Volume: 11 Issue: V Month of publication: May 2023

DOI: https://doi.org/10.22214/ijraset.2023.51499

A Survey on Intrusion Detection System using Machine Learning Techniques

Aishwarya B H[1], Bhavana S Akki[2], H M Harshitha[3], Navyashree R[4], Vedananda D.E[5]
[1, 2, 3, 4]*Research Scholar, Dept. of CS&E, JNNCE, Shimoga, India*
[5] *Assistant Prof., Dept. of CS&E, JNNCE, Shimoga, India*

Abstract: In every part of the world, there is a tremendous growth in digital literacy in the present era. People are trying to access internet-based applications with the use of digital machines. As a result, the internet as become a primary requirement for everyone, and most business transactions often take place conveniently across the network. On the other hand, intruders involved in making intrusions and doing activities such as capturing passwords, compromise on the route, collecting details of credit cards, etc. Many malicious activities are taking place over the network due to this intruding activity on the internet. The intrusion detection system (IDS) helps to find the attacks on the system and the intruders are detected. This paper has expected for an approach to develop IDS by using the principal component analysis (PCA) and the random forest classification algorithm. Where the PCA will help to shape the dataset by reducing the dimensionality of the dataset and the random forest will help in classification. Results obtained states that the proposed approach works more resourcefully in terms of accuracy as compared to other methods like SVM, Naïve Bayes, and Decision Tree.

Keywords: Intrusion detection system, Random Forest Approach, Principal Component Analysis.

I. INTRODUCTION

Nowadays, the involvement of the internet in normal life has been increased rapidly. The internet has made a crucial place in everyone's life. The use of the internet has become very crucial for everyone. So, with the increase in the use of the internet for personal activities, it is also necessary to keep secure the system from malicious activities. The evolution of malicious software (malware) poses a critical challenge to the design of intrusion detection systems (IDS). Malicious attacks have become more sophisticated and the foremost challenge is to identify unknown and obfuscated malware, as the malware authors use different evasion techniques for information concealing to prevent detection by an IDS.

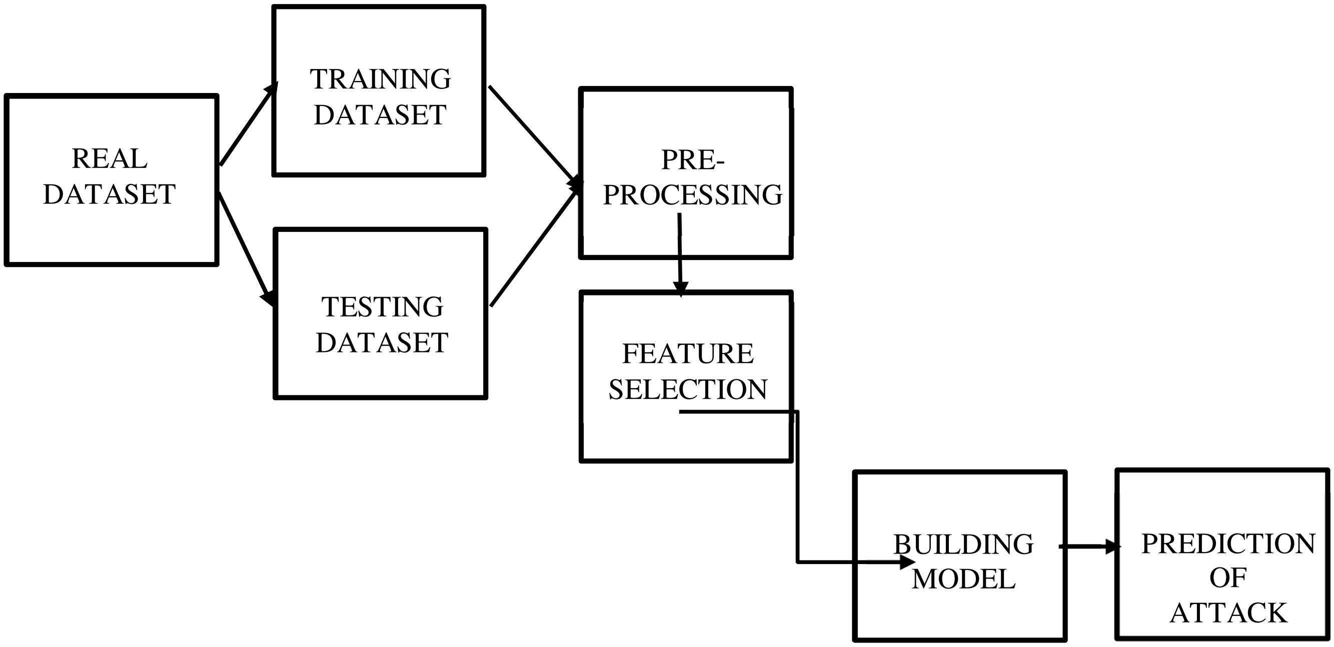

Fig. General methodology

As shown in above fig., these are the very common steps takes place in all intrusion detection system using machine learning models. It includes Data Collection, Data pre-processing, Feature selection, Building model and Prediction of attacks. Data Collection includes collecting data from various sources such as network traffic logs, system logs amd sensor data.

Data pre-processing includes cleaning, filtering and normalizing the collected data to remove any noise outliers or inconsistencies.

Feature Selection includes identifying the relevant features from the pre-processed data that are most important in intrusion detections.

Model Selection includes selecting approproate ML algorithms for the IDS such as decision tree, support vector machine, neural networks, random forest etc.

And the very last one is prediction of attacks, which detects and predicts the attack types.

II. RELATED WORK

The researchers, Sumaiya Thaseen Ikram et al. [1], have suggested an intrusion detection model that combines Principal Component Analysis (PCA) and Support Vector Machine (SVM) using the RBF kernel. The PCA technique reduces noisy attributes and selects the optimal attribute subset, which is then used to train the SVM classification models. The model's accuracy is improved by optimizing the SVM parameters C and ϒ for the RBF kernel using a proposed automatic parameter selection technique. The researchers evaluated the model's performance using two different datasets, NSL-KDD and gurekddcup. The results showed that the proposed model outperforms other classification techniques that use SVM as the classifier with PCA as the dimensionality reduction technique. Additionally, the proposed model requires fewer resources as the classifier input uses a reduced feature set, which reduces the training and testing overhead time.

Table-1: Comparison of Models

Algorithm Name	Datasets used	Precision	Recall	f1-score	Accuracy
PCA with SVM	NSL-KDD	0.9970	0.9970	0.9970	0.9970
	gurekddcup	0.997	0.999	0.998	0.996

Nada Aboueata et al. [2] conducted a study to evaluate the effectiveness of Support Vector Machine (SVM) and Artificial Neural Networks (ANN) machine learning techniques for detecting intrusions in the cloud environment. They compared the performance of different feature selection and parameter tuning schemes. Specifically, for SVM, they evaluated the effectiveness of features categorization, univariate, and PCA feature selection methods. They also adjusted two SVM parameters, namely the kernel function and penalty C parameter. The researchers found that the general-purpose feature category outperformed all other feature selection methods, achieving an accuracy of 90% with values of C as 20 and the Sigmoidal kernel function.

Table-2: Comparison of Models

Algorithm Name	Datasets used	Precision	Recall	f1-score	Accuracy
SVM	UNSW-NB-15	0.92	0.92	0.91	0.92
ANN		0.93	0.92	0.92	0.91

Erdo gan Dogdu et al. [3] proposed a novel method that combines big data technologies and deep learning techniques to enhance the performance of intrusion detection systems. They evaluated their approach using the UNSW-NB15 and CICIDS2017 datasets. The researchers used the homogeneity metric to rank and select the features, and they employed Deep Neural Networks (DNN), Random Forest (RF), and Gradient Boosted Tree (GBT) classifiers to classify the attacks in binary and multiclass modes. They conducted all experiments on Apache Spark with the Keras Deep Learning Library, while the RF and GBT classifiers were used from Apache Spark Machine Learning Library. The results showed high accuracy levels for DNN in binary and multiclass classification on the UNSW-NB15 dataset (99.19% and 97.04%, respectively) and low prediction times. The GBT classifier achieved the best accuracy (99.99%) for binary classification using the CICIDS2017 dataset and 99.57% accuracy for multiclass classification on the same dataset using DNN classifier.

Table-3: Comparison of Models

Algorithm Name	Datasets used	Accuracy
DNN	UNSW-NB-15	0.99
RF		0.98
GBT		0.97

Algorithm Name	Datasets used	Accuracy
DNN	CICIDS2017	0.97
RF		0.92
GBT		0.99

Shilpashree. S et al. [4] developed a decision tree-based model that enables system administrators to classify incoming traffic as malicious or non-malicious, regardless of the nature of the incoming data. They evaluated the model's effectiveness by comparing it with other strategies and tested it using both a 20% dataset and the full dataset. The experimental results showed that the decision tree-based model was effective in classifying incoming traffic and was powerful enough to handle the full dataset.

Table-4: Comparison of Models

Algorithm Name	Precision	Recall	f1-score	Accuracy
SVM	0.92	0.92	0.91	0.92
ANN	0.93	0.92	0.92	0.91

Ansam Khraisat et al. [5] conducted a comprehensive survey on intrusion detection system methodologies, types, and technologies, highlighting their advantages and limitations. They reviewed various machine learning techniques proposed for detecting zero-day attacks but noted their drawbacks, such as generating and updating information about new attacks, and high false alarms or poor accuracy. The authors also summarized recent research results and explored contemporary models to improve the performance of AIDS and overcome IDS issues. The study further examined popular public datasets used for IDS research, discussing their data collection techniques, evaluation results, and limitations. The authors highlighted the need for newer and more comprehensive datasets that contain a wide spectrum of malware activities, as normal activities are constantly changing and may become ineffective over time. Additionally, the study examined four common evasion techniques to determine their ability to evade recent IDSs. Developing IDSs that can accurately detect different types of attacks, including those that incorporate evasion techniques, remains a major challenge for this research area. There are several issues associated with these systems, such as high false positive rates, low detection rates, and difficulty in dealing with new and emerging threats. To improve the performance of anomaly-based IDS and overcome these issues, several contemporary models have been proposed. Some of these models are: Deep Learning-based IDS, Machine learning-based IDS, Ensemble-based IDS, Hybrid IDS, Cloud-based IDS, Big data analytics. These contemporary models offer promising solutions to improve the performance of anomaly-based IDS and overcome the issues associated with these systems.

Priyanka Sharma et al. [6] proposed a hybrid feature selection technique for the NSL-KDD dataset, followed by random forest classification on the reduced feature set. Their experiment results indicated that combining wrapper with filter methods and hybrid wrapper feature selection resulted in low accuracy, while filter-based feature selection using Gain ratio outperformed all other methods in terms of accuracy and time consumption in both testing modes. Hybrid feature selection techniques for the NSL-KDD dataset followed by the random forest approach can help improve the accuracy and efficiency of intrusion detection systems (IDS). The NSL-KDD dataset is a popular benchmark dataset for evaluating IDS, and the random forest approach is a widely used machine learning algorithm for classification tasks.

Some hybrid feature selection techniques that can be used for the NSL-KDD dataset are:

Correlation-based feature selection (CFS), Wrapper-based feature selection. After selecting the optimal subset of features using hybrid feature selection techniques, the random forest approach can be used for classification. The random forest approach is an ensemble learning technique that combines multiple decision trees to improve the accuracy of classification.

Some advantages of using the random forest approach for IDS are:

1) The random forest approach can handle high-dimensional datasets with many features.
2) The random forest approach can detect non-linear relationships between features and the target class.
3) The random forest approach can handle missing values and noisy data.

Hybrid feature selection techniques followed by the random forest approach can help improve the accuracy and efficiency of IDS using the NSL-KDD dataset. These techniques can help select only the most informative features and improve the classification accuracy of the IDS.

Jofrey L. Leevy and colleagues [7] have explored various intrusion detection datasets, including CICIDS2018, which is a large multi-class dataset with about 16 million instances, but is class-imbalanced. The authors searched for relevant studies based on this dataset and found that the reported performance scores were unusually high, possibly due to overfitting. They also observed that few studies addressed the class imbalance issue, which can distort experiment results, especially for big data. Another concern was the inadequate level of detail in data cleaning, which could impact the reproducibility of experiments. The authors identified several gaps in the current research, such as the absence of topics like big data processing frameworks, concept drift, and transfer learning. In a different context, the authors applied their approach to analyze Amazon.com reviews from various categories, which improved analytical precision and helped classify the reviews into various sentiments. Their approach produced accurate results on their datasets, and SVM was found to be the best algorithm for computing polarity on cryptic reviews.

Table 5: Summary of survey

TITLE	METHODOLOGY	ADVANTAGES	LIMITATIONS
Improving accuracy of Intrusion detection model using PCA and optimized SVM	Data pre- processing PCA technique SVM algorithm	Dimensionality reduction using PCA removes noisy attributes and retains the optimal attribute subset. Minimum training and testing overhead time.	This would not be applicable in a real time environment. SVM without normalization will increase calculation time.
Supervised Machine Learning Techniques for efficient network intrusion detection	Artificial Neural Networks Support Vector Machine	Clustering reduces the number of data points and as a result it reduces the training complexity significantly and showed better performance. The proposed classifier achieved an accuracy of 95.72% and false positive rate of 0.7%	Supervised machine learning models require large amounts of labeled data to train effectively. In the case of network intrusion detection, it can be difficult to obtain a large enough dataset that represents all possible types of attacks and normal network traffic.
Intrusion Detection System using Big data and Deep Learning Techniques	K-means Deep neural networks	The results show that the best accuracy results are obtained by DNN. This paper presented a method to improve the performance of intrusion detection systems by integrating big data technologies and deep learning techniques.	Big data is data that is difficult to store, manage or manipulate using traditional techniques. The characteristics of big data include volume, variety and velocity. The large volume of data is often associated with another challenge, which is variety, meaning different data source.
Random forest classification of NSL-KDD dataset using hybrid feature selection model.	Random forest classification algorithm.	Compared to a single decision tree algorithm random forest runs efficiently on large data sets with a better accuracy. Random forest is the best classification algorithm compared to others. This algorithm gives high accuracy.	Wrapper based hybrid feature selection and combining wrapper with filter method gives poor performance In terms of accuracy, recall, precision, roc and kappa and showing highest FPR and error.
A survey and analysis of intrusion detection models based on CSE-CIC-IDS2018 Big Data.	Deep learning Random forest classification algorithm	For the most part, they observed that the best performance scores for each study, where provided, were unusually high. The highest accuracies (100%) were obtained for the DDoS, and brute force attack types.	Topics such as big data preprocessing frameworks, concept drifts and transfer learning are missing from literature. The non-availability of an accuracy score for the collective attack types
Decision Tree: A Machine Learning for Intrusion detection	Decision Tree	The decision tree gives a higher accuracy of up to 82% for DoS attacks and 65% for probe. It requires less effort for data preparation during pre-processing.	It takes higher time to train the model. Improving Network performance is challenging.
Survey of intrusion detection systems: techniques, datasets and challenges	Artificial neural network Support vector machine	SIDS: Very effective in identifying intrusions with minimum false alarms (FA). AIDS: Could be used to detect new attacks. Could be used to create intrusion signature.	Cyber-attacks on ICSs is a great challenge for the IDS due to unique architectures of ICSs as the attackers are currently focusing on ICSs.

1) *PCA with SVM*: When used together, PCA can help to reduce the dimensionality of the data, making it easier and faster for SVM to classify the data. However, this approach may not always lead to the best performance, especially when there are nonlinear relationships between the features and the target variable.
2) *SVM with ANN*: When used together, SVM and ANN can complement each other, with SVM providing a robust classification algorithm and ANN providing the flexibility to learn complex relationships between the features and the target variable.

3) *Big Data and Deep learning:* When used together, deep learning can be used to process and analyze big data, providing insights and predictions that would be difficult to obtain using traditional techniques. However, deep learning requires large amounts of data and computational resources, which can make it difficult to implement in some settings.
4) *Decision Tree Models:* They are easy to interpret and can handle both categorical and continuous data. However, they can be sensitive to small changes in the data, which can lead to instability and overfitting.
5) *Hybrid Feature Selection technique with NSL-KDD*: They can be used to address the limitations of individual algorithms and provide more accurate and robust predictions. When applied to NSL-KDD and CICIDS2018 datasets, hybrid model selection techniques can help to improve the accuracy and efficiency of the intrusion detection system, leading to better protection against security threats and attacks.

III. CONCLUSION

It's difficult to say which technique is the "best" among these as each technique has its strengths and weaknesses, and the best technique depends on the specific requirements of the intrusion detection system and the characteristics of the data.

For example, PCA with SVM may be a good choice when there are many features in the dataset, and it's important to reduce the dimensionality of the data to improve the efficiency of the classification algorithm. SVM with ANN may be a good choice when the data has complex relationships that can't be captured by a linear SVM model alone. Big data with deep learning may be a good choice when the data is too large and complex to be processed using traditional techniques. Decision tree-based models may be a good choice when the data has a clear hierarchy of features that can be modeled using a tree-like structure. Hybrid model selection techniques may be a good choice when it's important to combine the strengths of multiple techniques to improve the accuracy and robustness of the system.

Therefore, the choice of the best technique depends on the specific needs of the system, the characteristics of the data, and the performance metrics used to evaluate the models. It's important to carefully evaluate the performance of different techniques on the specific dataset and select the one that provides the best results for the given requirements.

REFERENCES

[1] Sumaiya Thaseen Ikram1 and Aswani Kumar Cherukuri, " Improving Accuracy of Intrusion Detection Model Using PCA and Optimized SVM", CIT. Journal of Computing and Information Technology, Vol.24, No. 2, June 2016.

[2] Nada Aboueata, Sara Alrasbi, Andreas Kassler, Deval Bhamare, Aiman Erbad, "Supervised Machine Learning Techniques for Efficient Network Intrusion Detection", 978-1-7281-1856-7/19/$31.00 ©2019 IEEE.

[3] Osama Faker and Erdogan Dogdu. 2019. "Intrusion Detection Using Big Data and Deep Learning Techniques ". In 2019 ACM Southeast Conference (ACMSE 2019), April 18–20, 2019

[4] Shilpashree. S, S. C. Lingareddy, Nayana G Bhat, Sunil Kumar G, Decision Tree: " A Machine Learning for Intrusion Detection", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue- 6S4, April 2019 .

[5] Ansam Khraisat*, Iqbal Gondal, Peter Vamplew and Joarder Kamruzzaman, " Survey of intrusion detection systems: techniques, datasets and challenges", Springer open, Khraisat et al. Cybersecurity(2019)

[6] Priyanka Sharma, Rajni Ranjan Singh Makwana, "Random Forest Classification Of NslKdd Dataset Using Hybrid Feature Selection Model", Journal of Emerging Technologies and Innovative Research (JETIR), 2018 JETIR December 2018, Volume 5, Issue 12

[7] Jofrey L. Leevy* and Taghi M. Khoshgoftaar, "A survey and analysis of intrusion detection models based on CSE-CIC-IDS2018 Big Data", Springer open, (2020)

[8] J. Sanjay Rahul, J. Sai Keerthana, G. Tejaswini, R. N. S. Kalpana, "Intrusion Detection System Using Principal Component Analysis with Random Forest Approach", International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538 Volume 10 Issue V May 2022.

[9] G. Madhukar, G. Nantha Kumar, "An Intruder Detection System based on Feature Selection using Random Forest Algorithm", International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249-8958 (Online), Volume-9 Issue-2, December, 2019

[10] Mr Subhash Waskle, Mr Lokesh Parashar, Mr Upendra Singh, "Intrusion Detection System Using PCA with Random Forest Approach", International Conference on Electronics and Sustainable Communication Systems (ICESC 2020) IEEE Xplore Part Number: CFP20V66-ART; ISBN: 978-1-7281- 4108-4

[11] M. Al-Zewairi, S. Almajali, and A. Awajan. 2017. "Experimental Evaluation of a Multi-layer Feed-Forward Artificial Neural Network Classifier for Network Intrusion Detection System." New Trends in Computing Sciences (ICTCS), 2017 International Conference on. IEEE, Amman, Jordan 2017

[12] M. Belouch, S. El Hadaj, and Mo. Idhammad. 2017. " Two-Stage Classifier Approach Using Reptree Algorithm For Network Intrusion Detection." International Journal of Advanced Computer Science and Applications (ijacsa) 8.6 (2017).