Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Intrusion Detection Using Combination of GA Based Feature Selection and Random Forest Machine Learning Supervised Approach

[1]Sachin Sharma, [2]Shubhashish Goswami, [3]Gesu Thakur
[1]Assistant Professor, Computer Science & Engineering, School of Computer Science & Engineering, DevBhoomiUttarakhand University, Chakrata Road, Manduwala, Naugaon, Uttarakhand 248007
[2,3]Associate Professor, Computer Science & Engineering, School of Computer Science & Engineering, DevBhoomiUttarakhand University, Chakrata Road, Manduwala, Naugaon, Uttarakhand 248007
[1]socse.sachin@dbuu.ac.in, [2]coe@dbuu.ac.in, [3]head.ca@dbuu.ac.in

Article Info
Page Number: 216 – 232
Publication Issue:
Vol. 71 No. 3s (2022)

Abstract
As of late, the fast advancement of web innovation brings numerous serious organization security issues connected to vindictive interruptions. Interruption Detection System is viewed as one of the huge procedures to defend the organization from both outer and inward assaults. In any case, with the quick development of the IoT organization, cyberattacks are additionally evolving rapidly, and numerous obscure sorts are appearing in the contemporary organization climate. Thusly, the productivity of conventional mark based and oddity based Intrusion Detection System is inadequate. We propose a clever Intrusion Detection System, which utilizes a developmental strategy based include choice methodology and a Random Forest-based classifier. The development based include selector utilizes an imaginative Fitness Function to choose the significant elements and decreases aspects of the information, which raise the Ture Positive Rate and lessen the False Positive Rate simultaneously. With extraordinary high precision in multi-order errands and remarkable abilities of taking care of commotion in gigantic information situations, the Random Forest strategy is broadly utilized in peculiarity identification. This examination proposes a structure that can choose all the more consistent highlights and further develop the order results as contrasted and different innovations. The proposed structure is tried and investigated UNSW-NB15 datasets and NSL-KDD datasets. Different measurable outcomes and itemized correlation with different strategies are introduced inside this article.

Keywords: Genetic calculation, network security, NSL-KDD, irregular timberland choice tree, UNSW-NB15

Article History
Article Received: *22 April 2022*
Revised: *10 May 2022*
Accepted: *15 June 2022*
Publication: *19 July 2022*

1. Introduction

Because of the world's third modern transformation, PCs and systems administration advancements detonated in our everyday existence. With the advantageous they brought, these innovations likewise left us with concern and dangers. Infection, Trojan, and Worms can undoubtedly infuse into our framework. Touchy data can be spilled or captured by cyberattacks. And this large number of dangers actually raise with the advancement of data innovation. The customary guard framework can distinguish a few assaults, however as they differed a little, they can scarcely be perceived. Consequently, the entire business is looking for new components that can precisely catch and block those dangers and assurance our framework in a working and safe climate.

Guard components can be ordered into Intrusion Prevention Systems (IPS) and Intrusion Detection System (IDS). Interruption Detection System, as an entry, typically works at the wilderness of the organization. As per various strategies, abuse location and oddity recognition are two primary classes in interruption discovery. Abuse identification utilizes known assault strategies that have been characterized ahead of time. The framework decides the presence of these assaults to accomplish the identification cycle, which is likewise called include location [1]. Abuse identification is based on the current element library or component data set. It can recognize the interruption designs kept in the mark data set with high precision. Notwithstanding, abuse recognition neglects to distinguish the zero-day assault. All in all, while there are assaults which not exist in the mark data set, this identification framework can scarcely catch them. At the point when a caution is raised, and that implies a recorded mark has been identified, however note that the arrangement of marks could contain equivocal layouts that can be brought about by an aggressor as well as a genuine client. Abnormality discovery doesn't depend on the mark information base. It examinations the organization traffic by computing the deviation from the client's way of behaving to the ordinary profile. Inconsistency discovery can address the dependence issue on the mark information base, yet this technique might identify the ordinary organization conduct as an interruption, and the deception rate is relativity high.

Commonly, an interruption discovery framework comprises of two parts that combine. The primary part chooses just the fundamental highlights, and the subsequent part is for arrangement and pursues proficient choices. To accomplish the best presentation, these parts should work alongside one another to play out a low tedious and high exactness result.

Information pre-handling goes for an indispensable step toward the start of the entire distinguish process. Choosing the huge data and elements from the dataset can diminish the components of the crude dataset, which generally prompts better execution. The Genetic Algorithm (GA) is roused by a characteristic transformative hypothesis set forward by Darwin. Hereditary Algorithm is a usually involved technique for tracking down an improved and great answer for search issue. Center administrators in GA are roused by natural cycles like hybrid, transformation, and determination [2]. Wellness capability is viewed as the main piece of Genetic calculations. The Fitness Function assesses each posterity chromosomes, and afterward simply the most noteworthy scored one can get by to the following developmental round. The weakness of the past proposed Fitness Function in GA-based highlight choice model is essentially involving the Accuracy and chosen include numbers as boundaries. Overlooking the high False Positive Rate (FPR) of the interruption location usually brings about a low True Positive Rate (TPR).

In ongoing many years, AI has been progressively utilized as one more crucial part of the advanced Intrusion Detection System. For the most part, AI can be isolated into administered and unaided learning. Directed learning is a useful asset in dissecting the high-layered information and sorting out the secret example behind these measurements. Regulated advancing likewise has areas of strength for an of ordering high-layered information into explicit classes. Accordingly, this innovation can be utilized to perceive pernicious ways of behaving in network traffic. In light of the huge, high-layered areas of strength for and straight traffic information, some traditional AI techniques, for example, Probability-based Bayesian, Decision Tree, and Support Vector Machine (SVM), are shown to be less compelling in the order task. The outcomes have low exactness yet a high False Positive Rate, and the "layered blast" issue is inclined to happen.

Irregular Forest (RF) is a managed learning calculation. After the preparation cycle with given highlights and characterization results, a RF model can be acquired to order new datasets. Among a wide range of regulated learning calculations, Random Forest enjoys specific benefits in exactness and preparing speed. Likewise, great commotion handling capacity and high solidness pursue the Random Forest a famous decision in the Intrusion Detection System. There are various variables to assess the presentation of Random Forest, including precision, review rate, running time, and so on. We propose a model that joins the Genetic calculation with the Random Forest calculation to arrive at the best outcomes. Plus, a recently planned Fitness Function, which adds FPR as a punishment boundary, intends to cut the misleading problem rate (FAR) and increment the TPR simultaneously. Besides, F1-score is additionally utilized for adjusting the heaviness of the precession rate and the review rate. We primarily advance the exactness and time intricacy of Random Forest through boundary change and information dimensionality decrease. A steady number of chosen elements and low choice time are likewise treated as a significant execution marker.

Rest of the paper is coordinated as follows: In Section II, related works are evaluated, in Section III, the subtleties of the proposed Intrusion Detection System is given. Segment IV talks about the exploratory outcome in UNSW-NB15 [2] dataset contrasted with the NSL-KDD [3] dataset. Ends and a few potential future upgrades on this work are introduced in Section V.

2. Literature Survey

There is a lot of past explores in the writing that examined the Intrusion Detection System. Denning D.E[4] at first proposed the theoretical model of the interruption location framework in 1987. This paper right off the bat involves interruption discovery as a security guard strategy of the PC framework. The model is free of a particular working framework, application climate, and framework weakness and interruption type. A structure can be a phenomenal instance of planning interruption location application frameworks. Albeit the review rules in the proposed model can be set off by other obscure elements that are not oddities ways of behaving. Also, the way that whether the model can distinguish the most interruption before serious harm is done in any case should be demonstrated. Wu et al. [5] work seriously in data set interruption, particularly in irregularity discovery in view of information mining, the creator likewise utilizes affiliation rules to a forward execution in light of Trie tree. Aumreesh et al. [6] give a survey that underlines different kinds of Intrusion Detection System, for example, abuse based, inconsistency based, have based, network-based and cross breed based. It fundamentally centers around irregularity based and conduct based

alongside specialist based innovation in genuine organization traffic. S. Northcutt et al. [7] analyze the advantages and disadvantages of the peculiarity location approach and abuse identification approach individually. The creator brings up that the downside of the oddity identification approach is that when the Intrusion Detection System encounters another way of behaving interestingly, it raises the caution, which might be a misleading positive. Likewise, the misleading negative rate and False Positive Rate and irregularity discovery are generally a lot higher than abuse identification.

L. Haripriya and M.A. Jabbar [8] give a survey of utilizing Machine Learning (ML) innovations in the Intrusion Detection System. They likewise examine applications into a framework with ML, and the point by point correlation of different methodologies for the Intrusion Detection System utilizing ML is given. This paper showed that It is moderately difficult to prepare the ML models while a specific measure of traffic information is inadequate or not accessible. A valuable interruption location framework model purposes Artificial Neural Network (ANN) is introduced by BasantSubba et al. [9]. One impediment in their methodology is that the model they proposed requires huge preparation time. In any case, the general discovery execution of the brain organization won't be debased by the disappointment of adding new specialists to the past one. Container Shi Tang et al.[10] depict Filter and covering, which is the most well-known highlight choice calculation in their work. A blend of two calculations is likewise contrasted and the Genetic Algorithm based choice strategy, then comes out a resolution that GA has a lot higher effectiveness than Filter and Wrapper calculation in choosing highlights. S. Aksoy et al. [11] and B. Kavitha et al. [12] depict a fundamental strategy for choosing the necessary subset of elements by utilizing the Genetic Algorithm. They accept highlight choice can dispose of repetitive things, as well as impressively affect building proficient arrangement framework in additional means. Ketan Sanjay Desale and Roshani Ade [13] propose an imaginative element determination strategy that utilizing a technique in view of numerical crossing point standard and hereditary calculation. Moreover, a reach sort of component choice methods, for example, IG, CAE, and CFS, are tried. Their results of the other two routinely utilized classifiers, J48, and Naive Bayes (NB) are looked at. These articles give a genuine instance of involving the Genetic Algorithm as an element selector.

Yi Aung et al. [14] foster an IDS for recognizing network conduct by utilizing K-Means and RF. Diminishing the CPU and memory utilization is likewise one of their core interests. Besides, the half breed model shows a better than the framework just utilizing a solitary Random Forest calculation, explicitly in identification rightness and order precision. In this work, 10% of the KDDCUP99 [15] dataset is utilized to affirm the model exactness. Yaping Chang et al. [16] apply Random Forest to choose significant highlights and SVM to further develop the arrangement result. Furthermore, just 14 elements (altogether 41 highlights) are chosen to arrive at a higher assault identification rate, likewise utilizing the KDDCUP99[16] dataset. An information mining based interruption discovery structure consolidating abuse and irregularity recognition, which likewise applies the RF, is proposed by Mohammad Zulkernine and Jiong Zhang [17]. They use testing methods and ideal contentions in their structure to expand the identification rightness of minority interruptions. Albeit, the principal deficiency of their work is that the half breed framework can be subverted assuming interruptions are considerably more than ordinary information in a dataset. Second, some serious level comparative interruptions can't be accurately distinguished as exceptions by the framework. Third, their tests explores still work on the KDDCUP99 [15] dataset, which is

obsolete and can't really address the cutting edge extensive organization traffic. M. Zhao et al. [18] use GA to improve boundaries of Support Vector Machine all the while. The model chooses enhanced highlights and best SVM boundaries by connecting them into one chromosome. In any case, Fitness Function in their developmental cycle just permits the exactness and the True Positive Rate to evaluate each chromosome. Additional registering time is likewise expected in each developmental step.

3. Proposed methodology

The general framework engineering of proposed GA-RF IDS system is displayed in Figure 1. We utilize the Genetic Algorithm based highlight determination technique to choose valuable elements. In the Genetic Algorithm, various mixes of elements are called chromosomes and each chromosome will be assessed by the Fitness Function. As indicated by the wellness esteem, unquestionably the most elevated scored chromosome can get by to the following advancement round. The new chromosome will supplant the former one in the complete chromosome pool, which is known as the underlying populace. At the point when transformative circle stops, moderately trademark highlights are chosen out as a result of the Genetic Algorithm. Additionally, Random backwoods s utilized for additional element choice and results order. Irregular Forest is viewed as an amazing asset while managing complex information, whether in double grouping or multi-class characterization.

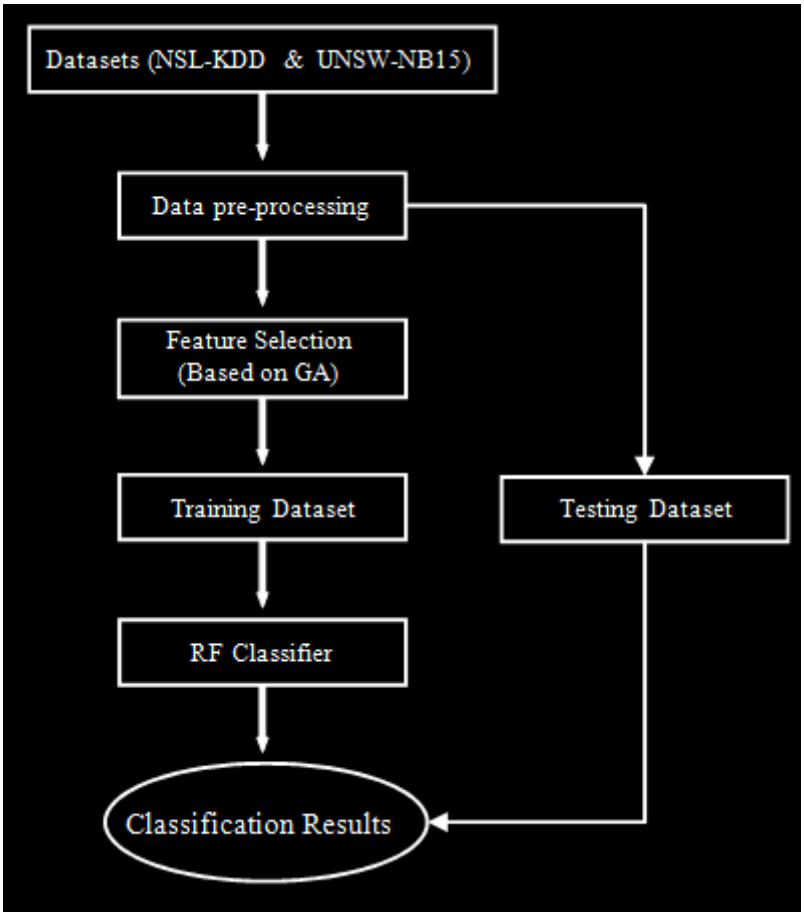


Figure 1: The design of the proposed GA-RF IDS

3.1 Brief Comparison of NSL-KDD and UNSW-NB15

As indicated by UNSW-NB15 [2] dataset, the NSL-KDD [3] dataset is viewed as an updated variant of the KDDCUP99[16]dataset. NSL-KDD[3] dataset eliminates the superfluous things in KDDCUP99 [16] and addresses the unbalancing issue among all records in both preparation dataset and testing dataset, which makes the discovery results more solid. NSL-KDD [3] preparing dataset covers 22 kinds of cyberattacks separated into four classes: Denial of Service (DOS), Probing Attack (PROBE), User to Root (U2R), and Remote to User (R2L). Table I presents the detail classifications of all assault types. Table I. likewise gives a short depiction of various classes. Figure 2 shows the circulation of typical traffic and 4 kinds of strange traffic. It plainly represents that the level of records in the dataset is contrarily corresponding to the quantity of records in every trouble level.

Table 1: Different types of attacks in KDD dataset

Class	Description	Attack Subclass
DoS	Restrict or deny a legitimate user request to a system	‘smurf’, ‘back’, ‘Neptune’,‘pod’, ‘teardrop’, ‘land’
PROBE	Identify and gather vulnerabilities exposed in a system or a network device	‘Ipsweep’, ‘nmap’, ‘portsweep’, ‘satan’
U2R	Pretend to be a legitimate user or gain unauthorized Root access to a system	‘loadmodule’, ‘buffer_overflow’, ‘rootkit’, ‘perl’
R2L	Gain unofficial local access from a remote machine	‘warezmaster’, ‘guess_password’, ‘imap’, ‘phf’, ’spy’, ‘multihop’, ‘ ftp_write ,‘wareaclient’

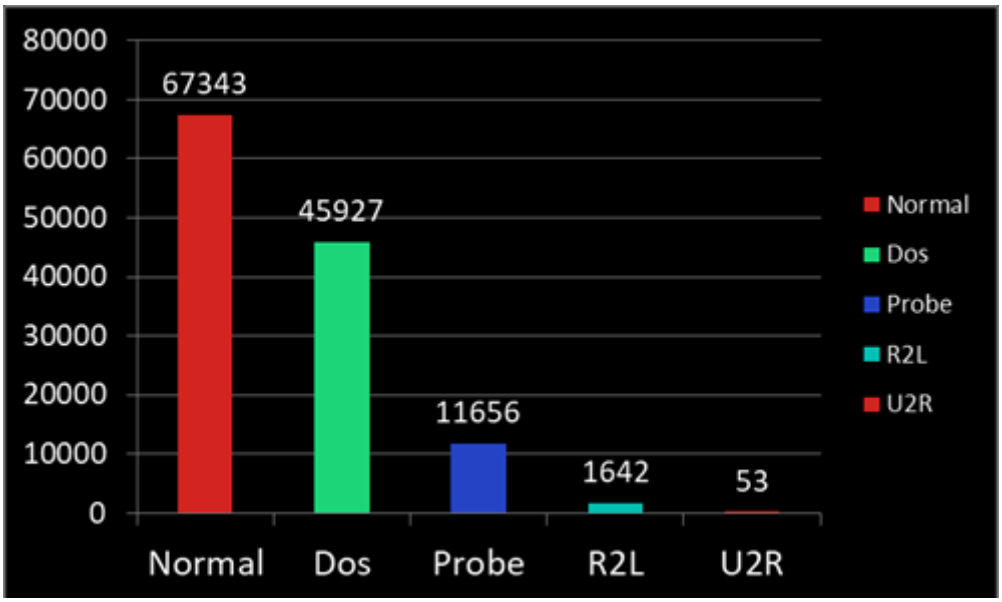


Figure 2: Circulation outline of class in KDD preparing dataset

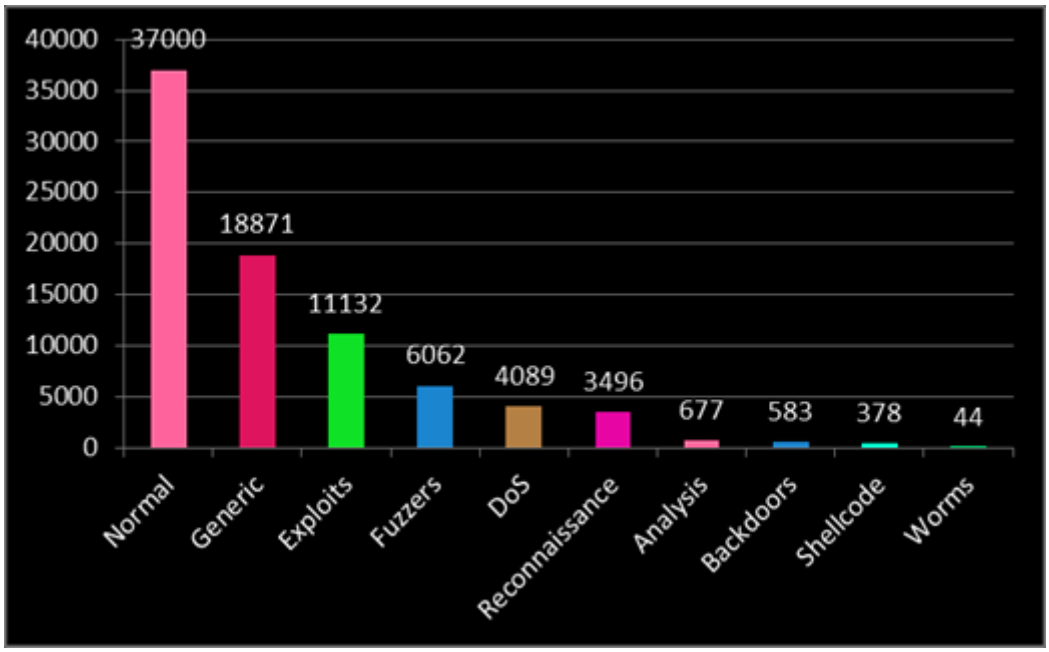


Figure 3: Assault circulation in UNSW-NB15 (preparing dataset)

Nonetheless, as indicated by UNSW-NB15 [2], NSL-KDD [3] dataset doesn't address the ongoing low impression assault situations. The UNSW-NB15[2] dataset is established to serve a comprehensive climate of the contemporary organization traffic, by laying out the engineered network utilizing the IXIA instrument, which can produce genuine current ordinary traffic and artificially strange traffic. UNSW-NB15 [2] dataset has 49 highlights however NSL-KDD [3] dataset just has 41 elements. Additionally, the additional highlights can be viewed as key elements and show benefits in past work. The records in UNSW-NB15 [2] are all arranged into ten gatherings, which areTypical, Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms. Table II gives itemized portrayal of all assault types in UNSW-NB15 [2]. Also, Figure 3 outlines the appropriation of the preparation dataset.

In this paper, the trial will be executed on each dataset, and results are introduced in Section IV.

Table 2: Data detailed information for different types of attacks in UNSW-NB15 dataset

Class	**Description**	**Attack Subclass**
Fuzzers	Attempts to suspend a program or network by providing randomly generated data.	24246
Analysis	Contains different attacks of port scan, spam and html files penetrations.	2677
Backdoors	A technique that bypasses system security to access a computer or its data.	2329
Dos	Restrict or deny a legitimate user request to a system	16353
Exploits	An attacker knows about a security problem in an operating system or software and uses the vulnerability to exploit that knowledge.	44525

where the Minimum worth and Maximum worth from every accessible datum xi addresses every piece of information.

3.2.Data Preprocessing steps

Pre-Processing changes the information in a uniform organization. It likewise used to eliminate the pointless information, which isn't needed for the proposed strategy and to finish the missing information.

3.2.1. 1-N encoding

To assess a model, UNSW-NB15 [2] and NSL-KDD [3] are utilized as benchmark dataset. Every one of the applicable investigations are performed utilizing the referenced datasets above. Additionally, utilizing just the material and pivotal elements to characterize the information source is fundamental. For improved consequences of component determination, NSL-KDD [3] dataset and UNSW-NB15 [2] dataset can't be utilized to prepare straightforwardly as the presence of non-numeric highlights in datasets. To beat this issue, non-numeric elements are changed over into numeric highlights by utilizing 1-n numeric coding. In this paper, every one of the non-numeric highlights like convention, administration, and banner have been changed over into numeric elements. For instance, the convention type highlight in NSL-KDD [3] comprises of 3 ostensible qualities which are tcp, udp, and icmp, the string esteem 'tcp' is supplanted by 1, 'udp' by 2 and 'icmp' by 3 et.

3.2.2. Normalization

Highlights in both datasets like "src-bytes", "dst-byts", "span" and so forth goes from 0 to 500000, which make the dataset unequal and ill suited to be handled. Unmatched records in the dataset will delude the classifier and result in a vague result. In this manner, these qualities or highlights ought to be standardized by utilizing the accompanying Max-Min (1) capability:

$$\frac{\boldsymbol{x_i - min}}{\boldsymbol{max - min}}$$

3.2.3. SMOTE Algorithms

Due to the minority of some particular cyberattack types, for example, R2L and U2R in NSL-KDD [3] dataset, Worms and Shellcode in UNSW-NB15 [2] dataset, standard classifier generally recognize those cyberattacks with extremely low exactness. Engineered Minority Oversampling Technique (SMOTE) is utilized to defeat this issue. Destroyed is viewed as a superior methodology in light of the Random Oversampling calculation. Basically the SMOTE Algorithm uses the K-Nearest Neighbor (KNN) to create the new examples, from a somewhat modest number of tests, planned to the first dataset. The calculation step is displayed beneath:

1) For each example x in the minority class $Smin$, compute the Euclidean Distance for every one of the rest in $Smin$ to get its K-Nearest Neighbor (KNN).
2) For every minority test x, arbitrarily select a few examples from its k-Nearest Neighbors, expecting that the chose neighbor is xn.
3) For each x , develop another example utilizing the accompanying recipe:

$$x_{new} = x + rand(0,1) \times |x - x_n|$$

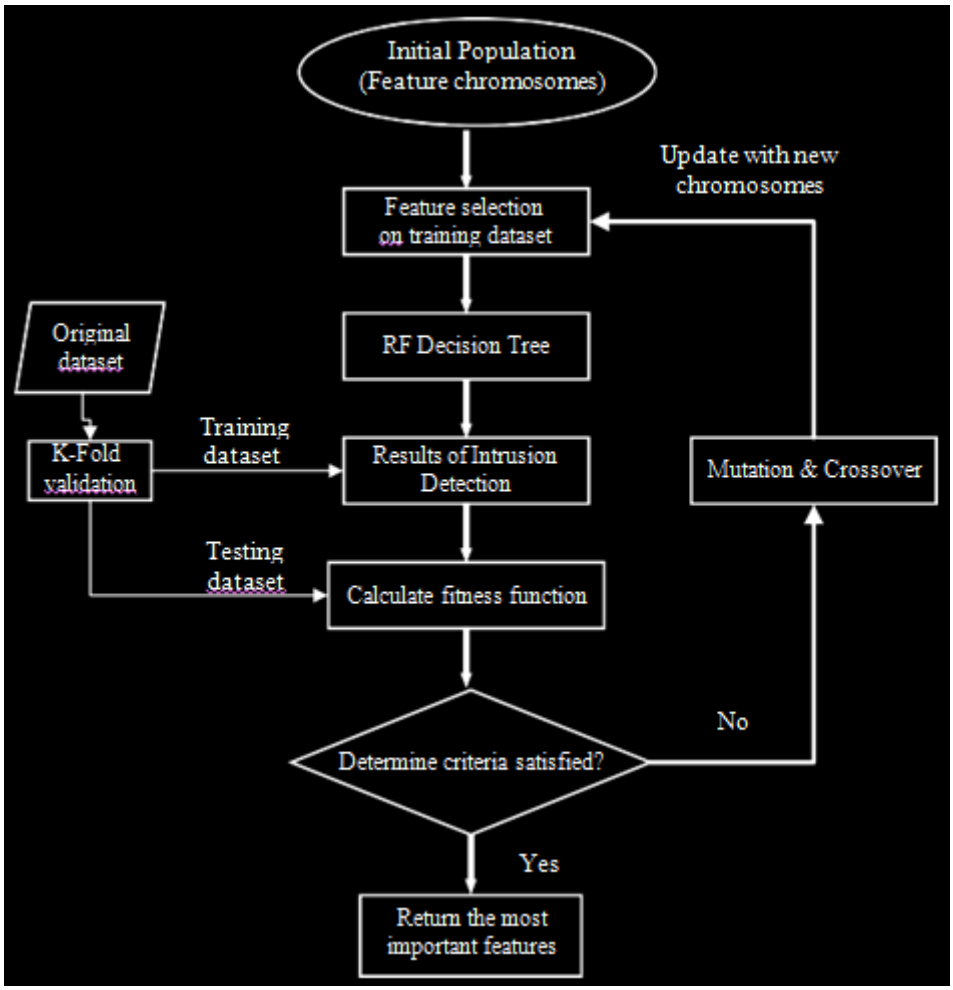


Figure 4: Work process graph of component determination

3.2.4. Hereditary Algorithm Based Feature Selection Method

In the proposed strategy, we utilize the Genetic Algorithm [19] as the foundation of the component determination technique. Figure 4 delineates the work process of our proposed include determination process. Introductory populace comprises of the component chromosomes. Highlights in NSL-KDD [3] and UNSW-NB15 [2] dataset are coded into twofold arrangement, for example, 110110111… 00101101. The chromosomes are produced arbitrarily. Then again, to incorporate however many classifications of assault as could reasonably be expected in both datasets, the quantity of the underlying populace is confined in 100 to 150. As indicated by the past investigates, the bigger the underlying populace it is, the more mind boggling the calculation is, and seriously registering time is required. In actuality, on the off chance that the above condition addresses the standardization cycle, starting populace is too little, the ideal execution of the calculation will be decreased, and it is not difficult to fall into the nearby ideal arrangement. Both unique datasets are isolated into preparing and testing datasets by utilizing the K-Fold approval strategy during the preparation interaction. Transformation rate and hybrid rate are kept steady in tests. In light of the arrangement results by the RF, Fitness Function assesses

3.2.5. False Negative (FN)

Incorrectly arrange the examples that initially have a place with positive classes into negative classifications.

Precision (3) is the level of information that is accurately anticipated. Exactness is determined as beneath:

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

each chromosome toward the finish of the cycle. At the point when any of the accompanying circumstances are fulfilled, the component extraction

$$f_1 - score = 2 \times \frac{precision \times recall}{precision + recall}$$

1) When the greatest number of preset cycles is reached, the inquiry is finished.
2) The greatest wellness esteem doesn't change for 10 progressive ages.

3.3. The Fitness Function

Wellness Function (8) is viewed as the most imperative and principal part of the hereditary calculation to assess a chromosome to get by. Toward the finish of each and every transformative step, the most elevated scored chromosome assessed by Fitness

$$fitness(c) = w_a + RF_{accur\ acy} + (w_b \times F_1 - score) - w_c \times FPR$$

In the information that anticipated being positive, the proportion of really certain information is called accuracy equation. In the really sure information, the proportion of information that anticipated being positive is called reviews equation. The equation of accuracy and review is displayed underneath:

Capability will supplant the lower scored one. A legitimate Fitness Function ought to save chromosomes with high fitting qualities and accelerate the iterative course of the hereditarycalculation. In addition, in the Intrusion identification framework situation, quite, not just the precision and the True Positive

Rate ought to be thought of, yet additionally the False Positive Rate ought to be remembered for the Fitness Function. Already, specialists select subsets with higher grouping precision

FPR equation is the pace of the bogus positive location determined by:also, less elements. In any case, they didn't take misleading identification in, so those component subsets would bring about higher deception rates, and the exhibition of the Intrusion Detection Systemwould corrupt.

3.5. Random Forest Decision Tree

Random Forest is viewed as an incorporated learning technique in light of decision trees. The Random Forest was proposed by Leo Breiman in 2001 to consolidate the packed away coordinated learning hypothesis [20] with the arbitrary subspace strategy [21]. RF is a notable classifier for directed learning. In the RF choice tree, every hub is grouped on the foundations of ideal component determination. This interaction go on until we arrive at the end measures. Every hub sorted as the moderately same sort of information. The quantity of votes decides the grouping result. The most casted a ballot leaf hub is viewed as the classification of the example. The still up in the air by way moving from root hub to leaf hub. The opposition of RF to commotion and anomalies settle numerous presentation issues as well as give us great dependability. The Non-Parametric nature of RF pursues it a superior decision for the order of high-layered information.

3.6. Proposed Fitness Function

We propose a creative Fitness Function (8) which utilizes three boundaries named Accuracy, F1-score and False Positive Rate (FPR) to assess every chromosome include.

1) True Positive (TP): Classify the examples that initially have a place with positive classifications into positive classifications.
2) True Negative (TN): Classify the examples that initially have a place with negative classifications into negative classifications.
3) False Positive (FP): Incorrectly order the examples that initially have a place with negative classifications into positive classes.

The recipe of the Fitness Function Eq. (8) is as underneath:

condition ()

In the proposed Fitness Function Eq. (8), wa loads for precision of Random Forest Decision Tree, wb loads for $F1 - score$ and wc loads for False Positive Rate. The $F1 - score$ is a proportion of test precision. It is the consonant mean of accuracy and review, which considers both accuracy and review of the order model to process. $F1 - score$ arrives at its best worth at 1 (wonderful accuracy and review) and most obviously terrible at 0.

We expect the high False Positive Rate prompts a False caution, which could make the Intrusion Detection System judge ordinary organization traffic to a malignant one. We propose to at the same time expand the TPR and decline the FPR. Thusly, we treat the False Positive Rate as a punishment boundary in our Fitness Function, and that implies a high False Positive Rate makes a lower worth of the entire Fitness Function. Each chromosome is assessed by the proposed Fitness Function toward the finish of each and every circle displayed in Figure 4 and just high scored chromosome can get by to next developmental round.

4. Experimental setup details, results and discussion

The testbed of our proposed technique is a Windows stage based PC of equipment setup having Intel Core i7-eighth era in 2.3GHz and 8 GB RAM. DEAP system (rendition 1.28) was utilized to play out the Genetic Algorithm under Python. Detail boundaries of the Hereditary Algorithm and Fitness Function are displayed in Table III.

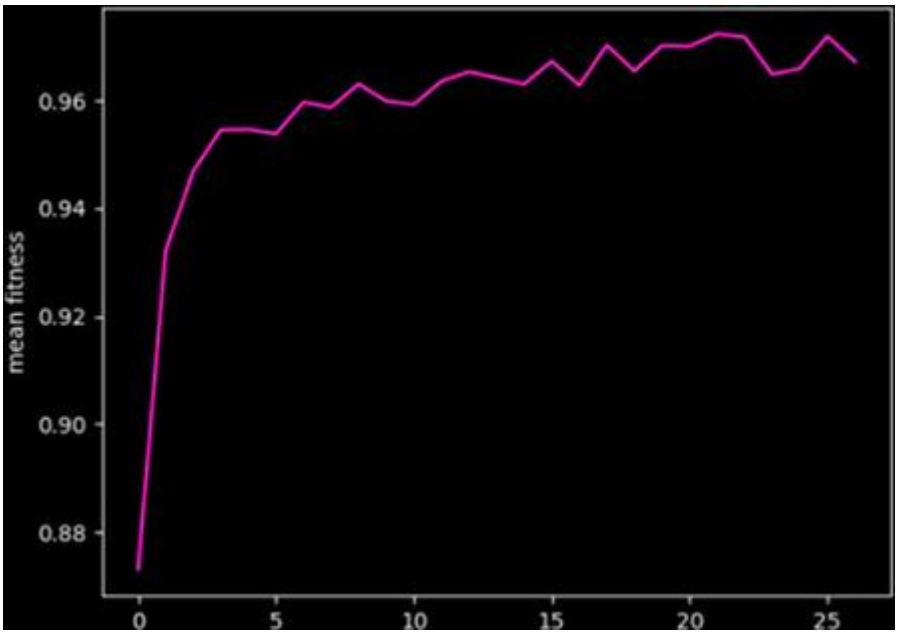
Table 3: Parameters details in GA and fitness function of the model

Evolution parameters	
Parameters Name	Number
Initial population	150
Mutation rate	0.01
Crossover rate	0.75
Selection type	Roulette wheel selection
Crossover type	Two-point crossover
Fitness Function parameters	
wa	0.6
wb	0.4
wc	100

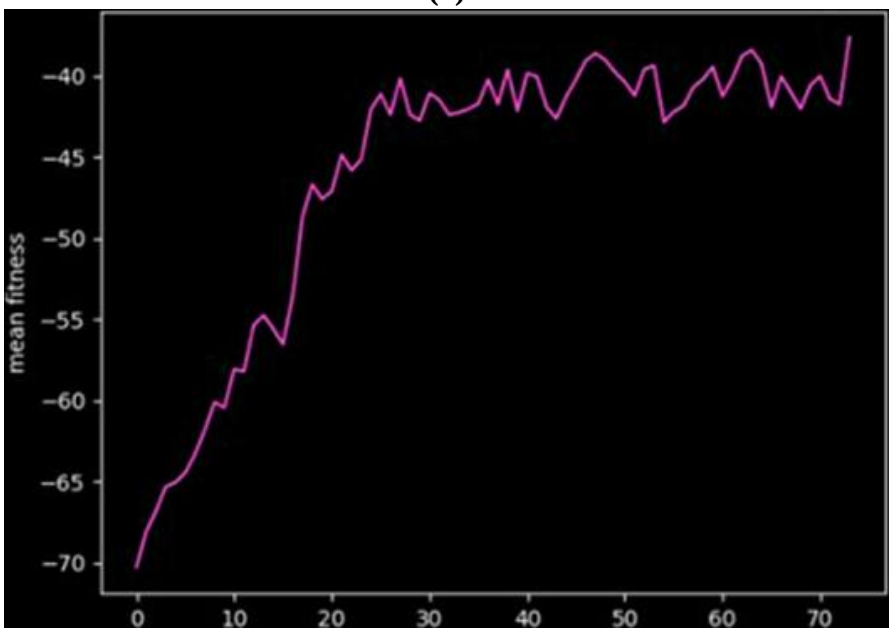
As per the Fitness Function (8), wellness score can be impacted by various upsides of boundaries. After many trials, we found mean wellness esteem arrived at its pinnacle when

wa =0.6 and wb =0.4. We characterized the DEAP system to be an issue of Maximization and set wc=100 to intensify the heaviness of FPR to accomplish the best outcome. The general mean wellness esteem in NSL-KDD [3] dataset (a) and UNSW-NB15 [2] dataset (b) are displayed in Figure 5. The X-hub addresses the N×10th age of the circle, and Y-pivot addresses the mean wellness upsides of every age. As displayed in Figure 5, with the course of chromosome choice, the capability diagram shows a vertical pattern and afterward continuously straightens out, and that implies right scored chromosomes are protected in populace and significant chose highlights are gradually turning out to be consistent.

F1 − $score$, Accuracy, Recall, Precision, and FPR for both NSL-KDD [3] Train dataset and UNSW-NB15 [2] Train dataset in double arrangement are displayed in Figure 6. What's more, the ROC Curve for both datasets is displayed in Figure 7.



(a)



(b)

Figure 5: (a) Mean wellness in NSL-KDD. (b) Mean wellness in UNSW-NB15

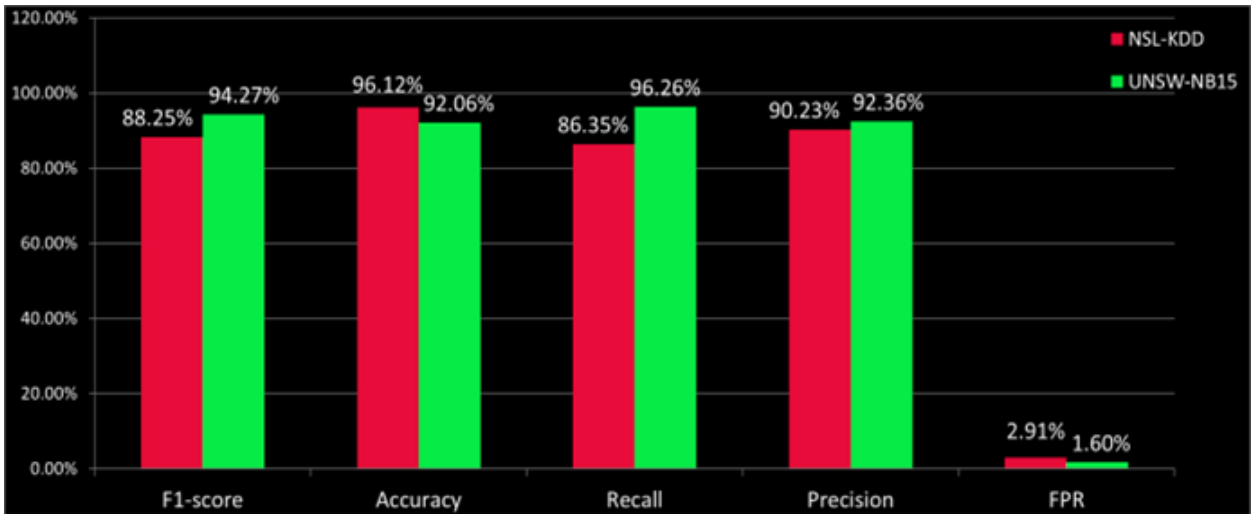
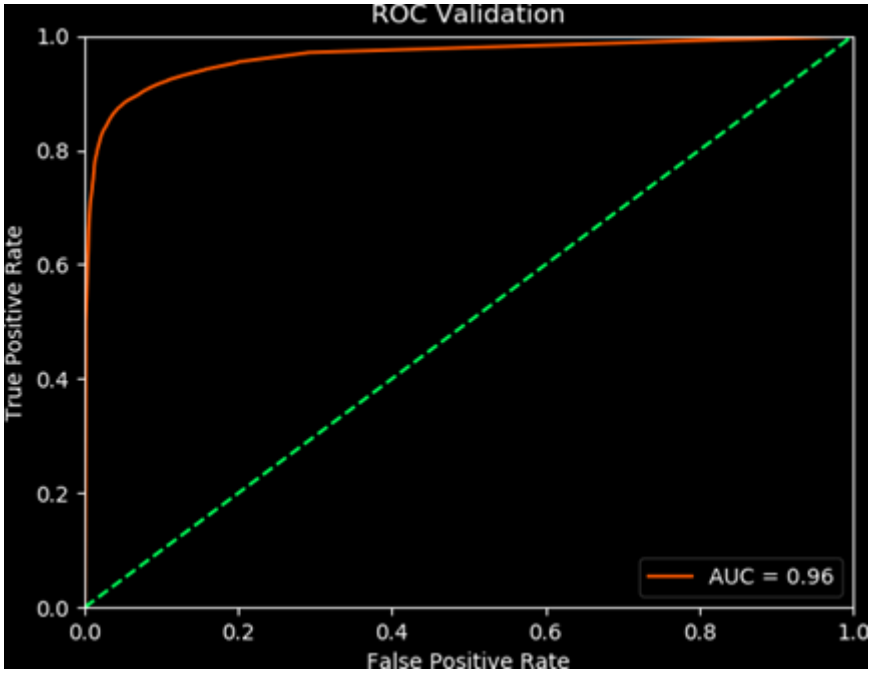
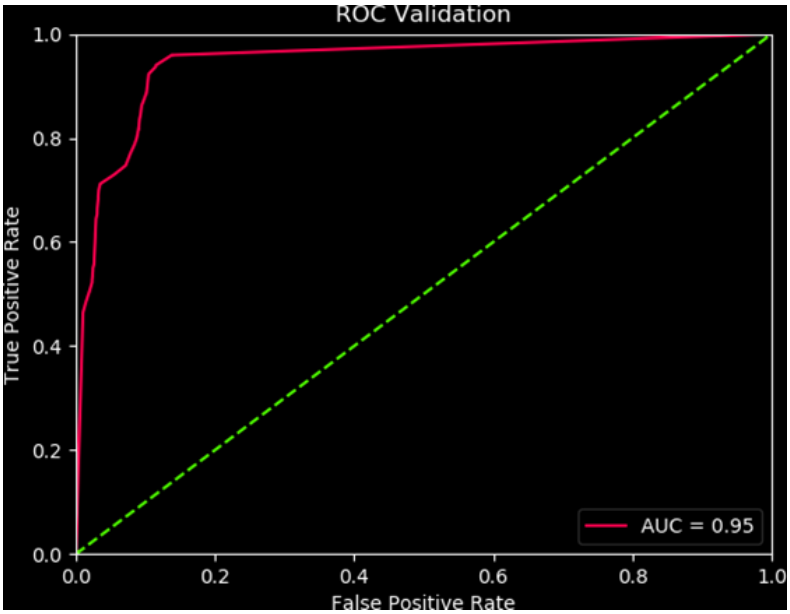


Figure 6: Assess list for NSL-KDD and UNSW-NB15

Consolidating the element choice outcomes from the Genetic Algorithm and the Random Forest, Table IV gathers significant highlights for double arrangement and multi-class grouping in NSL-KDD [3] dataset and the UNSW-NB15 [2] dataset.



(a)



(b)

Figure 7: (a) ROC Curve for NSL-KDD. (b) ROC Curve for UNSW-NB15

Exactness and AUC can mirror the ability to group of the classifier. Because of the lopsidedness issue in NSL-KDD [3] testing dataset and UNSW-NB15 [2] testing dataset, AUC number can show the ability to group of the structure all the more unbiasedly. Execution in NSL-KDD [3] dataset and UNSW-NB15 [2] dataset is displayed in Table V.

Table 4: Selected features form NSL_KDD and UNSW_NB15 datasets

Result with NSL-KDD dataset		
Class	Numbers	Selected Features
Normal	12	1,2,3,4,5,6,7,10,11,12,30,36
DOS	14	29,30,23,5,4,38,6,35,25,24,36,26,39,2
PROBE	15	36,5,35,33,12,2,40,37,6,3,32,27,41,30,26
R2L	11	23,3,5,33,12,24,10,36,32,37,6
U2R	12	1,24,33,32,36,23,6,10,14,17,5,13
Result with UNSW-NB15 dataset		
Normal	9	27,3,41,35,36,10,31,2,18
Reconnaissance	14	41,36,27,31,8,7,28,33,10,34,40,6,15,13
Exploits	8	41,31,27,28,7,2,13,14
Fuzzers	11	10,3,4,41,36,31,28,29,45,46,47
Worms	9	41,36,7,3,39,27,29,31,10
Generic	9	35,7,3,2,27,9,11,33,46
Shellcode	7	36,44,33,34,8,10,45
Dos	12	2,27,41,36,31,7,12,3,10,43,45,47
Analysis	7	27,2,35,7,12,28,36
Backdoor	10	35,27,2,33,14,9,17,25,23,42

Table 5: Method performance with NSL_KDD and UNSW_NB15 dataset

Result with NSK-KDD Testing dataset			
class	Accuracy (%)	FPR (%)	AUC
Normal	96.12	2.91	0.96
Dos	97.31	1.49	0.98
PROBE	94.58	1.39	0.96
R2L	90.79	0.07	0.92
U2R	88.21	0.11	0.85
Result with UNSW-NB15 Testing dataset			
Normal	92.06	1.6	0.95
Reconnaissance	91.24	0.6	0.94
Exploits	94.69	1.62	0.95
Fuzzers	86.04	2.1	0.91
Worms	98.81	1.14	0.98
Generic	99.25	0.39	0.99
Shellcode	95.43	2.49	0.97
Dos	94.03	2.06	0.9
Analysis	90.35	0.82	0.87
Backdoor	86.92	2.81	0.82

Contrasted and different innovations, our proposed GA-RF Intrusion Detection System shows more adequacy testing on NSL-KDD [3] dataset and the UNSW-NB15 [2] dataset, which can profoundly address the ongoing organization traffic state. The presentation correlation is displayed in Table VI.

Table 6: Comparison of performance with proposed method with other methods

Method	**Accuracy (%)**	**FPR (%)**	**DATASET**
ANN [9]	98.86(three layer)	-	NSL-KDD
GA-based J48[13]	91.86	-	NSL-KDD
GA-based NB [13]	89.5	-	NSL-KDD
K-mean RF [14]	99.8	-	10% of KDD'99
RS-GA-SVM [16]	88.2	2	KDD'99
RF-based IDS [17]	94.7	2	KDD'99
GA-RF (Proposed)	96.12	2.91	NSL-KDD
GA-RF (Proposed)	92.06	1.6	UNSW-NB15

5. Conclusion

In this paper, we propose an original Genetic Algorithm based highlight choice Intrusion Detection System which utilizes the Random Forest classifier. This developmental calculation is utilized to choose ideal elements for the interruption dataset. Another Fitness Function for the Genetic Algorithm is intended to accomplish high TPR and low FPR simultaneously. We likewise propose an upgraded Random Forest classifier, which consolidating the Genetic Algorithm based highlight determination strategy, and showing higher precision and AUC in both twofold class characterization and multi-class grouping. FPR is likewise lower than different strategies. Two benchmark datasets, NSL-KDD [3]

dataset and UNSW-NB15 [2] dataset, are run in tests, however the UNSW-NB15 [2] dataset is considered as a more powerful portrayal of current organization traffic. Destroyed calculation is utilized for both NSL-KDD [3] preparing dataset and UNSW-NB15 [2] preparing dataset, which can astoundingly further develop the recognition accuracy of minority assaults. The fundamental benefit of our proposed system is that it further develops the discovery exactness of the exemplary Random Forest by choosing fundamental highlights and diminishing preparation time.

Future work will be centered around GPU processing to abbreviate preparing time. Some profound learning calculations will likewise be considered to further develop location precision further.

References:

[1] W. K. Lee, S. J. Stolfo, and K. W. Mok, "A data mining framework for building intrusion detection models," in Proc. the 1999 IEEE Symposium on Security and Privacy, 1999, pp. 120-132.

[2] N. Moustafa, "UNSW-NBI5: A comprehensive data set for network intrusion detection systems (UNSW-NBI5 network data set)," in Proc. Military Communications and Information Systems Conference (MiICIS), 2015.

[3] S. Revathi and A. Malathi, "A detailed analysis on NSL-KDD dataset using various machine learning techniques for intrusion detection," International Journal of Engineering Research & Technology (IJERT), vol. 2, pp. 1848-1853, 2013.

[4] D. E. Denning, "An intrusion detection model," IEEE Transactions on Software Engineering, vol. 13, no. 2, pp. 222-232, 1987.

[5] W. Gongxing and H. Yimin, "Design of a new intrusion detection system based on database," in Proc. 2009 International Conference on Signal Processing Systems, 2009, pp. 814-817.

[6] A. K. Saxena, S. Sinha, and P. Shukla, "General study of intrusion detection system and survey of agent based intrusion detection system," in Proc. 2017 InternationalConference on Computing, Communication and Automation (ICCCA), 2017, pp. 421-471.

[7] S. Northcutt and J. Novak, "Network intrusion detection," IEEE Network, vol. 8, no. 3, pp. 26-41, 2003.

[8] L. Haripriya and M. A. Jabbar, "Role of machine learning in intrusion detection system: Review," in Proc. 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), 2018, pp. 925-929.

[9] M. B. Subba, S. Biswas, and S. Karmakar, "A neural network based system for intrusion detection and attack classification," in Proc. 2016 Twenty Second National Conference on Communication (NCC), 2016, pp. 1-6.

[10] P. S. Tang, X. L. Tang, and Z. Y. Tao, "Research on feature selection algorithm based on mutual information and genetic algorithm," in Proc. 2014 11th International Computer Conference on Wavelet Active Media Technology and Information Processing, 2014.

[11] S. Aksoy, "Feature reduction and selection," Department of Computer Engineering, Bilkent University, 2008.

[12] B. Kavitha, S. Karthikeyan, and B. Chitra, "Efficient intrusion detection with reduced dimension using data mining classification methods and their performance comparison," in Proc. International Conference on Business Administration and Information Processing, 2010, pp. 96-101.

[13] K. S. Desale and R. Ade, "Genetic algorithm based feature selection approach for effective intrusion detection system," in Proc. 2015 International Conference on Computer Communication and Informatics (ICCCI), 2015, pp. 1-6.

[14] Y. Y. Aung and M. M. Min, "An analysis of random forest algorithm based network intrusion detection system," in Proc. 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2017, pp. 127-132.

[15] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A detailed analysis of the KDDCUP99 dataset," in Proc. IEEE International Conference on Computational Intelligence for Security & Defense Applications, 2009.

[16] Y. Chang, W. Li, and Z. Yang, "Network intrusion detection based on random forest and support vector machine," in Proc. 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), 2017, pp. 635-638.

[17] J. Zhang, M. Zulkernine, and A. Haque, "Random-forests-based network intrusion detection systems," IEEE Transactions on Systems, Man, and Cybernetics, vol. 38, no. 5, pp. 649-659, Sept. 2008.

[18] M. Zhao, C. Fu, L. Ji, K. Tang, and M. Zhou, "Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes," Expert Systems with Applications, vol. 38, no. 5, pp. 5197-5204, 2011.

[19] K. Deb, An Introduction to Genetic Algorithms, pp. 293-315, 1999.

[20] S. W. Kwok and C. Carter, "Multiple decision trees," Machine Intelligence & Pattern Recognition, vol. 4, pp. 327-335, 2013.

[21] T. K. Ho, "The random subspace method for constructing decision forests," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pp. 832-844, Aug. 1998.