OPEN ACCESS

International Journal Of Modern Engineering Research (IJMER)

An Efficient approach for Network Intrusion Detection using Artificial bee colony optimization and Ensemble Classifier

Ruchi Mulay
M. Tech. Scholar (CSE)
ruchimulaycse@gmail.com
Columbia Institute of Engineering and Technology, Raipur C.G. India

Gargi Shankar Verma
Asso. Prof. (CSE)
gargi1981@gmail.com
Columbia Institute of Engineering and Technology, Raipur C.G. India

Abstract— *"Network intrusion detection system (NIDS)" monitors traffic on a network looking for doubtful activity, which could be an attack or illegal activity. The intrusion detection techniques based upon data mining are generally plummet into one of two categories: misuse detection and anomaly detection. In misuse detection, each instance in a data set is labeled as 'normal' or 'intrusive' and a learning algorithm is trained over the labeled data. As dataset used for NIDS dimension is very high hence applying Metaheuristic based (ABC) feature selection and Ensemble machine learning classifier may improve the accuracy, we have achieved average 99% accuracy.*

Keywords— *ID, NIDS, GWO , KDD, ML, DM*

I. INTRODUCTION

Intrusion detection is the process of classifying and responding to malevolent activities targeted at computing and network resources". An intrusion detection, also known as attack, mentions to a sequence of actions by use of which an intruder endeavors to gain control over a system. Network security is of vital significance in the present data communication. Programmers and interlopers can make numerous effective endeavors to cause the crash of the systems and web benefits by unapproved interruption.

"Network intrusion detection system (NIDS)" monitors traffic on a network looking for doubtful activity, which could be an attack or illegal activity. The intrusion detection techniques based upon data mining are generally plummet into one of two categories: misuse detection and anomaly detection. In misuse detection, each instance in a data set is labelled as 'normal' or 'intrusive' and a learning algorithm is trained over the labelled data. An intrusion detection system intention to differentiate between intrusion activity and normal actions. In doing so, conversely, an IDS can familiarize classification errors. A false positive is a gentle input for which the system speciously raises a notification. A false negative, in contrast, is a malevolent input that the IDS miscarries to report. The appropriately classified input data are typically mentioned to as true positives (suspicious attacks) and true negatives (normal traffic). There is a natural trade-off between distinguishing all malevolent events (at the outlay of floating alarms too repeatedly, i.e., having high false positives), and missing anomalies (i.e., having high false negatives, but not give out many false alarms).

In this paper we have proposed the use of wrapper feature selection approach using Artificial Bee Colony algorithm. In general, feature selection techniques are split up into two categories: filter and wrapper approach. Techniques that are not dependent on classifiers and work directly on data belong to the filter (heuristic) category. This strategy is often used to identify correlations between variables. The most trending filter methods are Chi-Square, gain ratio, information gain, ReliefF and hybrid ReliefF. Wrapper feature selection ap- proaches is a meta heuristic procedure, that uses classifiers like k-Nearest-Neighbor (kNN), Support Vector Machine (SVM), etc to detect interaction between variables. These wrapper methods can produce significantly more correct answer even though they take more time to compute than filter approaches. Moreover, filter-based approaches not compelling for high dimensional data sets. Lately, meta-heuristic-based approaches turned out to be successfully utilised to overcome FS bottle- necks [2]. Based on their source of inspiration, meta heuristic-based methods can be classified as chemical-based, physical-based, swarm-based, human-based, etc. The most well-known meta- heuristic-based method, known as swarm intelligence, is based on the fact that most animals live in groups, exhibit similar behaviors, and devote their majority of time to look for food for survival (cites: b3, b4). As there are several works relating to the feature selection method based on meta-heuristics. Below is an overview of the relevant works for the wrapper- based FS approach.

Further in section-II different literatures discussed, in section-III we have elaborated proposed method, in section-IV proposed approach has been validated and in last section we have concluded our research.

II. LITERATURE SURVEY

Cosimo Ieracitano et. al. Elsevier 2020 proposed IDS combines data analytics, statistical techniques with recent advances in machine learning theory to extract optimized and more correlated features. The validity of the proposed IDS is tested using the benchmark NSL-KDD database. Experimental results show that the designed IDS achieves better classification performance as compared to deep and conventional shallow machine learning as well as recently proposed state-of-the-art techniques.

Eesa, Orman, and Brifcani (2015) developed a feature selection model that utilizes the combination of ID3 classifier algorithm and bees algorithm. The model known as ID3-BA is designed to optimize the selection of the required subset of features in IDS. In this model, the bees algorithm is used for the generation of the required subset of features while the ID3 algorithm is used to construct the classifier. The ID3-BA model utilizes the KDD Cup99 (Knowledge Discovery and Data mining tools) dataset that contains 41 features for training and testing purposes.

Abhinav Kumra et. al. said that proposed method was triumphantly tested on the data log files and the database. The results of the proposed testimony are producing more accurate and irrelevant sets of patterns and the discovery time is less than other approach. As a naïve Bayesian network is a restricted network that has only two layers and assumes complete independence between the information nodes. This poses a limitation to this research work. In order to alleviate this problem so as to reduce the false positives, active platform or event based classification may be thought of using Bayesian network [OJCST 2017].

Kathleen Goeschel has shown that high accuracy may be maintained while reducing false positives using the proposed model composed of SVMs, decision trees, and Naïve Bayes. First, the SVM is trained based upon a new binary classification added to the dataset to specify if the instance is an attack or normal traffic. Second, attack traffic is routed through a decision tree for classification. Third, Naïve Bayes and the decision tree will then vote on any unclassified attacks. Future work is to write this model as a Java class such that it may be applied in other systems or applications. Further future work is to test this model on other network traffic data sets for more in-depth analysis [IEEE 2016].

Anna L. Buczak et. al. describes a focused literature survey of machine learning (ML) and data mining (DM) methods for cyber analytics in support of intrusion detection. Short tutorial descriptions of each ML/DM method are provided. Based on the number of citations or the relevance of an emerging method, papers representing each method were identified, read, and summarized. Because data are so important in ML/DM approaches, some well-known cyber data sets used in ML/DM are described. The complexity of ML/DM algorithms is addressed, discussion of challenges for using ML/DM for cyber security is presented, and some recommendations on when to use a given method are provided [IEEE 2016].

Bane Raman Raghunath et. al. focuses on two specific contributions: (i) an unsupervised anomaly detection technique that assigns a score to each network connection that reflects how anomalous the connection is, and (ii) an association pattern analysis-based module that summarizes those network connections that are ranked highly anomalous by the anomaly detection module. Fig.-1 shows the accuracy comparison of different approaches.

***Table 1** Comparison*

S. No.	Author /Year Publication	Description	Algorithm Used and Performance
1	DikshantGupta et. Al/ IEEE 2016	Paper includes the implementation of different data mining algorithms including Linear regression and K-Means Clustering to automatically generate the rules for classify network activities.	Linear regression-80% Accuracy K-Means Clustering-67% Accuracy
2	Upendra et. al./ IJCTCM 2012	Compared the performance measure of five machine learning classifiers such as Decision tree J48, BayesNet, Naive Bayes and ZeroR.	J48, BayesNet, ZeroR Approx. all algorithms gave 80% accuracy

3	A J M Abu Afza et. al/ IEEE 2013	Author presented a Dependable Network Intrusion Detection System (DNIDS) by integrating detection method with an intrusion-tolerant mechanism. To address the detection issue	CSI-KNN Algorithm 76% Accuracy
4	Neethu B/ IJCSNS 2013	Paper applies PCA for feature selection with Naïve Bayes for classification in order to build a network intrusion detection system	PCA 85% Accuracy but not time efficient
5	Upendra/ ijettcs 2013	In this paper author compared the performance of intrusion detection.	NB , C4.5 Accuracy 76%
6	Mostert, et. al. 2021 [1]	Developed an enhanced hybrid metaheuristic approach using grey wolf optimizer (GWO) and whale optimization algorithm (WOA) and reduces drawbacks of both algorithms	Multi-objective formulation of the FS will be addressed.
7	Mafarja, Majdi et. al. 2020 [2]	Proposed algorithm uses pigeon inspired optimizer for FS over IDS dataset. PIO solved optimization problems such as air robot path planning.	The proposed algorithm discretization process shows a faster convergence than the traditional way, as it uses cosine similarity.
8	Emary 2016 [3]	Novel binary gray wolf optimization (bGWO) is proposed for the feature selection task.	Proposed algorithm must be evaluated for larger number of iterations.

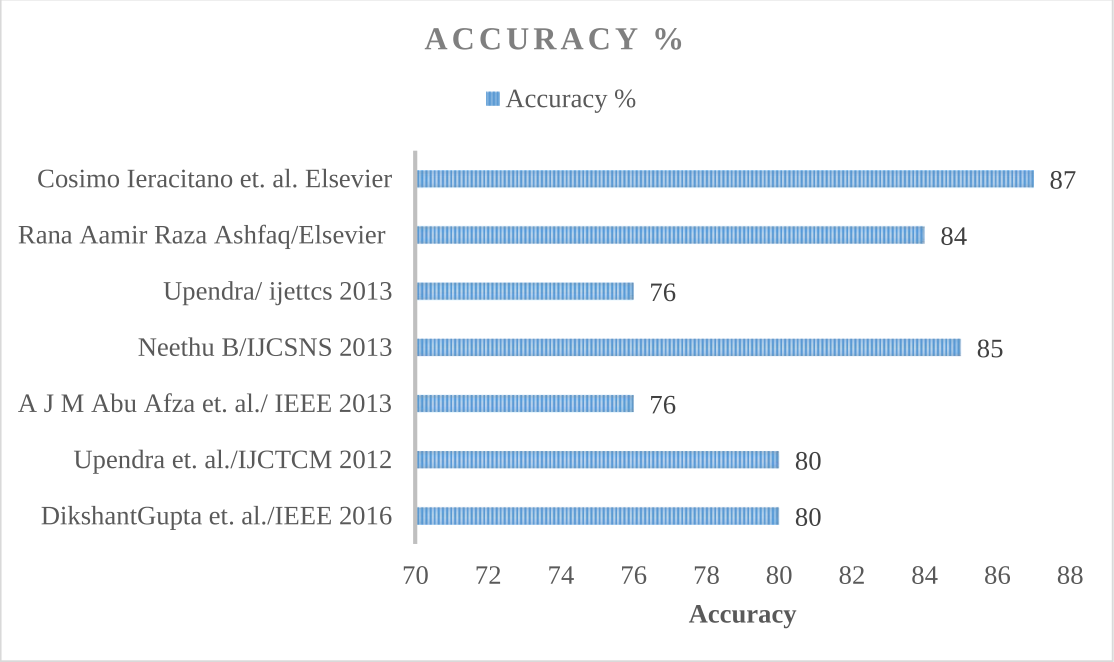

Figure 1 Accuracy Comparison

III. PROPOSED APPROACH

After gone through numerous literatures, come across conclusion that there is need of an efficient Network intrusion detection algorithm which should have higher value of precision i.e., algorithm should have high accuracy.

- Proposed system will use Ant bee colony algorithm for feature selection.
- To increase the accuracy proposed the use of ensemble machine learning classifier.

in this research Off-line anomaly detection using machine learning will be taken into account because from literature review section it can be concluded that still there is need if intelligent NIDS system will detect intruder efficiently with high precision value. There is some bottleneck identifier in earlier algorithm which are as follows:

- Due to very large dataset algorithm needed which should be time efficient.
- Earlier system having FNR (False Negative Rate) is high.
- Accuracy of anomaly detection is less.
- Feature selection problem is one of the most significant issues in data classification. The purpose of feature selection is selection of the least number of features in order to increase accuracy and decrease FPR, and the cost of the data classification.
- Various intrusion detection methods are proposed in the previous literatures but their performance metrics such as Accuracy, F-Score, True Positive Rate (TPR), False Positive Rate (FPR), Precision remains a problem, as generally improving accuracy also increases FPR.

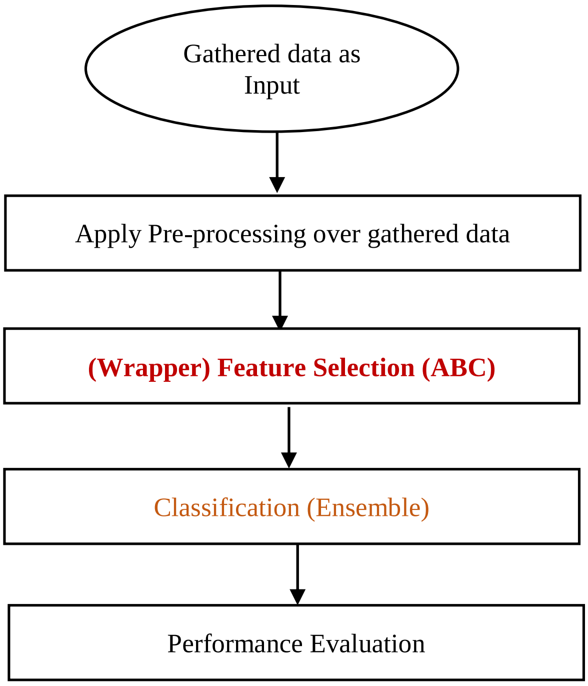
Gathered data as Input
Apply Pre-processing over gathered data
(Wrapper) Feature Selection (ABC)
Classification (Ensemble)
Performance Evaluation

Figure 2 Proposed Framework

Fig.-2 shows the proposed algorithm framework, in which after pre-processing of input data, it has been passed to Artificial Bee Colony (ABC) meta heuristic optimization algorithm for feature selection. Further reduced data passed to Ensemble classifier to validate the proposed prediction model.

A. Artificial Bee Colony

Artificial bee colony consists of 2 components: bees (the process of decision making for selection of position of food) and food source (position in space). It basically defines 2 types of behavior: nectar source identification and food source abandonment. Also, the bees are categorized into 3 groups- employed bees, onlooker bees and scout bees [13].

The onlooker bees wait at the hive to decide the choice of the source of food. The employed bees go to the earlier visited food source while the scout bees search around the hive. In the whole process of honeybee colonies, some of the bees randomly search for food around 2 the hive. After a food source is found, these bees bring back some nectar to the hive, deposit it and share the information of the nectar of food sources with the bees waiting in hive at dancing area.

The bee colony now enters a cycle of iterations and the following steps are followed: (1) after the information is shared, employed bee either becomes onlooker after the food source is abandoned or continue to forage the site visited earlier; (2) onlookers in hive will follow employed bees simultaneously based on the information received to forage further on some memorized sources of food; and (3) some of the scouts will start random search spontaneously [13].

The food sources are randomly initialized using the below expression:

$$a_i = l_k + rand(0,1)^*(u_k - l_k) \qquad (1)$$

Where is the solution in the population, is a randomly selected parameter index, and are upper and lower bound constraints for the solution search space of objective function to be optimized.
Important stage for ABC algorithm is information sharing which is achieved by influencing onlooker's behaviour which selects food source based on the probability:

$$Pb_i = fitness_i / \sum_{n=1}^{SN} fitness_i \qquad (2)$$

in order to calculate the fitness values, we use the following equation:

$$fitness_i = \begin{cases} \frac{1}{1+f} & if\ f \geq 0 \\ 1 + abs(f) & if\ f < 0 \end{cases} \qquad (3)$$

Greedy selection to update the solution

$$\left.\begin{matrix} a = a_{new} \\ f = f_{new} \end{matrix}\right\} \text{if } fitness_{\text{new}} > fitness_i(a_i) \text{And}$$
and remains the same if $fitness_{\text{new}} < fitness_i(a_i)$ (4)

Where f represents the food source. Onlookers will explore all the locations that seem promising and might have higher probability than other locations. The candidate food sources are then generated from the previously memorized ones as:
$$b_{i,j} = a_{i,j} + 2\,(r - 0.5)(a_{i,j} - a_{k,j}) \qquad (5)$$

A common problem of all the stochastic methods for optimization is that there is a poor balance between the exploitation and exploration results in weak optimization methods and thus suffers from either very slow convergence due to excessive exploration or premature convergence due to excessive exploitation [18].
ABC needs improvement as per some experiments especially in the way the new candidate food sources are generated, for it to be like other optimizers like PSO.
In order to improve the performance of ABC, the following alternative is used:

$$b_{i.j} = \begin{cases} a_{i,j} + \emptyset_{i,j}\,(a_{i,j} - a_{k,j}).\theta.\rho & if\ r_2 > z \\ a_{i,j} + \emptyset_{i,j}\,(a_{i,j} - a_{k,j}).2.\rho & if\ r_2 <= z \end{cases} (6)$$

Where $\emptyset_{i,j} = 2.(r_1 - 0.5)$ and k is a solution in neighbourhood of i, and is a random number in range [-1,1] and r1,r2 [0,1] are random numbers extracted from uniform distribution and

$$\rho = 0.5 - 0.25\,\frac{iter}{maxiter} \qquad (7)$$

θ is a number extracted from gaussian distribution & and represent current and total iterations respectively.
Practically, z is the parameter responsible for maintaining balance between gaussian & uniform distribution. Also, the search radius decreases with each iteration automatically.

Algorithm: Artificial Bee Colony (ABC)

```
1.  Initialize population of solutions $a_{i,j}$ using equation (1)
2.  Evaluate the population
3.  Run = 1
4.  Produce new food sources $b_{i,j}$ in the neighbourhood of $a_{i,j}$ for employed bees using equation (6)
5.  Apply greedy selection process between $a_{i,j}$ and $b_{i,j}$ by using equation (4)
6.  Calculate probability values $P_{bi}$ for $a_{i,j}$ with the help of their fitness values using the equations (1)
7.  Calculate the value of $fitness_i$ using eq. (3)
8.  Produce new positions bi for onlooker from the $a_i$ selected depending on $P_{bi}$ and evaluate by using equation (2)
9.  Apply greedy selection process for the onlookers between $a_{i,j}$ and $b_{i,j}$ by using equation (4)
10. Find the abandoned sources, if exists, and replace with new random solution $a_i$ for scout using equation
        $a_{i,j} = min_j + rand(0,1)^*(max_j - min_j)$
11. Memorise the position of best food source achieved so far.
```

```
12.      Run = Run + 1
13.      Repeat till Run = Maximum runs (iterations).
```

B. Ensemble Method: Boosting

Ensemble learning is a machine learning paradigm where multiple models (often called “weak learners”) are trained to solve the same problem and combined to get better results [17]. The main hypothesis is that when weak models are correctly combined, we can obtain more accurate and/or robust models.

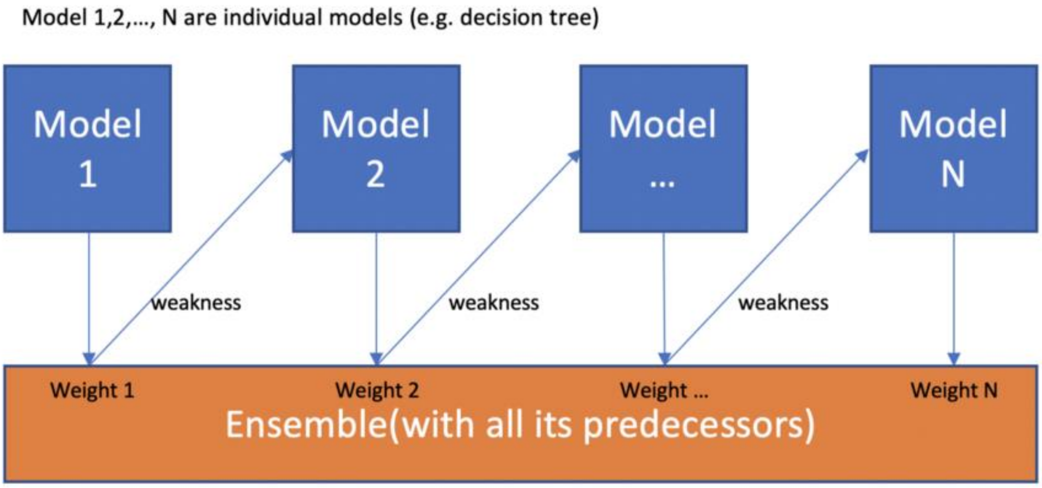
Model 1,2,..., N are individual models (e.g. decision tree)
Model 1
Model 2
Model ...
Model N
weakness
weakness
weakness
Weight 1
Weight 2
Weight ...
Weight N
Ensemble(with all its predecessors)

Figure 3 Boosting: Homogeneous ML Algorithms in a Sequential Way

Very roughly, bagging will mainly focus at getting an ensemble model with less variance than its components whereas boosting and stacking will mainly try to produce strong models less biased than their components (even if variance can also be reduced) [1].

IV. RESULT AND DISCUSSION

As a source dataset for experimental evaluation, we have used KDD dataset. KDD Dataset: This is the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between ``bad'' connections, called intrusions or attacks, and ``good'' normal connections. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment.

***Table 2** Dataset Description*

S. No.	Features	Samples
1.	42	65837
Url: https://www.unb.ca/cic/datasets/nsl.html		

A. Performance Metric

- Accuracy (A): Measures the proportion of correctly classified network traffic [13].

A= TP+TN/(TP+TN+FP+FN)

- Precision (P) : Measures the proportion of predicted attacks that were actual attacks, which indicates the relevance of a model’s predictions.

P=TP/(TP+FP)

- Recall/TPR/Sensitivity/Detection Rate (R): Recall, which corresponds to TPR, measures the proportion of actual attacks that were correctly predicted, reflecting a model’s ability to identify malicious activity.

R=TP/P = TP/(TP+FN)

- F1-Score: Overall, the most trustworthy metric is the F1-score, also referred to as F-measure.

○ It calculates the harmonic mean of Precision and Recall, considering both FP and FN.

○ A high F1-score indicates that malicious activity is being correctly identified and there are low false alarms.

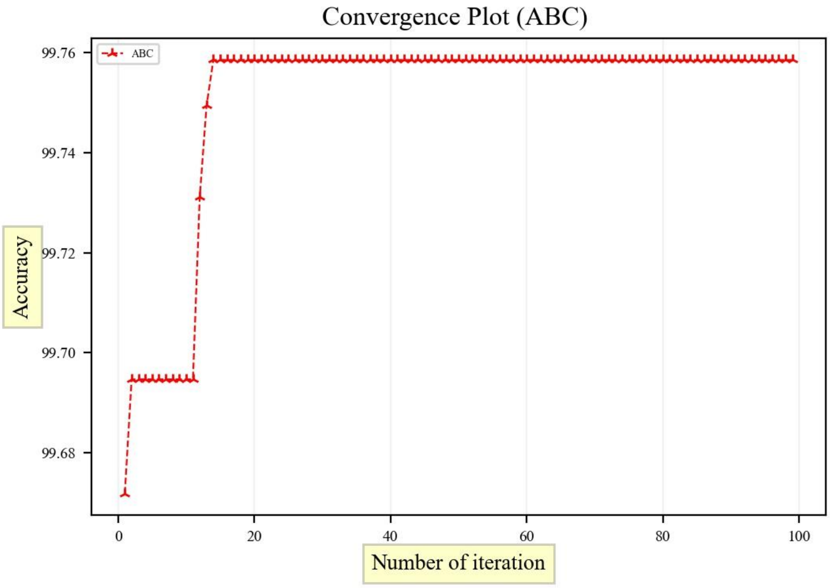

Figure 4 Convergence Plot (Accuracy ABC)

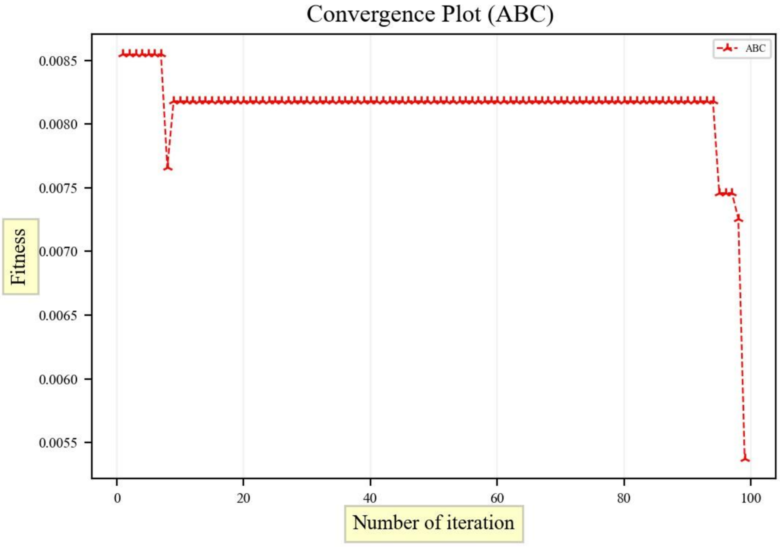

Figure 5 Convergence Plot (Fitness ABC)

Table 3 Evaluation

	Hold Out	**Ensemble (ADA Boost)**	**Ensemble (ADA Boost) FS**	**Ensemble (Random Forest)**	**Ensemble (Random Forest) FS**	**SVM**	**SVMFS**
Accuracy	0.400	0.965	0.971	0.985	**0.998**	0.891	0.901
	0.300	0.969	0.964	0.983	**0.998**	0.618	0.920
	0.200	0.971	0.972	0.984	**0.998**	0.833	0.920
	0.100	0.965	0.971	0.990	**0.997**	0.894	0.904
Precision	0.400	0.965	0.971	0.985	**0.998**	0.894	0.901
	0.300	0.969	0.964	0.983	**0.998**	0.738	0.920
	0.200	0.971	0.972	0.984	**0.998**	0.857	0.920
	0.100	0.965	0.971	0.990	**0.997**	0.895	0.906
Recall	0.400	0.965	0.971	0.985	**0.998**	0.891	0.901

	0.300	0.969	0.964	0.983	**0.998**	0.618	0.920
	0.200	0.971	0.972	0.984	**0.998**	0.833	0.920
	0.100	0.965	0.971	0.990	**0.997**	0.894	0.904
Fscore	0.400	0.965	0.971	0.985	**0.998**	0.891	0.901
	0.300	0.969	0.964	0.983	**0.998**	0.528	0.920
	0.200	0.971	0.972	0.984	**0.998**	0.834	0.919
	0.100	0.965	0.971	0.990	**0.997**	0.894	0.904
Note:	**FS-**	**Feature Selection**					
	SVM-	**Support Vector Machine**					

Table-1 shows the evaluation of different machine learning classifiers. Figure 4 and Figure 5 shows the convergence graph of ABC algorithm.

V. CONCLUSION

An intrusion detection framework expectation to separate between interruption action and ordinary activities. In doing as such, then again, an IDS can acquaint order blunders. A false positive is a delicate contribution for which the framework probably raises a notice. A false negative, interestingly, is a malicious information that the IDS prematurely delivers to report. The suitably grouped info information are ordinarily specified to as evident positives (suspicious assaults) and genuine negatives (ordinary activity). From result we can conclude that after feature selection, Ensemble (Random Forest) machine leaning classifier has achieved 99% accuracy which is significant then existing state of the art approaches. Proposed approach has been validated on different values of hold out.

REFERENCES

[1]. Mostert, Werner, Malan, Katherine M, & Engelbrecht, Andries P. (2021). A Feature Selection Algorithm Performance Metric for Comparative Analysis. Algorithms, 14(3), 100. https://doi.org/10.3390/a14030100.

[2]. Mafarja, Majdi, Qasem, Asma, Heidari, Ali Asghar, Aljarah, Ibrahim, Faris, Hossam, & Mirjalili, Seyedali. (2020). Efficient Hybrid Nature-Inspired Binary Optimizers for Feature Selection. Cognitive Computation, 12(1), 150–175. https://doi.org/10.1007/s12559-019-09668-6.

[3]. Emary, E, Zawbaa, Hossam M, & Hassanien, Aboul Ella. (2016). Binary grey wolf optimization approaches for feature selection. Neurocomputing (Amsterdam), 172, 371–381. https://doi.org/10.1016/j.neucom.2015.06.083

[4]. EESA, Adel Sabry, ORMAN, Zeynep, & BRIFCANI, Adnan Mohsin Abdulazeez. (2015). A new feature selection model based on ID3 and bees algorithm for intrusion detection system. Turkish Journal Of Electrical Engineering & Computer Sciences, 23, 615–622. https://doi.org/10.3906/elk-1302-53

[5]. Vieira, S. M., Mendonça, L. F., Farinha, G. J., & Sousa, J. M. (2013). Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients. Applied Soft Computing, 13(8), 3494-3504.

[6]. Ieracitano, C., Adeel, A., Morabito, F. C., & Hussain, A. (2020). A novel statistical analysis and autoencoder driven intelligent intrusion detection approach. Neurocomputing, 387, 51-62.

[7]. Gupta, D., Singhal, S., Malik, S., & Singh, A. (2016, May). Network intrusion detection system using various data mining techniques. In 2016 International Conference on Research Advances in Integrated Navigation Systems (RAINS) (pp. 1-6). IEEE.

[8]. Mukherjee, S., & Sharma, N. (2012). Intrusion detection using naive Bayes classifier with feature reduction. Procedia Technology, 4, 119-128.

[9]. Afza, A. A., & Uddin, M. S. (2014, March). Intrusion detection learning algorithm through network mining. In 16th Int'l Conf. Computer and Information Technology (pp. 490-495). IEEE.

[10]. Neethu, B. (2013). Adaptive intrusion detection using machine learning. International Journal of Computer Science and Network Security (IJCSNS), 13(3), 118.

[11]. Mukherjee, S., & Sharma, N. (2012). Intrusion detection using naive Bayes classifier with feature reduction. Procedia Technology, 4, 119-128.

[12]. Riad, A. M., & Fahmy, M. M. Intrusion Detection System Based on Data Mining techniques. Egyption Informatics Journal, Faculty of Computers and Information, 7

[13]. DikshantGupta et. al./Network Intrusion Detection System Using various data mining techniques/IEEE 2016.

[14]. Upendra et. al./An Empirical Comparison and Feature Reduction Performance Analysis of Intrusion Detection/IJCTCM 2012

[15]. A J M Abu Afza et. al./Intrusion Detection Learning Algorithm through Network Mining/ IEEE 2013

[16]. Neethu B/Adaptive Intrusion Detection Using Machine Learning/IJCSNS 2013

[17]. Upendra/An Efficient Feature Reduction Comparison of Machine Learning Algorithms for Intrusion Detection System/ijettcs 2013

[18]. Abhinav kumra et. al./Intrusion Detection System Based on Data Mining Techniques/OJCST 2017