Available online at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/cose

Computers & Security

TC 11 Briefing Papers

A novel combinatorial optimization based feature selection method for network intrusion detection

Anjum Nazir[a,b], *Rizwan Ahmed Khan*[a,*]

[a] *Faculty of IT, Barrett Hodgson University, Karachi, Pakistan*
[b] *FAST National University, Karachi Pakistan*

ARTICLE INFO

Article history:
Received 23 January 2020
Revised 27 November 2020
Accepted 26 December 2020
Available online 31 December 2020

Keywords:
Intrusion detection
Machine learning
Feature selection
Metaheuristics

ABSTRACT

The advancements in communication technologies and ubiquitous accessibility to a wide array of services has opened many challenges. Growing numbers of cyberattacks show that current security solutions and technologies do not provide effective safeguard against modern attacks. Intrusion is one of the main issue that has gone viral and can compromise the security of a network of any size. Intrusion Detection / Prevention Systems (IDS / IPS) are used to monitor, inspect and possibly block attacks. However, traditional intrusion detection techniques like signature or anomaly (network behavior) based approaches are prone to many weaknesses. Advancements in machine learning algorithms, data mining and soft computing techniques have shown potential to be used in IDS. All of these technologies, specially machine learning algorithms have to deal with the issue of high dimensionality of data /network traffic data as high dimensional data makes data sparse in hyper-space which restricts different algorithms scaling and generalization capabilities. Secondly, the problem magnitude also grows exponentially when IDS needs to make decision in a real time environment. One of the solution is to tackle this issue is to use feature selection techniques to reduce dimensionality of data. Feature selection is a process of selecting the optimal subset of features from a large feature-set to improve classification accuracy, performance and cost of extracting features. In this paper, we proposed a wrapper-based feature selection method called 'Tabu Search - Random Forest (TS-RF)'. Tabu search is used as a search method while random forest is used as a learning algorithm for Network Intrusion Detection Systems (NIDS). The proposed model is tested on the UNSW-NB15 dataset. The obtained results compared with other feature selection approaches. Results show that TS-RF improves classification accuracy while reducing number of features and false positive rate simultaneously.

1. Introduction

Internet has changed daily lives by providing economical, fast and reliable access to different types of services. Internet usage has spiked unprecedentedly in recent years

* Corresponding author.
E-mail address: rizwan17@gmail.com (R.A. Khan).
https://doi.org/10.1016/j.cose.2020.102164

Feldmann et al. (2020); Trevisan et al. (2020). Feature rich and widespread availability of Internet has also introduced different security problems like cyber espionage Hore and Raychaudhuri, targeted attacks Neupane et al. (2019), etc. Generally there are different reasons and motivations behind a cyberattack. Louvieris et. al. Louvieris et al. (2013) uncovered several reasons of increased growth of cyberattacks. One basic reason they pointed out is due to easy access and availability of hacking tools, users do not require sophisticated skills or deep knowledge to launch an attack. They also suggested that this is one of the leading factor for increase in cyberattacks.

Ensuring fundamental security triad (Confidentiality, Integrity and Availability) of an information system is becoming a real challenge. Confidentiality refers to protecting information asset against unintentional, unauthorized disclosure, integrity assures the correctness or accuracy of the stored or transmitted information, while availability defines the system or service is available to use. Organizations deploy different software and hardware-based security solutions like firewalls, antivirus (AV) solutions, Intrusion Detection and Prevention Systems (IDS / IPS) to defend against different security attacks. However, statistics published by Symantec Internet Security Threat Report (ISTR) 2019 Symantec (2019) reveals that these solutions are not robust to provide security against existing and new threats. Firewall is considered as first defense line that allows or denies network traffic on the basis of IP addresses and port numbers. Traditional firewalls significantly lacks advanced features such as threat prevention, Deep Packet Inspection (DPI), etc. DPI is a technique by which a firewall or a security solution can inspect and analyze message payload Dharmapurikar et al. (2003). It is extremely useful to monitor and inspect actual content being transferred. Antivirus (AV) software serves as a second defense line. It detects and blocks the execution and spread of any virus or malware. Malware or a malicious software is any program or file which is harmful for a *'computer system'* mal (2020). It stores patterns or strings (sequence of bytes) of known viruses in its database, which is known as virus signature Kaspersky: (2018). The main limitation of antivirus is it cannot detect or block the execution nor the spread of any new (unseen) virus whose signature is not updated yet in virus database.

On the other hand, intrusion detection / prevention system can perform deep packet inspection and looks inside packets' payload for any type of intrusion. Any unwanted or unauthorized activity that can compromise the security of a system or a network is known as intrusion worm (2020). The main goals of intrusion detection systems are (i) increase attack detection rate and (ii) decrease false alarm rate (false positive rate) Prasad et al. (2020). In literature Debar et al. (2000); Liao et al. (2013) IDS are classified based on different characteristics, however two main categories of intrusion detection systems are (i) misuse or signature-based and (ii) anomaly-based IDS. Misuse or signature-based IDS works like antivirus software. It performs deep packet inspection and looks inside packet's payload for any malignancy. Despite of high accuracy, signature-based IDS suffers from different problems such as (i) it cannot detect new attacks (ii) high search complexity, (iii) scalability, (iv) difficult to keep up to date against new intrusion. Anomaly-based IDS (A-IDS) learns and build a network / traffic profiles also known as baseline, which is used to find deviations Genereux et al. (2019). It is useful approach to detect new attacks, however it also suffers from different problems for example (i) high false positive rate, (ii) high training time for model building (iii) accuracy and validation of built model etc.

To achieve optimal security requirements of a network, researchers investigated machine learning approaches to develop an IDS that can detect novel attacks with high accuracy and low false positive rate. Machine learning uses mathematical and statistical models to establish and recognize patterns in large datasets Ripley (2007). Pattern recognition is the key to make predictions on unseen data. There are number of areas where machine learning has proved its potential such as cancer diagnosis Kourou et al. (2015), genetics and genome sequence analysis Libbrecht and Noble (2015), textual data classification Tong and Koller (2001), face recognition Chopra et al. (2005) affect analysis Khan et al. (2019a, 2012, 2019b).

Intrusion detection systems employ machine learning algorithms to classify packets as normal or an attack based on the information they carry. The information is extracted from the packets in the form of features. Network packets contain hundreds of features or attributes. These features help machine learning algorithm to make correct prediction. However, not all features contribute well in classification process. These features are known as irrelevant or weakly relevant features John et al. (1994). These features not only consume large amount of computational resources during processing but also slow down classifiers' detection speed and effects accuracy and false positive rate (FPR). Therefore, feature selection is performed to select best optimal features to improve detection accuracy, FPR, training and testing speed. As most of the information that is required to a classifier is centered around strong / relevant features, it also helps classification process, reduces processing and storage cost and enhance the understanding of data Mohammadi et al. (2019).

In this paper we proposed a new wrapper based feature selection method "Tabu Search - Random Forest (TS-RF)". Wrapper based feature selection method is discussed in Section 2.2. TS-RF is based on Tabu Search (TS) metaheuristic optimization algorithm Glover (1989, 1990) and Random Forest (RF) ensemble classifier Feng et al. (2015). In the proposed work, we used Tabu Search for feature searching while Random Forest as a learning method. Results are presented and discussed in Section 5, which show that TS-RF outperformed legacy and recently proposed feature selection methods by:

1. Improving classifier accuracy,
2. Reducing feature space / feature vector by more than 60%. Thus reducing computational complexity of the model.
3. Reducing misclassification rate among different attack types by removing features causing confusion / ambiguity.
4. Significant improvement in detection accuracy for attacks having low sample count in the dataset.
5. Reducing false positive rate.

Rest of the paper is organized as follows. Section 2.2 presents common feature selection approaches used in network intrusion detection. Section 3 organizes literature review / related work. In Section 4 we presented

our proposed feature selection algorithm. Experiments and results are presented in Section 5 followed by conclusion and future work.

2. Dimension reduction in network intrusion detection

Intrusion detection datasets contain large amount of data with high dimensions. Dimension refers to number of features in the dataset. As discussed in Section 1, dimension reduction (feature selection) techniques play an incredibly important role to overcome the issues caused by 'curse of dimentionality' Bauer and Kohler (2019). Dimension reduction not only improves classification accuracy but also improves space-time complexity of machine learning classifier Kambhatla and Leen (1997). Dimension reduction or feature selection techniques main goal is to select best relevant features that can truly represent complete dataset without compromising classifier accuracy. It tries to remove redundant and irrelevant features from the dataset. In the next secion we will discuss about feature selection.

2.1. Feature selection basics

In this section we will present components of feature selection algorithm. Feature selection algorithm comprises of two main parts (i) feature search strategy and (ii) feature evaluation criteria.

2.1.1. Feature search

Feature search process finds optimal set of features from the dataset or examples. Its main objective is to remove irrelevant and redundant features. Features that do not contribute well in classification process are known as irrelevant features while redundant features have high degree of correlation among them. A feature is considered as relevant if it carries some information about the target or class. Relevant features can be further divided into strongly relevant features and weakly relevant features. The objective of feature search is to identify relevant features to maximize classification accuracy, reduce false positive rate etc. For example if $X = \{x_1, x_2, x_3, \ldots, x_n\}$ is a set of 'n' input features, then $Y_m \subset X_n$ such that $m < n$ and $f_{eval}(Y_m)$ is maximized, where 'm' is the number of features in the optimal feature vector and $f_{eval}(Y_m)$ is the cost or fitness of the optimal feature set.

One basic technique to find optimal feature set is to apply exhaustive search. Exhaustive search technique will definitely find optimal features. However, search complexity of exhaustive search is very high 2^n, where n is the number of feature, the problem becomes NP-hard problem[1]. In literature, researchers used other techniques for feature searching, for example branch and bound algorithm Atashpaz-Gargari et al. (2018), sequential search methods Castañeda Gonzalez et al. (2018), random search and metaheuristics Jiménez et al. (2017).

Branch and bound algorithm makes problem easier than exhaustive search however it is not practical to apply it if $n > 30$ Pudil et al. (1994). On the other hand, sequential search methods like forward, backward, bidirectional selection usually use greedy techniques to find best feature set. The biggest problem of these techniques is they do not guarantee to achieve global optimal solution. Random search methods such as genetic algorithm add some randomness in the search procedure to help to escape from local optimum. In this paper we used Tabu Search metaheuristics for feature searching as it does not suffer from search space complexity. Tabu search maintains tabu-list to keeps track of its move. Tabu search is discussed in Section 4.1.

2.1.2. Feature evaluation

In feature evaluation, each feature or subset of features are evaluated on the basis of their relevancy to the class label. Irrelevant or redundant features are identified through this process. Feature evaluation requires 'evaluation criteria' that must be followed for feature selection. Different evaluation criteria may yield different set of features average precision, accuracy, false positive rate, and Receiver Operating Characteristic (ROC) analysis are few examples of different evaluation criteria studied in the literature Liu and Motoda (2007).

2.2. Feature selection methods

There are three common types of feature selection methods which are discussed below.

2.2.1. Filter method

Filter based feature selection methods use 'intrinsic' property of the data to select features. They do not depend upon the any learning algorithm (a classifier or clustering algorithm) as wrapper methods do. They use different feature ranking techniques that actually identify the relevancy of features and decides weather to select or discard this feature. Rank or score is calculated by different techniques such as dataset statistical properties e.g. entropy Dash et al. (2002), Laplacian score He et al. (2006) etc. Filter algorithms are usually computationally less expensive than wrapper and embedded approaches, however one common drawback for filter methods is that they are adequate only for independent features.

2.2.2. Wrpaper method

Wrapper-based feature selection methods consist of three components which are (i) search strategy, (ii) predictor function (learning process) and (iii) evaluation or fitness function. Search strategy selects the subset of features to be evaluated as discussed in Section 2.1.1. Predictor function uses any classification algorithm that is used to determine the performance of the selected features against the objective or fitness function. The performance of wrapper method is better than filter-based selection approaches, however it is more time consuming as compared to filter method.

2.2.3. Embedded method

Embedded methods combine the best qualities of filter and wrapper techniques discussed above. It actually introduces an *interaction* between feature search strategy and learning process (predictor function). In embedded methods, a learning algorithm takes advantage of its own variable selection process and performs feature selection and classification

[1] Non-deterministic polynomial time.

simultaneously. Due to this it reaches towards fast convergence and finds optimal solution quickly as compared to wrapper technique. Secondly, embedded methods work by adding a penalty against complexity to reduce the degree of overfitting or variance of a model by adding more bias.

3. Related work

For any machine learning experiment dataset plays a critical role. In literature we found different datasets purposely built for IDS study. IDS datasets are classified into (i) network and (ii) host datasets. Network datasets contain normal and attack traffic while host datasets contains host or computer activities over a period of time. In this paper our focus is on network-based IDS so we will restrict our discussion to network datasets only.

On this subject, there are several network-based IDS datasets published in last two decades Hindy et al. (2018). Usually a dataset consists of number of attributes known as input vector or feature vector and an output (class label). A comprehensive summary of network datasets can be found in Ring et al. (2019). We focused our research on UNSW-NB15 Moustafa and Slay (2015) dataset because this is one of the most recent dataset that contains modern attacks. Legacy datasets like KDDCup99 dataset (2018), NSL-KDD Tavallaee et al. (2009) etc. cannot meet current research requirements due to dynamically changing security and operational demands of a network. Despite having dataset intrinsic weaknesses, they also do not have modern day normal traffic nor they have updated attack patterns.

UNSW-NB15 dataset is recently proposed by Moustafa et. al. Moustafa and Slay (2015) to address these issues. It is a hybrid dataset which consists of real modern normal network activities coupled with synthetic updated attacks. This is the main motivation to use UNSW-NB15 dataset for our research. Dataset contains nine type of attacks with one normal traffic. This section presents literature review on feature selection techniques for network based IDS dataset. We selected notable papers published in last 5 years.

In Eesa et al. (2015b) Eesa et. al. presented a novel feature selection method based on Cuttlefish Optimization Algorithm (COA). Authors used Decision Tree (DT) classifier on KDDCup99 dataset (2018) dataset to evaluate different parameters such as accuracy, false positive, detection rate, ROC etc. They suggested that accuracy and detection rate increased with the reduction of features. However, they did not provide any detail about final optimal features set. Essa et. al. Eesa et al. (2015a) also proposed a wrapper-based feature selection algorithm that uses bees optimization algorithm. The model is known as ID3-BA (Bee Algorithm). In this work, bee algorithm is used to generate initial and following population of sub-optimal feature set in each iteration. They used ID3 Cheng et al. (1988) as a fitness function. The whole model is tested on KDDCup99 dataset. The main limitation of both the works is the lack of results on new datasets like UNSW-NB15 Moustafa and Slay (2015)

Guha et. al. Guha et al. (2016) presented an attack detection method for cloud computing infrastructure. They used Artificial Neural Networks (ANN) with genetic feature selection. The proposed work is tested on NSL-KDD Tavallaee et al. (2009) and UNSW-NB15 dataset. Simulation results demonstrate better detection accuracy. However, it does not present any details about final optimal feature set. In addition to this, the authors did not compare the results of proposed model with current state of the art techniques.

Praneeth et. al. Nskh et al. (2016) applied Principal Component Analysis (PCA) for feature selection on KDDCup99 dataset. Selected features were tested against different kernels (such as linear, polynomial, radial basis) of Support Vector Machine (SVM) Tong and Koller (2001). The results suggested that the performance of radial-basis kernel was better than linear and polynomial kernels. In similar vein, A. Hadri et. al. Hadri et al. (2016) used PCA and fuzzy-PCA for feature selection process. They used KDD-Cup99 dataset for their simulation. Simulation results showed that fuzzy-PCA performed better than normal PCA for feature selection. A. Rama and W. Gata Syarif and Gata (2017) also used PCA technique for feature selection. They used KNN Peterson (2009) and proposed a hybrid classifier based on binary Particle Swarm Optimization (PSO) Ali et al. (2018). Results show that proposed algorithm has accuracy around 99% while KNN accuracy remain around 97% on KDD-Cup99 dataset. Authors did not provide any details about the final feature set. S. Zhao et. al. Zhao et al. (2017) also used PCA for feature selection but they used different classifiers in their study. Their results represents Softmax regression performed better than KNN.

M. A. Zewairi et. al. Al-Zewairi et al. (2017) work focused towards classifier designing however they also used Gedeon method for feature selection. They did not provide any comparison about the performance of the classifier before or after feature selection. They used UNSW-NB15 dataset for their studies. Similarly P. Mishra et. al Mishra et al. (2017) also studied their work on UNSW-NB15 dataset. They used Recursive Feature Elimination (RFE) and chi-square + RFE techniques for feature selection. They tested subset of features on number of different classifiers. C. Khammassi and S. Krichen Khammassi and Krichen (2017) worked on UNSW-NB15 and KDD-Cup99 dataset. They created genetic algorithm based wrapper method, known as GA-LR. They used logistic regression as a learning algorithm.

Muna A.H. et. al. Muna et al. (2018) used deep auto-encoder on NSL-KDD and UNSW-NB15 datasets. Their main focus was the development of new classifier for Industrial Internet Control System (IICS). One weakness in their work is these datasets are not developed for IICS.

Selvakumar, B et. al. Selvakumar and Muneeswaran (2019) used nature-inspired firefly algorithm for feature selection of KDD-Cup99 dataset. They used Bayesian Network (BN) and C4.5 for classification. The weakness in their is due to the use of outdated KDDCup99 dataset. Kasongo et. al. Kasongo and Sun (2020) proposed Feed-Forward Deep Neural Network (FFDNN) wireless IDS system using a Wrapper Based Feature Extraction Unit (WFEU). They used different classifiers like SVM. naive-bayes, decision tree and k-nearest neighbor on UNSW-NB15 dataset. Comparison results show that WFEU-FFDNN has higher accuracy as compared to other schemes.

Table 1 – Tabu Search algorithm parameters.

Parameter	Value
Tabu List Size	7
No. of neighbor	6
Aspiration Level	0.05
No. of Iterations	100

4. Proposed feature selection method (TS-RF)

In this section we will describe our proposed feature selection technique Tabu Search - Random Forest (TS-RF). Tabu Search (TS) was first introduced by Fred Glover Glover (1989, 1990) in 1997 as a general iterative metaheuristics for solving combinatorial optimization problems. Metaheuristics have gained noticeable popularity and importance in past two decades. It provides äcceptable solution" in reasonable time where other optimization techniques face trouble. Metahueristics are generally classified in two groups (i) single solution-based metaheuristics and (ii) population-based metaheuristics. Single solution based techniques construct a single solution and tries to improve it in every iteration whereas population based metaheuristics generate multiple solutions in each iteration, select the best available solution and tries to improve it further e.g. genetic algorithm. Tabu Search is a single solution-based optimization algorithm.

4.1. *Overview of Tabu search*

Tabu search is a form of local neighborhood search. Each feasible solution, $S \in \Omega$, has an associated set of neighbors, $N(S) \subseteq \Omega$, where Ω is a set of feasible solutions. A solution $S' \in N(S)$ can be reached from S by an operation called a "move" to S'. TS moves from a solution to its best admissible neighbor, even if this causes the objective or fitness function to deteriorate. This is non-greedy behaviour of tabu search which is also helpful to avoid trapping in local optima.

To avoid cycling and trapping into local minima, solutions that were recently explored are declared as "forbidden" or "tabu". These recently visited solutions are kept in a list which is known as tabu list for a number of iterations. For every move, tabu list is consulted each time. If a move is present in the tabu list, then algorithm discards that move and forwards to next iteration. The tabu status of a particular solution can be overridden when certain criteria (aspiration criteria or level) is satisfied. Sometimes intensification and diversification strategies are used to improve the search. In the first case, the search is accentuated in promising regions of the feasible domain. In the second case, an attempt is made to consider solutions in a broad area of the search space. TS algorithm parameters are mentioned in Table 1

4.2. *Random forest (RF)*

Random Forest is basically a tree based classifier that builds a set of decision trees randomly. The classification of an input sample is determined by the majority classification by the ensemble. It utilizes different ensemble techniques Banfield et al. (2007) that is why it is also categorized as ensemble classifier. RF can perform classification and regression tasks. The performance of RF classifier is generally better than decision tress on unseen data Khan et al. (2013).

4.3. *Fitness function*

Third important component of a wrapper-based feature selection method is fitness or evaluation function (also known as cost function or objective function). Fitness function defines the objective or goal to achieve. It is calculated for each feasible solution in the search space $f : S \rightarrow \mathbb{R}$. Eq. (1) is proposed to calculate the fitness of each solution. The inspiration for Eq. (1) is taken from the behavior of artificial neural network, where weights are assigned to each input that may influence the result.

$$f(cost) = b + \sum_{i=1}^{n} x_i w_i \tag{1}$$

where, x_i : represents the objective we want to optimize,
w_i : weight associated with the objective

Our aim is to (i) improve classification accuracy, (ii) reduce False Positive Rate (FPR) and (iii) reduce the number of features. Therefore, the problem becomes multi-objective optimization problem with conflicting objectives as generally improving accuracy also increases FPR. Therefore, we converted conflicting objectives into non-conflicting objectives with help of Eq. (2) which calculates classification error in terms of accuracy.

$$Error(e) = 100 - Accuracy \tag{2}$$

Eq. (1) is a generic equation. The customized expanded version of Eq. (1) is presented in Eq. (3). In Eq. (3) we replaced x_i with actual objectives we want to optimize. Here x_1 can be replaced by 'e' which represents classification error, x_2 can be replaced by 'e' which represents number of features and x_3 can be replaced by 'fpr' which shows weighted false positive rate.

$$f(cost) = w_1 * e + w_2 * n + w_3 * fpr \tag{3}$$

In Eq. (3) $b = 0$ and $w_1 = 0.333$, $w_2 = 0.333$, $w_3 = 0.333$. Equal weights are assigned to each objective to avoid any influence.

4.4. *Tabu search - Random forest (TS-RF) algorithm*

Step by step working of TS-RF is described below while graphical representation of the steps / proposed algorithm is presented in Fig. 1 and its pseudocode is presented in Algorithm 1 .

1. **Parameter Initialization:** Initially, algorithm related parameters are initialized, this includes parameters presented in Table 1.
2. **Cost of initial solution:** Initial feature vector includes features based on random selection or complete features list. Initial feature list can serve as a base feature vector until new best feature list becomes available. We included all features in the initial solution and calculated the cost as per Eq. (3).

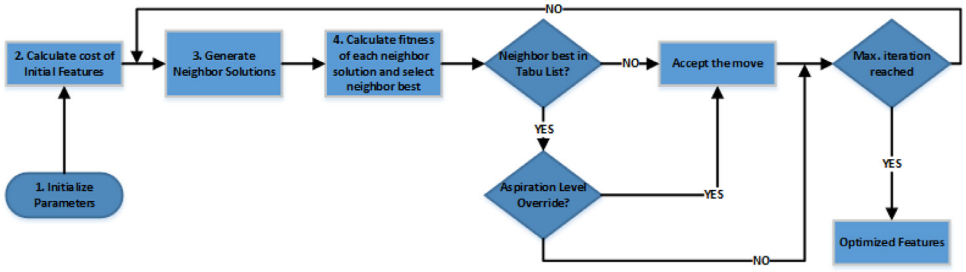

Fig. 1 – Flowchart representation of proposed feature selection method .

Algorithm 1: Tabu Search - Random Forest (TS-RF) Algorithm.

Parameters:
Ω: Set of feasible solutions
S: Current solution
S^*: Best admissible solution
f(cost): Objective function
N(S): Neighborhood of S
V^*: Sample of neighborhood solution
TL: Tabu List
AL: Aspiration Level
Q: Fixed number of iterations

```
Start with initial feasible solution S ∈ Ω
Initialize TL and AL
for Q do
    Generate neighbor solutions V* ⊂ N(S)
    Calculate fitness Cost(V*) and find best S* ∈ V* if Move S* from S is not in TL then
        Accept the move and update current solution S
        Update TL and AL
        Increment iteration number
    else
        if cost(S*) < AL then
            Accept the move and update current solution S
            Update TL and AL
            Increment iteration number
```

3. **Neighborhood solutions:** Neighbors are generated by randomly adding or deleting a feature from the feature vector. This process is known as "move". How many neighbor-solutions should be generated depends upon dataset. We generated 6 neighbor-solutions for each round by applying $\sqrt{n}$ rule where $n = 43$.
4. **Fitness Calculation:** We calculated the fitness of each neighbor-solution generated in step 3 by using Eq. (3). If the fitness (cost) of any move (neighbor-solution) is better than current best, then that move (neighbor-solution) is candidate solution and can become global best solution.
5. **Tabu move and Tabu List:** Tabu search maintains tabu list to keep tracks of last *x* number of moves (refer Section 4.1). If the new move is not in tabu list, then move will be accepted and global best solution will be updated by candidate solution.
6. **Aspiration criterion:** Aspiration criterion is a mechanism used to temporarily override the tabu list 'rule'. If the cost of forbidden move is less aspiration level, then tabu rule will be overridden and move will be accepted.
7. **Stopping Criteria:** There are three common stopping criteria which are (i) stop when predefined number of iterations reached, (ii) stop when no further improvement observed after certain number of iterations and (iii) stop when objective function reached at required threshold. We used fixed number iterations approach and executed the simulation for 100 times.

5. Experiments and results

In this section we will discuss the experiments performed to assess proposed feature selection algorithm and their results. This section comprises of two parts, in Section 5.1 we presented comparison of our proposed feature selection algorithm 'TS-RF' with legacy techniques. Results are obtained by comparing proposed approach with different machine learning classifiers. Machine learning algorithms can be classified into different families, based on how they operate. For example parametric, non-parametric, neural networks, probability or tree-based etc algorithms are commonly used in the literature. We included Naive Bayes (NB), K-Nearest Neighbor (kNN), Random Forest (RF) and Multi-layer Perceptron (MLP) for comparison.

In Section 5.2 we compared 'TS-RF' with state of the art feature selection techniques proposed in recent studies. We assessed our proposed method on UNSW-NB15 Moustafa and Slay (2015) dataset. Rationales of using UNSW-NB15 dataset are already discussed in Section 3.

5.1. *Comparison of 'TS-RF' with legacy feature selection techniques*

As discussed in Section 2.2 the objective of feature selection is to identify those features in the dataset that contain most information. There are many approaches which are being used over the years for this subject that can be considered as legacy techniques now. We focused on entropy and correlation based techniques such as Gain Ratio Quinlan (1986), Chi-Square and Pearson Correlation Liu and Motoda (2007). Gain Ratio is an entropy-based Sethna (2006) feature selection method that measures the importance or relevance of a feature with respect to the corresponding class. Chi-Square and Pearson Cor-

Table 2 – Results of Feature Selection Methods: Gain Ratio and Pearson Correlation based methods calculate the importance of each feature in terms of *"Rank or weight"* ranging between [0, 1]. For Chi Square method we normalized the output between [0, 1]. This table present attributes and their corresponding weights for all three filter based feature selection technique and TS-RF. TS-RF column consists of binary value '0' or '1', zero means the attribute is not present in the final optimal feature set and has been removed.

Feature	Pearson Correlation Method	Gain Ratio Method	Chi Square Method	TS-RF
ct_dst_sport_ltm	0.395	0.377	0.320699574	1
ct_dst_src_ltm	0.363	0.175	0.305822512	1
ct_src_dport_ltm	0.363	0.262	0.286043105	1
sttl	0.36	0.505	0.282999734	1
ct_srv_dst	0.35	0.19	0.32740075	1
ct_srv_src	0.348	0.172	0.306081844	1
ct_dst_ltm	0.337	0.195	0.288217833	1
xServ	0.24	0.38	0.291684417	1
is_sm_ips_ports	0.096	0.225	0.005158653	1
smean	0.095	0.25	0.745379259	1
dloss	0.054	0.188	0.219287557	1
ct_flw_http_mthd	0.051	0.121	0.038753231	1
dbytes	0.043	0.219	0.491079696	1
sbytes	0.024	0.272	1	1
sloss	0.022	0.18	0.178023703	1
response_body_len	0.014	0.182	0.082514424	1
ct_src_ltm	0.312	0.153	0.24359123	0
ct_state_ttl	0.306	0.432	0.28834849	0
xState	0.284	0.328	0.165153803	0
swin	0.268	0.278	0.093005441	0
dwin	0.263	0.271	0.090222665	0
xProt	0.245	0.291	0.283548039	0
rate	0.225	0.17	0.302254754	0
dttl	0.214	0.456	0.283484592	0
dtcpb	0.208	0.271	0.09022709	0
stcpb	0.208	0.271	0.090244809	0
dmean	0.207	0.2	0.364841579	0
dload	0.203	0.205	0.276768666	0
ackdat	0.164	0.254	0.216142431	0
tcprtt	0.161	0.259	0.22409221	0
synack	0.141	0.251	0.225966996	0
sload	0.098	0.22	0.676250756	0
sinpkt	0.092	0.132	0.245189132	0
dpkts	0.067	0.207	0.291838911	0
dur	0.062	0.156	0.29300127	0
djit	0.052	0.174	0.206908102	0
ct_ftp_cmd	0.047	0.123	0	0
is_ftp_login	0.047	0.123	0	0
spkts	0.045	0.168	0.228175389	0
trans_depth	0.041	0.117	0.018255278	0
sjit	0.033	0.168	0.215903334	0
dinpkt	0.03	0.219	0.270306487	0

relation are correlation-based techniques which measure the relationship or dependency between two variables like input feature and class label. Chi-Square method is based on χ^2 test. It calculates the independence of two events such as A and B if $P(AB) = P(A)P(B)$ or $P(A|B) = P(A)$ and $P(B|A) = P(B)$. Pearson correlation finds linear relationship or dependency between two features to calculate the importance of a particular feature.

Table 2 presents the results achieved after feature selection process. It shows attributes names and their corresponding weights or ranks calculated by each algorithm. The weight is normalized between [0, 1]. TS-RF column consists of binary value '0' or '1', zero means the attribute is not present in the final optimal feature set and has been removed.

Data presented in Table 2 is used for further analysis. In the first stage we selected all those features whose weights are greater than or equal to 0.2. The selected features are then feed to machine learning classifier for classification. Results of this setup is presented in Table 2. In the second stage, we repeated the experiment and selected all those features whose weights are greater than or equal to 0.15. Results obtained from these experiments are presented in Figs. 2, 3 and Tables 3, 4. Findings of these experiments are summarized below.

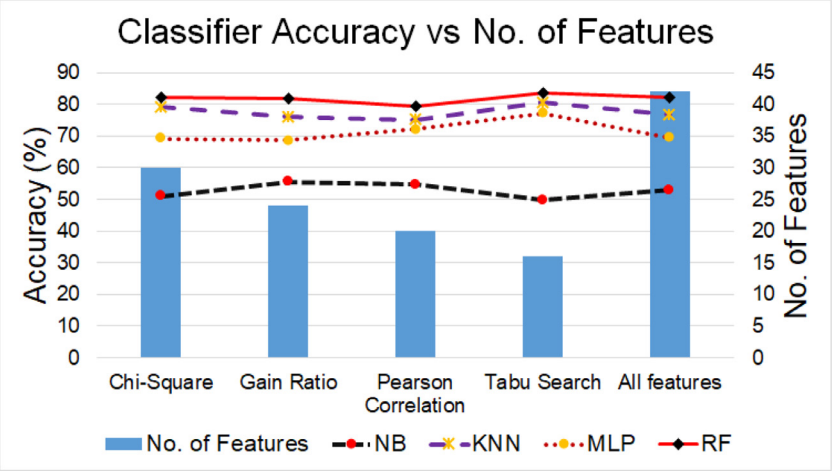

Fig. 2 – This figure represents classifier accuracy vs no. of features selected, (when cutoff value is 0.2).

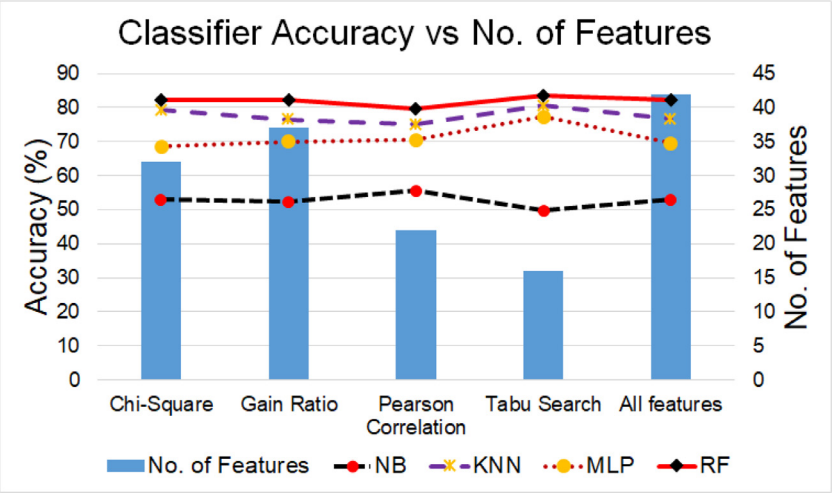

Fig. 3 – This figure represents classifier accuracy vs no. of features selected, (when cutoff value is 0.15).

Table 3 – Impact of feature selection techniques on time complexity of classifiers at Cutoff=0.2, +ve and -ve values show percentage improvement and decrement respectively.

Feature Selection Method	RF (%)	NB (%)	KNN (%)	MLP (%)
Chi-Square	+11.512	+25.76	No improvement	+0.90
Gain Ratio	+33.30	+46.01	-60	+8.02
Pearson Correlation	+37.47	+57.67	-40	-67.55
Tabu Search	+39.44	+45.40	+20	+95.40

Table 4 – Impact of feature selection techniques on time complexity of classifiers at Cutoff=0.15, +ve and -ve values show percentage improvement and decrement respectively.

Feature Selection Method	RF (%)	NB (%)	KNN (%)	MLP (%)
Chi-Square	+9.42	+26.38	+40	-87.70
Gain Ratio	+6.88	+12.88	+40	-2.24
Pearson Correlation	+27.37	+44.17	+60	+4.44
Tabu Search	+39.44	+45.40	+20	+95.40

Table 5 – Comparison of TS-RF with State of the Methods. Results show that TS-RF achieved higher accuracy with reduced no. of features.

Reference	Year	Classifier	Feature Selection	Accuracy	FPR	Number of Features	$f(cost)$ Eq. (3)
K. Khammassi Khammassi and Krichen (2017)	2017	Decision Tree	GA-LR	81.42	6.39	21	15.30
Moustafa et. al. Moustafa and Slay (2017)	2017	LR	Hybrid Association Rules	83	14.2	11	14.05
Tama et. al. Tama et al. (2019)	2019	TSE-IDS	Hybrid (PSO, ACO and GA)	91.27	8.73	19	12.19
Kasongo et. al. Kasongo and Sun (2020)	2020	RF	WFEU	77.16	N/A	22	$14.93 + \Delta fpr$
Kumar et. al. Kumar et al. (2020)	2020	N/A	Rule based	84.83	12.5	13	13.54
Proposed (TS-RF)	2020	RF	TS-RF	83.12	3.7	16	12.18

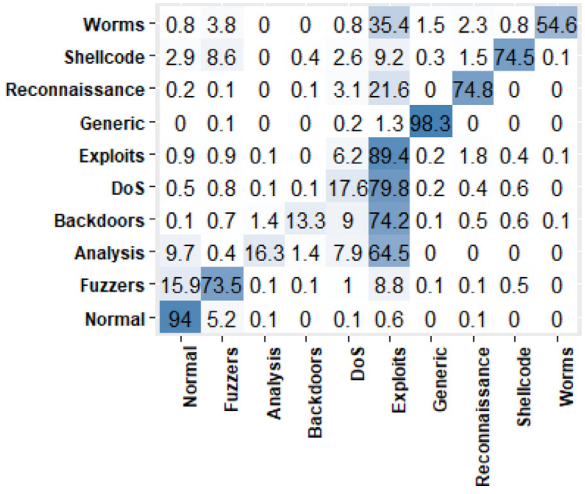

Fig. 4 – Confusion Matrix for Random Forest classifier after application of TS-RF.

- Tabu Search - Random Forest (TS-RF) produced better results, it reduced number of features up to 62% comparing with total number of features.
- It also increased classification accuracy by 2% for random forest, 4% for K nearest neighbor and upto 7% for MLP which is a significant improvement.
- We also observed that there is a slight decrease in the accuracy for NB classifier. But at the same time, time complexity for TS-RF is also significantly reduced.
- Time complexity for MLP is exceptionally reduced by 95% while for RF, NB and KNN classifiers it is reduced up to 40%, 45% and 20% respectively.
- From these results it is evident that TS-RF selected only important features from the dataset which not only improved classification accuracy but also it reduced time complexity.

We observed that classification accuracy of random forest is highest among all classifiers. Fig. 4 shows the confusion matrix of RF classifier for features selected by TS-RF. From this confusion matrix we can analyze:

- TS-RF has successfully removed those features which were creating confusion among classes and increasing misclassification rate.
- Classes having very limited sample size like Worm traffic (0.0007%) and Shellcodes (1%) of the total distribution have responded extraordinarily. Worm misclassification error is reduced by more than 30% while shellcodes accuracy has improved around 13% compared to full features RF accuracy.
- The reduction in misclassification will help IDS to correctly classify and block attacks.
- It will also help security analyst to spend less time to investigate intrusion events and prepare an appropriate action response.

5.2. *Comparison with state of the art feature selection methods*

In this section we will present comparison of results of our proposed feature selection algorithm TS-RF with the recent techniques published in last three years. We considered research papers which included UNSW-NB15 dataset in their study also we filter all those papers that provided details about their final optimal feature set.

In Table 5 we presented comparison results. Cost for all techniques is calculated by using Eq. (3) which includes classification accuracy, false positive rate and number of features. Simulation results show that TS-RF has lowest cost as compared to all other techniques. Results show that Tama et al. (2019) has highest accuracy which is 91.27%, however its false positive rate is also high and number of features in the final dataset. Similarly, in Moustafa and Slay (2017) and Kumar et al. (2020) number of features in the final dataset is less than TS-RF technique, but their false positive rate is also high which increased overall cost of techniques. Results show that TS-RF has outperform all recent studies performed on UNSW-NB15 dataset.

6. Conclusion & future work

In this paper we proposed a new wrapper-based feature selection algorithm, Tabu Search - Random Forest (TS-RF)

for intrusion detection. TS-RF wrapper exploits Tabu search metaheuristic algorithm for feature searching / weighting and Random forest as a learning algorithm. Our study comprises of two phases, initially we compared TS-RF with three legacy feature selection approaches which are Gain Ratio, Chi-Square and Pearson Correlation. Results show that TS-RF not only improved classification accuracy but also reduced time-complexity of the model.

In the second phase, we compared TS-RF with recent feature selection techniques published in the literature. We proposed a weighted-sum based cost or fitness function that includes three input parameters which are (i) classification accuracy, (ii) false positive rate and (iii) number of features. The cost function is use to find overall effectiveness of the technique. Our results show that TS-RF has lowest cost among all other recent techniques.

In future we plan to address class imbalance problem present in UNSW-NB15 dataset because it does not only impact the classifier accuracy but also it increases misclassification rate and false positives. Extraction of new discriminant / salient features from the original data by applying deep learning may also help to reduce misclassification rate and improve accuracy. Another possible avenue is to work on the application of metaheurisitc algorithms like TS or other evolutionary computing or nature inspired algorithm for threat detection and feature selection because these optimization algorithms have also shown better results in other domains.

Declaration of Competing Interest

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

CRediT authorship contribution statement

Anjum Nazir: Conceptualization, Validation, Methodology, Visualization, Software, Formal analysis, Investigation, Data curation, Writing - original draft. **Rizwan Ahmed Khan:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Visualization, Writing - review & editing.

REFERENCES

Al-Zewairi M, Almajali S, Awajan A. Experimental Evaluation of a Multi-layer Feed-forward Artificial Neural Network Classifier for Network Intrusion Detection System. In: 2017 International Conference on New Trends in Computing Sciences (ICTCS). IEEE; 2017. p. 167–72.

Ali MH, Mohammed BADA, Ismail A, Zolkipli MF. A new intrusion detection system based on fast learning network and particle swarm optimization. IEEE Access 2018;6:20255–61.

Atashpaz-Gargari E, Reis MS, Braga-Neto UM, Barrera J, Dougherty ER. A fast branch-and-bound algorithm for u-curve feature selection. Pattern Recognit 2018;73:172–88.

Banfield RE, Hall LO, Bowyer KW, Kegelmeyer WP. A comparison of decision tree ensemble creation techniques. IEEE Trans Pattern Anal Mach Intell 2007;29(1):173–80.

Bauer B, Kohler M, et al. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. Ann Stat 2019;47(4):2261–85.

Cheng J, Fayyad UM, Irani KB, Qian Z. Improved Decision Trees: A Generalized Version of Id3. In: Machine Learning Proceedings 1988. Elsevier; 1988. p. 100–6.

Chopra S, Hadsell R, LeCun Y, et al. Learning a Similarity Metric Discriminatively, with Application to Face Verification. In: CVPR (1); 2005. p. 539–46.

Dash M, Choi K, Scheuermann P, Liu H. Feature Selection for Clustering-a Filter Solution. In: 2002 IEEE International Conference on Data Mining, 2002. Proceedings. IEEE; 2002. p. 115–22.

dataset K.c. 2018. https://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html, accessed:-06-28.

Debar H, Dacier M, Wespi A. A revised taxonomy for intrusion-detection systems. Annales des télécommunications 2000;55:361–78. Springer.

Dharmapurikar S, Krishnamurthy P, Sproull T, Lockwood J. Deep Packet Inspection Using Parallel Bloom Filters. In: 11th Symposium on High Performance Interconnects, 2003. Proceedings.. IEEE; 2003. p. 44–51.

Eesa AS, Orman Z, Brifcani AMA. A new feature selection model based on id3 and bees algorithm for intrusion detection system. Turkish Journal of Electrical Engineering & Computer Sciences 2015a;23(2):615–22.

Eesa AS, Orman Z, Brifcani AMA. A novel feature-selection approach based on the cuttlefish optimization algorithm for intrusion detection systems. Expert Syst Appl 2015b;42(5):2670–9.

Feldmann A., Gasser O., Lichtblau F., Pujol E., Poese I., Dietzel C., Wagner D., Wichtlhuber M., Tapidor J., Vallina-Rodriguez N., et al. The lockdown effect: Implications of the covid-19 pandemic on internet traffic. 2020. ArXiv preprint arXiv:2008.10959.

Feng Z, Mo L, Li M. A Random Forest-based Ensemble Method for Activity Recognition. In: 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2015. p. 5074–7.

Genereux SJ, Lai AK, Fowles CO, Roberge VR, Vigeant GP, Paquet JR. Maidens: mil-std-1553 anomaly-based intrusion detection system using time-based histogram comparison. IEEE Trans Aerosp Electron Syst 2019;56(1):276–84.

Glover F. Tabu searchpart i. ORSA Journal on Computing 1989;1(3):190–206. doi:10.1287/ijoc.1.3.190.

Glover F. Tabu searchpart ii. ORSA Journal on computing 1990;2(1):4–32.

Castañeda Gonzalez J, Alvarez-Meza A, Orozco-Gutierrez A. An Enhanced Sequential Search Feature Selection Based on Mrmr to Support Fcd Localization. In: Iberoamerican Congress on Pattern Recognition. Springer; 2018. p. 487–95.

Guha S, Yau SS, Buduru AB. Attack Detection in Cloud Infrastructures Using Artificial Neural Network with Genetic Feature Selection. In: 2016 IEEE 14th Intl Conf on Dependable, Autonomic and Secure Computing, 14th Intl Conf on Pervasive Intelligence and Computing, 2nd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech). IEEE; 2016. p. 414–19.

Hadri A, Chougdali K, Touahni R. Intrusion Detection System Using Pca and Fuzzy Pca Techniques. In: 2016 International Conference on Advanced Communication Systems and Information Security (ACOSIS). IEEE; 2016. p. 1–7.
He X, Cai D, Niyogi P. Laplacian Score for Feature Selection. In: Advances in neural information processing systems; 2006. p. 507–14.
Hindy H., Brosset D., Bayne E., Seeam A., Tachtatzis C., Atkinson R., Bellekens X.. A taxonomy and survey of intrusion detection system design techniques, network threats and datasets. 2018. arXiv:1806.03517.
Hore S., Raychaudhuri K. Cyber Espionagean Ethical Analysis. In: Innovations in Computational Intelligence and Computer Vision. Springer. p. 34–40.
Jiménez F, Sánchez G, García JM, Sciavicco G, Miralles L. Multi-objective evolutionary feature selection for online sales forecasting. Neurocomputing 2017;234:75–92.
John GH, Kohavi R, Pfleger K. Irrelevant Features and the Subset Selection Problem. In: Machine Learning Proceedings 1994. Elsevier; 1994. p. 121–9.
Kambhatla N, Leen TK. Dimension reduction by local principal component analysis. Neural Comput 1997;9(7):1493–516.
Kasongo SM, Sun Y. A deep learning method with wrapper based feature extraction for wireless intrusion detection system. Computers & Security 2020;92:101752.
Kaspersky:. antivirus fundamentals: Viruses, signatures, disinfection. 2018. https://www.kaspersky.com/blog/signature-virus-disinfection/13233/, accessed:-05-16.
Khammassi C, Krichen S. A ga-lr wrapper approach for feature selection in network intrusion detection. computers & security 2017;70:255–77.
Khan RA, Crenn A, Meyer A, Bouakaz S. A novel database of children's spontaneous facial expressions (liris-cse). Image Vis Comput 2019a;83:61–9.
Khan RA, Meyer A, Konik H, Bouakaz S. Human Vision Inspired Framework for Facial Expressions Recognition. In: 2012 19th IEEE International Conference on Image Processing; 2012. p. 2593–6. doi:10.1109/ICIP.2012.6467429.
Khan RA, Meyer A, Konik H, Bouakaz S. Framework for reliable, real-time facial expression recognition for low resolution images. Pattern Recognit Lett 2013;34(10):1159–68.
Khan RA, Meyer A, Konik H, Bouakaz S. Saliency-based framework for facial expression recognition. Frontiers of Computer Science 2019b;13(1):183–98.
Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17.
Kumar V, Sinha D, Das AK, Pandey SC, Goswami RT. An integrated rule based intrusion detection system: analysis on unsw-nb15 data set and the real time online dataset. Cluster Comput 2020;23(2):1397–418.
Liao HJ, Lin CHR, Lin YC, Tung KY. Review: intrusion detection system: a comprehensive review. J Netw Comput Appl 2013;36(1):16–24. doi:10.1016/j.jnca.2012.09.004.
Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat. Rev. Genet. 2015;16(6):321.
Liu H, Motoda H. Computational methods of feature selection. CRC Press; 2007.
Louvieris P, Clewley N, Liu X. Effects-based feature identification for network intrusion detection. Neurocomputing 2013;121:265–73.
Mishra P, Pilli ES, Varadharajan V, Tupakula U. Out-vm Monitoring for Malicious Network Packet Detection in Cloud. In: 2017 ISEA Asia Security and Privacy (ISEASP). IEEE; 2017. p. 1–10.
What is a computer worm, 2020. https://us.norton.com/internetsecurity-malware.html, accessed: -10-04.
Mohammadi S, Mirvaziri H, Ghazizadeh-Ahsaee M, Karimipour H. Cyber intrusion detection by combined feature selection algorithm. Journal of information security and applications 2019;44:80–8.
Moustafa N, Slay J. Unsw-nb15: A Comprehensive Data Set for Network Intrusion Detection Systems (Unsw-nb15 Network Data Set). In: 2015 Military Communications and Information Systems Conference (MilCIS); 2015. p. 1–6. doi:10.1109/MilCIS.2015.7348942.
Moustafa N., Slay J.. A hybrid feature selection for network intrusion detection systems: Central points. 2017. ArXiv preprint arXiv:1707.05505.
Muna AH, Moustafa N, Sitnikova E. Identification of malicious activities in industrial internet of things based on deep learning models. Journal of Information Security and Applications 2018;41:1–11.
Neupane RL, Neely T, Calyam P, Chettri N, Vassell M, Durairajan R. Intelligent defense using pretense against targeted attacks in cloud platforms. Future Generation Computer Systems 2019;93:609–26.
Nskh P, Varma MN, Naik RR. Principle Component Analysis Based Intrusion Detection System Using Support Vector Machine. In: 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). IEEE; 2016. p. 1344–50.
Peterson LE. K-Nearest neighbor. Scholarpedia 2009;4(2):1883.
Prasad M, Tripathi S, Dahal K. An efficient feature selection based bayesian and rough set approach for intrusion detection. Appl Soft Comput 2020;87:105980.
Pudil P, Novovičová J, Kittler J. Floating search methods in feature selection. Pattern Recognit Lett 1994;15(11):1119–25.
Quinlan JR. Induction of decision trees. Mach Learn 1986;1(1):81–106.
Ring M, Wunderlich S, Scheuring D, Landes D, Hotho A. A survey of network-based intrusion detection data sets. Computers & Security 2019;86:147–67.
Ripley BD. Pattern recognition and neural networks. Cambridge university press; 2007.
Selvakumar B, Muneeswaran K. Firefly algorithm based feature selection for network intrusion detection. Computers & Security 2019;81:148–55.
Sethna J. Statistical mechanics: Entropy, order parameters, and complexity, vol. 14. Oxford University Press; 2006.
Syarif AR, Gata W. Intrusion Detection System Using Hybrid Binary Pso and K-nearest Neighborhood Algorithm. In: 2017 11th International Conference on Information & Communication Technology and System (ICTS). IEEE; 2017. p. 181–6.
Symantec. In: Tech. rep., Symentec Corporaton. Internet Security Threat Report (Vol. 24); 2019.
Tama BA, Comuzzi M, Rhee KH. Tse-ids: a two-stage classifier ensemble for intelligent anomaly-based intrusion detection system. IEEE Access 2019;7:94497–507.
Tavallaee M, Bagheri E, Lu W, Ghorbani AA. A Detailed Analysis of the Kdd Cup 99 Data Set. In: Computational Intelligence for Security and Defense Applications, 2009. CISDA 2009. IEEE Symposium on, IEEE; 2009. p. 1–6.
Tong S, Koller D. Support vector machine active learning with applications to text classification. Journal of machine learning research 2001;2:45–66.
Trevisan M, Giordano D, Drago I, Munafò MM, Mellia M. Five years at the edge: watching internet from the isp network. IEEE/ACM Trans. Networking 2020;28(2):561–74.
worm W.i.a.c.. 2020. https://www.rsaconference.com/industry-topics/blog/network-intrusion-methods-of-attack, accessed:-10-04.

Zhao S, Li W, Zia T, Zomaya AY. A Dimension Reduction Model and Classifier for Anomaly-based Intrusion Detection in Internet of Things. In: 2017 IEEE 15th Intl Conf on Dependable, Autonomic and Secure Computing, 15th Intl Conf on Pervasive Intelligence and Computing, 3rd Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech). IEEE; 2017. p. 836–43.

Engr. Anjum Nazir is currently pursuing PhD in Computer Science from National University FAST, Pakistan. He has more than a decade of industry experience in Cybersecurity. Currently he is working as Assistant Professor at Barrett Hodgson University, Karachi, Pakistan. His research interests include computer security, machine learning and cognitive computing.

Dr. Rizwan Ahmed Khan has received PhD in Computer Science from Université Claude Bernard Lyon 1, France in 2013. He has worked as postdoctoral research associate at Laboratoire d'InfoRmatique en Image et Systemes d'information (LIRIS), Lyon, France. Currently he is working as Professor at Barrett Hodgson University, Karachi, Pakistan. His research interests include machine learning, computer vision, image processing, pattern recognition, explainable AI (XAI) and human perception.