IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 5, May 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|

N: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 5, May 2023||

| DOI:10.15680/IJIRSET.2023.1205252 |

Implementation of Intrusion Detection System Using RandomForest and Support Vector Machine Algorithms

Falguni Ghatkar[1], Sakshi kharche[2], Priyanka Doifode[3] ,Jagruti Khairnar[4] , Prof. Neelam Kumar[5]

Student, Department of Computer Engineering, Shree Ramchandra College of Engineering Pune, Maharashtra, India[1]

Professor, Department of Computer Engineering, Shree Ramchandra College of Engineering Pune, Maharashtra, India[5]

ABSTRACT: Intrusion detection techniques is based on intrusion model, which can be divided into two major types, misuse detection and anomaly detection. In this paper we are detecting the anomaly present in given dataset. In recent years Machine Learning (ML) algorithms has been gaining popularity in Intrusion Detection system (IDS) In Machine learnig has a algorithm is Support Vector Machines (SVM) And Random forest(RF) has become one of the popular ML algorithm used for intrusion detection due to their good generalization nature and the ability to overcome the curse of dimensionality.This paper represent the study that gaining high accuracy and random forest algorithm to SVM for intrusion detection. Apart Detection Calculate a term of accuracy, Precision, Recall, F-measure And improve the accuracy

KEYWORDS: SVM, Random Forest, Machine Learning, Supervised Learning.

I. INTRODUCTION

Cyber-attacks pose a significant and persistent threat to individuals, organizations, and nations in today's interconnected world. These malicious activities encompass a wide range of techniques and strategies aimed at exploiting vulnerabilities in computer systems, networks, and digital infrastructure. Machine learning algorithms have emerged as a powerful tool in enhancing IDS accuracy and effectiveness. By leveraging the vast amount of network data, machine learning algorithms can learn and identify patterns indicative of intrusions [2] Techniques such as artificial neural networks, support vector machines, and random forests are commonly employed to train IDS models using labeled datasets. These models can then analyze network traffic in real-time, accurately distinguishing between normal and malicious activities.Machine learning-based IDS offers the advantage of adaptability, as they can learn from new attack patterns and update their detection capabilities accordingly. Machine learning has emerged as a powerful tool for solving complex problems across various domains. Achieving high accuracy is a fundamental goal in machine learning, as it directly impacts the performance and reliability of predictive models. This research paper delves into the realm of improving accuracy in machine learning by exploring advanced techniques and methodology [1]

II.OBJECTIVES

In supervised learning, the goal of IDS (Intrusion Detection Systems) is to precisely identify and categorise instances of harmful or unauthorised activity within a computer network. IDS is a security tool that monitors and analyses network activity, system logs, and other pertinent data sources to assist safeguard computer systems and networks. IDS typically uses a machine learning algorithm in the context of supervised learning that has been trained on a labelled dataset.[6] The IDS algorithm's goal is to identify patterns and traits in known intrusions or attacks so that it can recognise and categorise similar occurrences in real time.

The main goal is to precisely identify intrusions or harmful activity.[3]

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 5, May 2023||

| DOI:10.15680/IJIRSET.2023.1205252 |

II. REVIEW OF LITERATURE

Falguni Ghatkar et al in," To Study Different Types of Supervised Learning Algorithm" paper states that A Concepts and principles underlying machine learning, including supervised, Un-supervised, and reinforcement learning. It then delved into an extensive discussion of popular machine learning.algorithms such as linear regression, decision trees, support vector machines, random forests Among Others. Each algorithm was critically evaluated based on its strengths, weaknesses, and suitability for different types of data and problem domains[1]

Guo Pu, Lijuan Wang et al. in A Hybrid Unsupervised Clustering-Based Anomaly Detection Method Paper presented an unsupervised anomaly detection algorithm SSC-OCSVM, which combined sub_space clustering and one class support vector machine to detect attacks without any prior knowledge. They also evaluated the proposed algorithm utilizing the well_known public NSL-KDD network attacks dataset. In addition, they compared its performance with three other clustering algorithms for unsupervised detection available in the literature [2]

Mausumi Das Nath, Tapalina Bhattasali et al. Anomaly Detection Using Machine Learning Approaches They have initially tried to carry out the classification process with a single classifier in their presented work. It is noted that it works best in individual scenarios. It may not yield noteworthy results in accuracy and computation when many features are present in the dataset. In such a scenario, the ensemble approach performs comparatively better, especially in handling vast network traffic where there is a challenging task of multiple types of anomalies and at the same time unknown. Here, the computation time is also less as Naïve Bayes works fast in real-time. Integrating multiple base classifiers, the ensemble approach provides promising detection results by compensating for their hazard. Accuracy has measured in percentage [3]

WHu, Y Liao et al. Robust Anomaly Detection Using Support Vector Machines In this paper,They have proposed a new approach, based on Robust Support Vector Machines, to anomaly detection in computer security. Experiments with the 1998 DARPA BSM data set show that RSVMs can provide good generalization ability and effectively detect intrusions in the presence of noise. The running time of RSVMs can also be significantly reduced as they generate fewer support vectors than the conventional SVMs [4]

Chee-Wooi Ten Anomaly Detection for Cybersecurity of the Substations. The contribution of this paper is a new substation anomaly detection algorithm that can be used to systematically extract malicious "footprints" of intrusion-based steps across substation networks. An impact factor is used to evaluate how substation outages impact the entire system.[5]

III. REQIUREMENT

- **Software Requirement**

➢ **Anaconda**
Anaconda is an open-source distribution of the Python and R programming languages for data science that aims to simplify package management and deployment.

➢ **Python**
Python is a computer programming language often used to build websites and software, automate tasks, and conduct data analysis. Python is a general-purpose language, meaning it can be used to create a variety of differentprograms and isn't specialized for any specific problems.

➢ **Tkinter**
Tkinter is the de facto way in Python to create Graphical User interfaces (GUIs) and is included in all standard Python Distributions. In fact, it's the only framework built into the Python standard library. Its Is used For The GUI.

- **Hardware Requirement**

➢ **Processor**

- Intel core i5 or above. 64bit, Quad-core.
- 2.5 GHz minimum per core

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 5, May 2023||

| DOI:10.15680/IJIRSET.2023.1205252 |

- **RAM**
- 8 GB or MORE

- **Hard Disk**
- Windows 10
- 10 GB of available space or more

IV. SYSTEM ARCHITECTURE

System architecture in machine refers to the overall structure of the components and processes involved in a machine learning system. It encompasses the arrangement of hardware, software, data, and algorithms to enable the training, evaluation, and deployment of machine learning models.

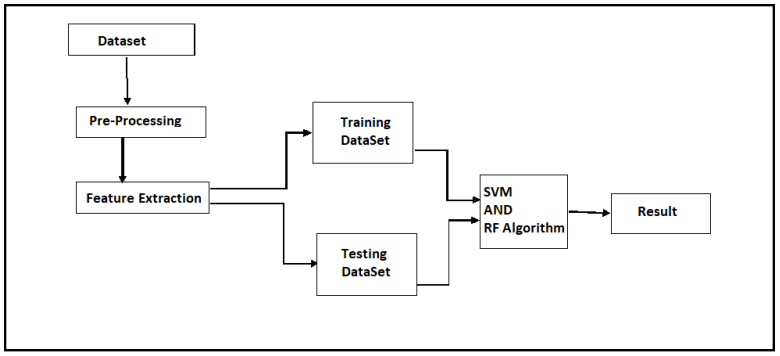


Fig 1 Intrusion Detection System Architecture

A. Dataset

Dataset refers to a collection of structured or unstructured data that is used for training, validating, and testing machine learning models. A dataset typically consists of a set of input data samples and corresponding output labels

B. Pre-Processing

Pre-processing refers to the modifications done to our data before we give it to the algorithm. A technique for converting unclean data into clean data sets is data pre-processing. In other words, if data are gathered from various sources, they are gathered in an unprocessed way that prevents analysis.

C. Feature Selection

Feature extraction is a step in the dimensionality reduction process that separates and compacts a starting set of raw data into more manageable categories.

D. Training Dataset

Data are split into two sets, one is the training set and the other is the testing set. Training set data are used to train the model

E. Testing Dataset

The input for the model is the testing set.

F. Proposed model using RF and SVM:

Once the RF and SVM model is trained and evaluated, it can be deployed in a real-time monitoring system. Network or system activities are continuously fed into the IDS, and the SVM and RF model classifies each instance as normal or intrusive based on the learned patterns.

G. Result

Attack names were mapped to one of the two classes, 0 for Normal, 1 for Attack. If Attack name is mapped with normal class then intrusion is not detected. And If Attack name is mapped with Attack class then intrusion are detected.

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 5, May 2023||

| DOI:10.15680/IJIRSET.2023.1205252 |

V. METHODOLOGY

Machine Learning: The field of study known as machine learning enables computers to learn without being explicitly programmed. One of the most intriguing technologies that has ever been developed is machine learning.[7]The ability to learn is what, as the name suggests, gives the computer a more human-like quality Machine learning isbeing actively used right now, possibly in a lot more ways.

i. **Supervised Learning:** In Supervised Learning, algorithms learn from labeled data. The algorithm decides the label to use after comprehending the input.[8]
Based on patterns and associations between the patterns and the new, unlabeled data, the algorithm decides which label should be applied to new data. Supervised learning (SL) is a machine learning paradigm used for situations when the available data consists of labelled examples, where each data point has a corresponding label. The two categories into which supervised learning can be divided are classification and regression. [8]. We are used the two major algoritam as follow

1. Support Vector Machine
2. Random forest

ii. **Suport Vector Machine(SVM) :** A splitting hyperplane defines SVM, a discriminative classifier. In order to linearly classify incursion, SVMs use a kernel function to translate the training data into a higher-dimensional space. SVMs are well renowned for their capacity to generalise and are most useful when there are a lot of attributes and few data points. By using a kernel, such as a linear, polynomial, Gaussian Radial Basis Function (RBF), or hyperbolic tangent, many types of separating hyperplanes can be produced. Many features in IDS datasets are redundant or have less of an impact on classifying data items into the appropriate groups. Therefore, when training an SVM, features selection should be taken into account. Multiple class classification is another application for SVM[1]

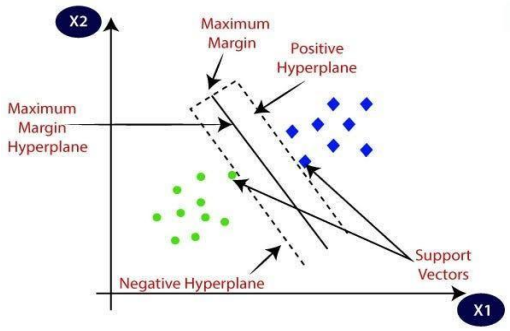


Fig.Support Vector Machine

Steps of Suport Vector Machine Step 1: Import the dataset in model
Step 2: Investigate the information to learn how they appear.
Step 3: Prepare pre-process the data.
Step 4: Separate the data into labels and characteristics.
Step 5: Set up training and test sets for the data.
Step 6: Develop the SVM algorithm.
Step 7: Make some estimates.
Step 8: Evaluate the algorithm's output.

iii. Random Forest(RF):

The random forests is an ensemble of unpruned classification or regression trees. Random forest generates many classification trees. Each tree is constructed by a different bootstrap sample from the original data using a tree classification algorithm. After the forest is formed, a new object that needs to be classified is put down each of the tree

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 5, May 2023||

| DOI:10.15680/IJIRSET.2023.1205252 |

in the forest for classification. Each tree gives a vote that indicates the tree's decision about the class of the object. The forest chooses the class with the most votes for the object [1]

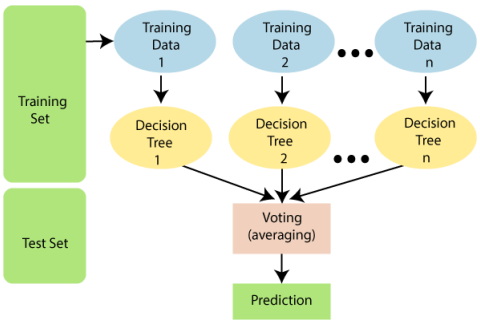


Fig. Random Forest

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.
The Working process can be explained in the below steps and diagram:
Step-1: Randomly select K data points from the Dataset
Step 2: Create the decision trees linked to the subsets of data that have been chosen.
Step 3: Choose N as the decision tree size that you want to build.
Step 4: Repeat steps 1 and 2
Step 5: By searching up each decision tree's predictions for the new data points, add new data points to the category that obtains the most votes

iv. Performance Metrics:

IDS has a variety of classification metrics, some of which go by several names. The confusion matrix for a two-class classifier is shown in Table 6 and can be used to gauge an IDS's effectiveness. The examples in a predicted class are represented by each column of the matrix, whereas the occurrences in an actual class are represented by each row [10] IDS are often assessed using the following common performance metrics:

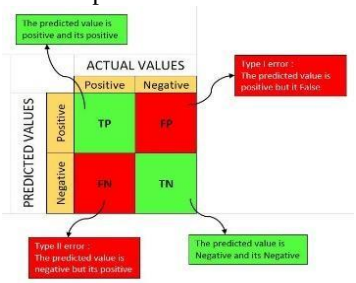


Fig. Performance Matrics

- **True Positive Rate (TPR):**
  The percentage of positive data points that are genuinely positive.

- **True negatives (TN)**:
  Data points marked as negative but truly negative are known as "true negatives" (TN).

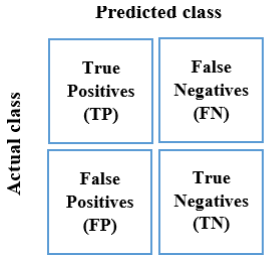
- **False Positive Rate (FPR):**
  The percentage of positive data items that are actually negative

- **False Negative Rate (FNR):**
  Data items that have been labelled as negative but are in fact positive

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 5, May 2023||

| DOI:10.15680/IJIRSET.2023.1205252 |



Measure	formula
Accuracy	(TP+TN)/(TP+FP+FN+TN)
Precision	TP/ (TP+FP)
Recall	TP/ (TP+FN)
F-Measure	2*Precision*Recall/ (Precision+Recall)

Fig. Evalutaion Parameter

Precision:

Precision is a metric that quantifies the number of correct positive predictions made. Precision, therefore, calculates the accuracy for the minority class [10] It is calculated as the ratio of correctly predicted positive examples divided by the total number of positive examples that were predicted [11]

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

Recall

Recall is a metric that quantifies the number of correct positive predictions made out of all positive predictions thatcould have been made.Unlike precision that only comments on the correct positive predictions out of all positive predictions, recall provides an indication of missed positive predictions [11]

$$Recall = \frac{TP}{P}$$

F1-Score

Precision and recall can be combined into one metric using F-metric, which covers both characteristics.
Precision and memory don't give the complete tale on their own. We can have great precision but poor recall, or vice versa, great precision but poor recall. A way to communicate both concerns with a single score is offered by the F-measure.. Once precision and recall have been calculated for a binary or multiclass classification problem, the two scores can be combined into the calculation of the F1-Score[11]

$$F\ Score = \frac{2TP}{2TP + FP + FN}$$

Accuracy

Accuracy is calculated as the number of correct predictions divided by the total number of predictions made by themodel [11]

$$accuracy = \frac{(TP + TN)}{TP + FP + TN + FN}$$

VI. EXPERIMENTAL RESULTS

In studies comparing SVM and Random Forest for IDS, both algorithms have shown promising results. SVM is known for its ability to handle complex data and find optimal hyperplanes, resulting in high accuracy rates. Random Forest, with its ensemble learning approach, can effectively handle large datasets and reduce overfitting, leading to robust performance. And detected the intrusion with the help of algorithm. Attack name are Mapped with the 0 then intrusion are not detected. And Attack name is mapped with 1 then intrusion are detected.

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 5, May 2023||

| DOI:10.15680/IJIRSET.2023.1205252 |

- *Accuracy*

SVM is slightly more accurate, but more expensive in terms of time. RF produces similar accuracy in a much faster manner if given modeling parameters. These classifiers can contribute to an IDS system as one source of analysis and increase its accuracy

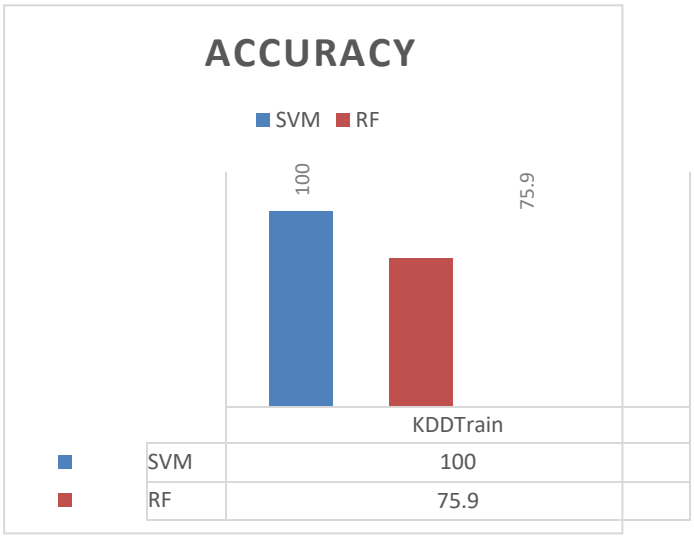


Fig. Accuracy Of Algorithm

The confusion matrix is a useful tool for evaluating the performance of a Support Vector Machine (SVM) in an Intrusion Detection System (IDS). It provides a comprehensive breakdown of the classification results by showing the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). From the confusion matrix, various evaluation metrics can be derived, such as accuracy, precision, recall, and F1 score. By analyzing the confusion matrix, researchers and practitioners can gain insights into the strengths and weaknesses of the SVM-based IDS model, enabling them to further optimize and fine-tune the system's performance.

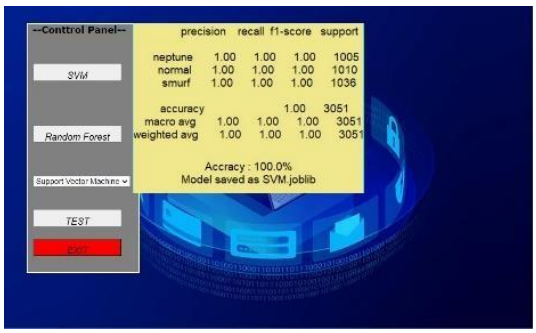


Fig. Accuracy of SVM

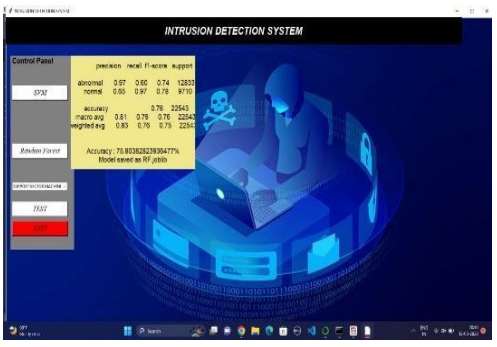


Fig. Accuracy Of Random Forest

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 5, May 2023||

| DOI:10.15680/IJIRSET.2023.1205252 |

The confusion matrix is a valuable tool for evaluating the performance of the Random Forest algorithm in an Intrusion Detection System (IDS). It provides a comprehensive breakdown of the predictions made by the model, comparing them to the ground truth labels. The confusion matrix consists of four key metrics: true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). TP represents the number of correctly predicted instances of malicious activity, while TN represents the correctly predicted instances of normal activity. FP indicates the instances where the model incorrectly classified normal instances as malicious, and FN represents the instances where the model incorrectly classified malicious instances as normal. By examining the values in the confusion matrix, we can derive various performance measures such as accuracy, precision, recall, and F1 score, which provide insights into the strengths and weaknesses of the Random Forest algorithm in identifying intrusions within a given IDS implementation.

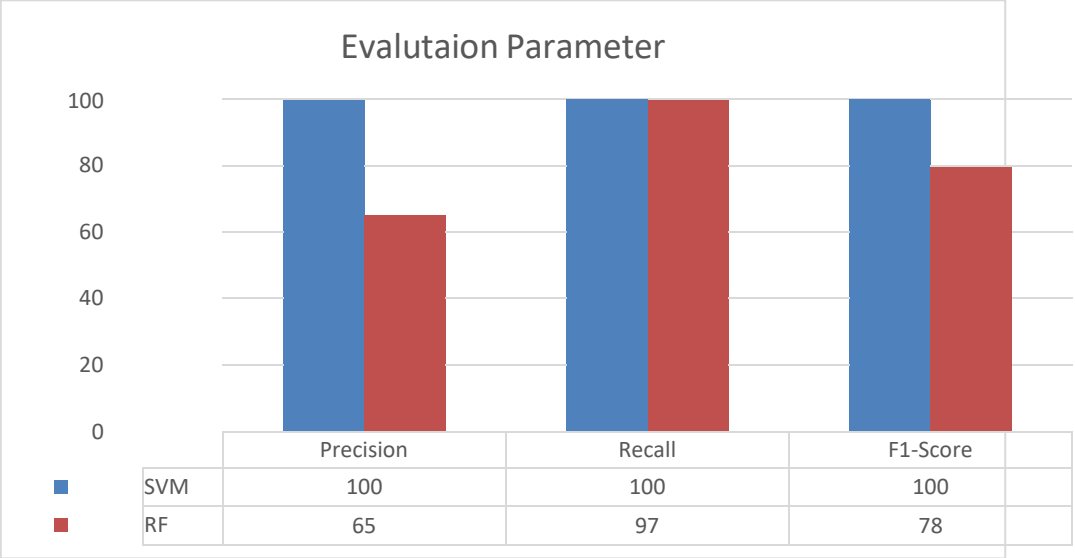


Fig. Graphical Representation of Evalutaion Parameter

VII. CONCLUSION

We are successful implementation of Intrusion Detection Systems (IDS) using Support Vector Machines (SVM) and Random Forest (RF) algorithms has shown promising results in terms of accuracy [7] VM is a powerful machine learning algorithm that works well with high-dimensional data and can effectively separate different classes. It has been widely used in IDS for its ability to handle complex datasets and find optimal hyperplanes to classify normal and malicious network traffic. Random Forest, on the other hand, is an ensemble learning method that combines multiple decision trees to make predictions. It is known for its robustness against overfitting and its ability to handle large datasets. Random Forest can be applied to IDS by creating an ensemble of decision trees trained on various features extracted from network traffic data. By using SVM and Random Forest algorithms together, IDS can benefit from the strengths of both methods. SVM can handle complex feature spaces and provide accurate classification, while Random Forest can improve the overall performance by reducing the risk of individual decision trees' biases.

REFERENCES

[1] Falguni Ghatkar, Sakshi kharche, Priyanka Doifode, Jagruti Khairnar, Prof. Neelam Kumar," To Study Different Types of Supervised Learning Algorithm" May 2023, International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), Volume 3, Issue 8, May 2023,PP-25-32

[2] A Hybrid Unsupervised Clustering-Based Anomaly Detection MethodGuo Pu, Lijuan Wang, Jun Shen, and Fang DongTSINGHUA SCIENCE AND TECHNOLOGY ISSNll1007-0214 02/11 pp146–153 DOI: 10.26599/TST.2019.9010051Volume 26, Number 2, April 2021

|e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

||Volume 12, Issue 5, May 2023||

| DOI:10.15680/IJIRSET.2023.1205252 |

[3] Anomaly Detection Using Machine Learning ApproachesMausumi Das Nath, Tapalina Bhattasali St. Xavier's College (Autonomous), Kolkata, India, m.dasnath@sxccal.edu,Azerbaijan Journal of High Performance Computing, Vol 3, Issue 2, 2020, pp. 196-206

[4] Cyber Security Challenges: Designing Efficient Intrusion Detection Systems and Antivirus Tools Srinivas Mukkamala, Andrew Sung and Ajith Abraham Department of Computer Science, New Mexico Tech, USASchool of Computer Science and Engineering, Chung-Ang University, Korea.

[5] C. R. SRINIVASAN, B. RAJESH, P. SAIKALYAN, K. PREMSAGAR, AND E. S. YADAV, "A REVIEW ON THE DIFFERENT TYPES OF INETWORK (INTRUSION NETWORK)," J. ADV. RES. DYN. CONTROL SYST., VOL. 11, NO. 1, PP. 154–158, 2019.

[6] "A Hybrid Unsupervised Clustering-Based Anomaly Detection Method", Guo Pu, Lijuan Wang, Jun Shen, and Fang Dong, TSINGHUA SCIENCE AND TECHNOLOGYISSNll1007-021402/11pp146–153 DOI:10.26599/TST.2019.9010051 Volume 26, Number 2, April 2021.

[7] "Anomaly Detection Using Machine Learning Approaches", Mausumi Das Nath, Tapalina Bhattasali. St. Xavier's College (Autonomous), Kolkata, India, m.dasnath@sxccal.edu, tapalina@sxccal.edu.Azerbaijan Journal of High Performance Computing, Vol 3, Issue 2, 2020.

[8] Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection in Wireless Network using KDDCUP'99 and NSLKDD datasets Shilpashree S*1, Dr. S C Lingareddy* Research Scholar, Alpha College of Engineering, Bengaluru, Affiliated to VTU, Belagavi.

[9] Survey of intrusion detection systems: techniques, datasets and challenges Ansam Khraisat* , Iqbal Gondal, Peter Vamplew and Joarder Kamruzzaman.

[10] Intrusion Detection using Machine Learning Techniques: An Experimental Comparison Kathryn-Ann Tait1 , Jan Sher Khan2 , Fehaid Alqahtani3 , Awais Aziz Shah4 , Fadia Ali Khan5 , Mujeeb Ur Rehman5 , Wadii Boulila6,7 and Jawad Ahmad1 1School of Computing, Edinburgh Napier University, United Kingdom.

[11] Analysis of Classification Techniques for Intrusion Detection Umair Ahmad, Hira Asim Department of Software Engineering University of Management and Technology Lahore, Pakistan

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details