSED INTRUSION-MINER

| Attack Type | Algorithm | TP-Rate (%) | FP-Rate (%) | Accuracy |
|---|---|---|---|---|
| DOS | BayesNet | 94.6 | 0.2 | 90.62 |
| | J48 | 96.8 | 1 | 92.06 |
| | JRip | 97.4 | 0.3 | 92.3 |
| | SMO | 96.4 | 0.8 | 91.65 |
| | IBk | 96.7 | 0.8 | 92.22 |
| | **Intrusion-Miner** | **98.2** | **0.2** | **96.1** |
| U2R | BayesNet | 30.3 | 0.3 | 90.62 |
| | J48 | 12.2 | 0.1 | 92.06 |
| | JRip | 12.8 | 0.1 | 92.3 |
| | SMO | 13.3 | 0.1 | 91.65 |
| | IBk | 22.3 | 0.1 | 92.22 |
| | **Intrusion-Miner** | **35.4** | **0.1** | **96.1** |
| R2L | BayesNet | 5.2 | 0.6 | 90.62 |
| | J48 | 0.1 | 0.5 | 92.06 |
| | JRip | 0.1 | 0.4 | 92.3 |
| | SMO | 0.1 | 0.4 | 91.65 |
| | IBk | 7.8 | 0.6 | 92.22 |
| | **Intrusion-Miner** | **12.2** | **0.4** | **96.1** |
| PROBE | BayesNet | 83.8 | 0.13 | 90.62 |
| | J48 | 75.2 | 0.2 | 92.06 |
| | JRip | 83.8 | 0.1 | 92.3 |
| | SMO | 74.3 | 0.3 | 91.65 |
| | IBk | 72.4 | 0.2 | 92.22 |
| | **Intrusion-Miner** | **85.6** | **0.2** | **96.1** |

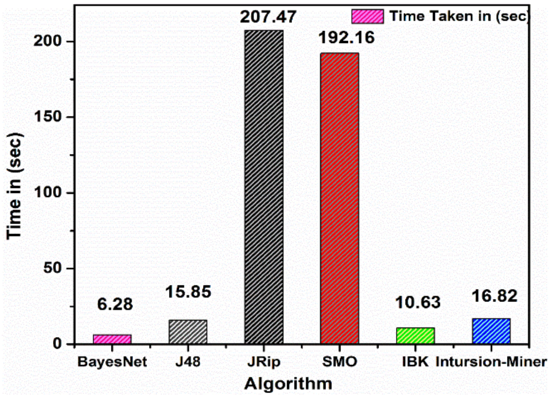

Fig. 6. Time Taken (in seconds) by Various Algorithms to Build the Model.

### E. Discussion on Results

It is a general accepted fact that the accuracy of the algorithm, the overall TP-Rate, and FP-Rate are among the most important parameters for measuring the performance of a classification algorithm. However, when dealing with highly imbalanced dataset like the one used in KDD-99, it is also required to note the performance of the classifier for each class individually as well. According to our hypothesis, it is obvious from Table I that it is not possible to detect all attack categories with a single algorithm giving low false alarm rate and high probability of detection. It gives us an idea that for different attack categories different algorithms could be used. Simulation results shown in Table I clearly depict that some algorithms show better performance towards a specific attack category. For instance, most of the algorithms produce significant TP rates–like 95% for DoS category. While for U2R and R2L type of attacks, the accuracy of all the classifiers is significantly lower than other classes of attack. This is because the classifiers generally tend to learn using the majority results; consequently, the records having class having a very small number are often miss-classified. Having said all that, the proposed Intrusion-Miner outperforms all the other algorithms used in this paper in terms of classification results.

While selecting an algorithm, Training Time (TT) of each algorithm is important and needs to be taken into the account. To build real-time network intrusion detection system, it is important to consider because it gives an idea which algorithm is suitable for real time environment. In this respect, the proposed Intrusion-Miner is also able to yield the classification model while taking an acceptable training time to build the model.

## V. Conclusion

In this paper, a hybrid learning approach called Intrusion-Miner has been proposed with the help of probabilistic BayesNet and IBK for better classification. Final result from the analysis of KDD 99 dataset using weka shown that it gave optimum performance during simulation that shows through tables, figures and graph to have a clear understanding for researchers. Simulation results proved that proposed Intrusion-Miner has better outcome as compared to other state-of-the-art algorithm and each classifier compared in term of AA, FP, TP and TT of algorithm. We also evaluated the speed of building classification model in which Intrusion-Miner has higher speed. From experimental results, it can be clearly observed that proposed approach not only achieve remarkable improvement in performance but also help in implementation of real time system applications and maximizing detection rate.

For future work, we recommend researchers to investigate other optimizing techniques for IDS that further improve the overall accuracy. In future, we plan to expand our work and use some other datasets as well.

## References

[1] M.H. Bhuyan, D.K. Bhattacharyya, J.K. Kalita, Network anomaly detection: methods, systems and tools. IEEE communications surveys & tutorials, vol. 16, no. 1, pp. 303-336, 2014.

[2] A.K. Ghosh, J. Wanken, F. Charron. Detecting anomalous and unknown intrusions against programs. in Computer Security Applications Conference, 1998. Proceedings. 14th Annual. 1998, IEEE.

[3] A.A. Ghorbani, W. Lu, M. Tavallaee, Network intrusion detection and prevention: concepts and techniques. Springer Science & Business Media, vol. 47, 2009.

[4] C. Kreibich, and J. Crowcroft, Honeycomb: creating intrusion detection signatures using honeypots. ACM SIGCOMM computer communication review, vol. 34, no. 1, pp. 51-56, 2004.

[5] E. De la Hoz, et al. PCA filtering and probabilistic SOM for network intrusion detection. Neurocomputing, vol. 164, pp. 71-81, 2015.

[6] S. Aljawarneh, M. Aldwairi, M.B. Yassein, Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model. Journal of Computational Science, vol. 25: pp. 152-160, 2018.

[7] U. Ravale, N. Marathe, P. Padiya, Feature selection based hybrid anomaly intrusion detection system using K means and RBF kernel function. Procedia Computer Science, vol. 45: pp. 428-435, 2015.

[8] S.A. Aljawarneh, R.A. Moftah, A.M. Maatuk, Investigations of automatic methods for detecting the polymorphic worm's signatures. Future Generation Computer Systems, vol. 60: pp. 67-77, 2016.

[9] M. Nezakatolhoseini, and M.A. Taherkhani, A Framework for Performance Evaluation of ASIPS in Network-Based IDS. arXiv preprint arXiv:1211.0620, vol. 4, no. 5, 2012.

[10] D.E. Denning, An intrusion-detection model. IEEE Transactions on software engineering, vol. 2, pp. 222-232, 1987.

[11] D. Anderson, T. Frivold, A. Valdes, Next-generation intrusion detection expert system (NIDES): A summary. 1995.

[12] D. Zhao, Q. Xu, Z. Feng. Analysis and design for intrusion detection system based on data mining. in Education Technology and Computer Science (ETCS), 2010 Second International Workshop on. 2010. IEEE.

[13] G. Nadiammai, M. Hemalatha, Effective approach toward Intrusion Detection System using data mining techniques. Egyptian Informatics Journal, vol. 15, no. 1, pp. 37-50, 2014.

[14] E. De la Hoz, et al., PCA filtering and probabilistic SOM for network intrusion detection. Neurocomputing, vol. 164, pp. 71-81, 2015.

[15] S.N. Pawar, R.S. Bichkar, Genetic algorithm with variable length chromosomes for network intrusion detection. International Journal of Automation and Computing, vol. 12, no. 3, pp. 337-342. 2015.

[16] F. Kuang, et al., A novel SVM by combining kernel principal component analysis and improved chaotic particle swarm optimization for intrusion detection. Soft Computing, vol. 19, no. 5, pp. 1187-1199, 2015.

[17] I. Ahmad, et al., Enhancing SVM performance in intrusion detection using optimal feature subset selection based on genetic principal components. Neural computing and applications, vol. 24, no. 7-8, pp. 1671-1682, 2014.

[18] C. Guo, et al., A distance sum-based hybrid method for intrusion detection. Applied intelligence, vol. 40, no. 1, pp. 178-188. 2014.

[19] S. Mukherjee, N. Sharma, Intrusion detection using naive Bayes classifier with feature reduction. Procedia Technology, vol. 4: pp. 119-128, 2012.

[20] S.K. Sharma, et al. An improved network intrusion detection technique based on k-means clustering via Naïve bayes classification. in Advances in Engineering, Science and Management (ICAESM), 2012 International Conference on. 2012. IEEE.

[21] Z.-S. Pan, et al. Hybrid neural network and C4. 5 for misuse detection. in Machine Learning and Cybernetics, 2003 International Conference on. 2003. IEEE.

[22] A.P. Muniyandi, R. Rajeswari, R. Rajaram, Network anomaly detection by cascading k-Means clustering and C4. 5 decision tree algorithm. Procedia Engineering, vol. 30, pp. 174-182, 2012.

[23] M. Panda, M.R. Patra, Network intrusion detection using naive bayes. International journal of computer science and network security, vol. 7, no. 12, pp. 258-263, 2007.

[24] Y. Li, L. Guo, An active learning based TCM-KNN algorithm for supervised network intrusion detection. Computers & security, vol. 26, no.7-8, pp. 459-467, 2007.

[25] Y.B. Bhavsar, and K.C. Waghmare, Intrusion detection system using data mining technique: Support vector machine. International Journal of Emerging Technology and Advanced Engineering, vol. 3, no. 3, pp. 581-586, 2013.

[26] L. Dhanabal, S. Shantharajah, A study on NSL-KDD dataset for intrusion detection system based on classification algorithms. International Journal of Advanced Research in Computer and Communication Engineering, vol. 4, no. 6, pp. 446-452, 2015.

[27] R.K. Kovarasan, and M. Rajkumar, An effective intrusion detection system using flawless feature selection, outlier detection and classification. Progress in Advanced Computing and Intelligent Engineering, vol. 713, pp.203-213, 2019.

[28] M. Turk, A. Pentland, Eigenfaces for recognition. Journal of cognitive neuroscience, vol. 3, no. 1, pp. 71-86, 1991.

[29] S. Stolfo, J., et al., Cost-based modeling for fraud and intrusion detection: Results from the JAM project. 2000, COLUMBIA UNIV NEW YORK DEPT OF COMPUTER SCIENCE.

[30] J. Jabez., S. Gowri, S. Vigneshwari, J. A. Mayan, and S. Srinivasulu. Anomaly Detection by Using CFS Subset and Neural Network with WEKA Tools, Information and Communication Technology for Intelligent Systems, vol 107, pp. 675-682, 2019.

## AUTHORS' PROFILE

**Samra Zafar** received the BS degree in computer science from COMSATA University Islamabad, Pakistan, in 2015. Currently Master student of Computer Science with the School of Computer Science and Technology, Dalian University of Technology. Her research interests include data mining, machine learning, WSN and software defined network.

**Muhammad Kamran** received the MS and PhD degrees in computer science from the National University of Computer and Emerging Sciences, Islamabad, Pakistan, in 2008 and 2012, respectively. Currently, he is an Assistant Professor with the College of Computer Science and Engineering at University of Jeddah, Jeddah, Saudi Arabia. His research interests include machine learning, evolutionary computation techniques, information security, health-informatics, big data analytics, and decision support systems.

**Xiaopeng Hu** received the PhD degree in computer science from Imperial College London, U.K., in 2005. He is currently a Professor of Computer Science with the School of Computer Science and Technology, Dalian University of Technology. His research interests include computer vision, machine learning, and sensor fusion.