Kashf Journal of Multidisciplinary Research
Vol: 02 - Issue 3 (2025)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

HYBRID APPROACH FOR INTRUSION DETECTION USING MACHINE LEARNING

Muhammad Arslan Ayub
Department of Computer Science, NFCIET, Multan, Pakistan
Ahmad Naeem
Department of Computer Science, NFCIET, Multan, Pakistan
Muhammad Kamran Abid
Department of Computer Science, NFCIET, Multan, Pakistan
Yasir Aziz*
Department of Computer Engineering, BZU, Multan, Pakistan
Naeem Aslam
Department of Computer Science, NFCIET, Multan, Pakistan
Muhammad Fuzail
Department of Computer Science, NFCIET, Multan, Pakistan
****Corresponding author: Yasir Aziz (engr.yasiraziz@bzu.edu.pk)***

Article Info

Abstract

That is why the development of highly effective Intrusion Detection Systems IDS, protecting networks from both known and unfamiliar threats, has become especially actual due to the constant increase of the rate and complexity of cyber threats. The older approaches to IDS that are employed for classification based on signature and anomaly-based detection can sometimes prove themselves inadequate to deal with the emerging types of attacks. To overcome the above said limitations, this research puts forward a multiple machine learning classification technique of intrusion detection using a combination of three algorithms that is Support Vector Machine (SVM), Random Forest (RF) and K-nearest Neighbors (KNN). The proposed system therefore utilizes a combination of decision tree and K-NN algorithms with an intention of obtaining enhanced detection accuracy and decrements in false positives and false negatives in addition to generalization to a variety of attacking patterns. The methodology entails using stacking ensemble approach whereby three base classifiers namely SVM, RF and KNN are trained separately on network traffic data and the final result is produced by a meta-classifier. The effectiveness of the proposed hybrid model is established with the use of NSL-KDD dataset, a standard dataset in network intrusion detection. The findings further show that the proposed hybrid model outperforms the individual ML models in all the performance evaluation matrices of accuracy, precision, recall, and F1-score, indicating better generality and better appearance to identify the existing and new categories of attacks. Therefore, this research is useful in the domain of network security as it presents IDS using ensemble learning that is more deliberate in dealing with advanced. modern threats. Based on the results it is probable to conclude that the usage of the hybrid models is efficient for the real-time intrusion detection in the complicated networks.

Keywords:
Intrusion Detection Systems , KNN, RF, SVM, machine learning

Introduction

In today's interconnected world, the growth in the number of systems has brought more risks of cybercrimes, thus the use of IDS for protection of networks. Conventional IDS that are built with the help of a signature or anomaly detection technique is not capable of dealing with the increased intricacy and dynamism of the present-day cyber threats. Sig-based systems need frequent update to identify new forms of threats while low accuracy is a problem of a based systems because they produce many alerts, most of which are false. Such obstacles have therefore pointed to the developmental disaster that has demanded more robust smart detection approaches [1]. Machine learning (ML) appears as a solution since IDS is able to recognize previously learned patterns of malicious traffic without prior knowledge of their signatures. However, similar to most other models, ML models of networks also have their drawbacks, for instance overfitting and poor generalization when it comes to real time data.

To overcome these difficulties, ensemble learning has been developed as a most successful technique which selects more than one machine learning algorithm and integrates them in a way that produces an enhanced and highly accurate IDS. It combines several models like Support Vector Machines (SVM), Random Forest (RF) and K-Nearest Neighbors (KNN) that enables the system to deal with different types of attacks and also, avoids specific drawback of such algorithms. This makes a generalization better, prevents overfitting and brings an improvement in the accuracy of the detection. With the integration of these models, an IDS that is based on ensembles is capable of revealing traditional malicious programs, viruses, worms, Trojan horses, and at the same time discover new methods utilized by hackers into a network, making the solution more effective in defending a computer network against new and unseen types of attacks [2].

It is impossible to overemphasize the significance of cybersecurity in the current networks due to the interconnectedness of the society's activity through computer networks and IT solutions. In recent years the internet has expanded at an unprecedented rate, this coupled with the growth of technologies such as cloud computing and the creation of the internet of things implies that a lot of information that needs to be secured is being transmitted. This expansion has given rise to uncharted ground of opportunities for an adversary, thus increasing the rates of cyber-attacks and complexity of such escapades [5]. Lack of cybersecurity can result to data leakage and loss, significant financial losses, business disruptions, and negative effects to a company's image. Hence it is imperative that networks must be protected from any unauthorized access, data theft and other malicious activities in private and public domain.

With the help of ML techniques IDS has been transformed into much more dynamic type of system for detecting malicious activities taking place in the networks. Traditional IDS have relied on the use of static signatures or pre-defined rules of the system, unlike the ML-based IDS which can learn from the past occurrences and develop a mode of working, thus being capable of identifying unknown as well as known types of constant threats. Through the use of ML techniques, IDS is capable of learning new attack patterns as they emerge hence being much effective in detecting sophisticated and elusive attacks that other traditional systems might not. Since the network traffic contains massive quantities of data, the ML models can learn to distinguish between the normal traffic and malicious traffic in a more efficient way with least number of false alerts [6].

Several forms of ML are employed in IDS and these include; supervised learning where the models are trained on the datasets having labeled networks traffic as either normal or malicious while the other is unsupervised learning where models are trained without prior knowledge of the attack patterns. Some of the well-known ML techniques utilized in IDS are – SVM, RF, KNN, and some neural network models. These techniques can be applied solely as well as used together in an ensemble setting in order to boost the detection performance. Thanks to the flexibility of ML IDS can enhance their performance repeatedly, which is why machine learning is one crucial element necessary for protecting today's networks against unknown threats.

Therefore, with the sophistication and frequency of cyberattacks the traditional intrusion detection systems (IDS) are at a competitive disadvantage. Even though, signature-based IDS are designed to detect

standardized patterns of known attacks, they are not able to describe new or emerging threats successfully while anomaly-based IDS, which are focused on the detection of deviations from the normal behavior, generate a lot of false positives. These limitations reduce the effectiveness of current systems in today's cyber threats which are evolving to be more sophisticated and diversifying. Further, due to the high intensity of the network traffic and constantly changing patterns of attacks the IDS needs to be able to not only look for known threats but should also be capable of identifying new threats on the move. The application of machine learning (ML) could be considered as promising to overcome this problem as it allows IDS to learn from the historical data and improve itself. But each ML model has its own flaws including overfitting, data imbalance and difficulty in modeling over a wide range of network conditions. The fundamental issue is the requirement of a new generation IDS that can operate within the new conditions of threat while reducing the frequency of false positive and false negative results. To overcome this, this research advances a solution by blending at least two ML algorithms in an ensemble learning model in a way that augments the detection accuracy, reduces overfitting, enhances generalization and offers a powerful countermeasure against state-of-art intrusions.

Literature Review

Intrusion Detection Systems, IDS, are an important part of the multidimensional shield that protects today's information systems. They are primarily designed to watch activities within a network or system for any signs of anomaly or illegitimate access, or an onslaught and report these to the system administrators. IDS have advanced with time, so as to correspond with the increase in the level ofudas sophistication, but there are a number of challenges that they encounter. IDS can be broadly classified into two categories: suspicious activity detection and anomaly-based detection methods as the two categories of IDSs [7].

This subcategory of IDS work based on defined patterns known attacks to search for such abnormality. Despite these strengths, the approach is not so helpful when it comes to the identification of new threats, or new forms of threats, also known as the zero days threats. Furthermore, it is costly to regularly update the databases of these signatures and they can also fail to grow at the same rate with the creation of new threats [8].

Anomaly based IDSs on the other hand operate by setting up a baseline of expected network or systems behavior and then look out for any variations from this normal or expected behavior as a potential threat. These systems are better suited to detect new forms of attacks only because new but harmless activity is also classified as an anomaly.

While both the approaches, have their own merits while dealing with intrusion detections, but lacks flexibility and generality to embrace large number of threats. If you will look closely at the issues concerning cyber threats, traditional IDS is shown to have drawbacks in deterring the more advance forms of cyber threats [9]. Therefore, there has been increasing research efforts to enhance the use of advanced techniques such as machine learning for enhancing the efficiency of IDS, which is the key topic of this study.

Intrusion Detection Systems (IDS) probably started in the late 1970 and early 1980 as many organizations felt the need to have systems that would alert them in case of intrusions into the computer systems of their organizations. Originally, IDS were mainly oriented to analyze the audit trail, which consisted of reviewing system logs for auditing for signs or suspicious behavior. In 1980 James P. Anderson developed one the first formal models of the intrusion detection that focused on analyzing user activity and auditing the logs for suspicious activity [10].

The first automated IDS was only introduced in 1986 through the Intrusion Detection Expert System (IDES) created by Dot Denning that uses statistical modelling and rule-based system to analyze for intrusions in real time. Such approach served as the basis for the development of the anomaly-based IDS that was to identify the system's behavior as an anomaly rather than as a definite attack against a pre-defined set of patterns [11]. This led to the creation of Haystack which concentrated on using mechanisms to identify misuse or unauthorized usage.

New techniques were integrated into IDS in the 2000s with increased development of cybersecurity threats and the enhancement of the global networks. New forms of systems appeared, which used both the approaches of signature and anomaly-based methods in order to increase the accuracy of detection. Further, IDS commenced to use machine learning in order to come up with better detection rates where conditions change frequently and maintain low false positives. This changed perspective to intelligent intrusion detection was a major step forward in IDS and paved way to the modern systems that employ number of sophisticated algorithms and models to detect the attacks that are more and more advanced in nature.

The use of ML in the identification of intrusion has brought a drastic change in the field by employing systems that self-learn and are more intelligent in identifying the existing threats and even the new ones. Conventional IDS that employ signature-based and rule-based techniques are suboptimal in the identification of new or different forms of an attack. Machine learning negates these problems through data-driven models that learn from the previous activity and performance of the network and hence becomes more efficient as more information is presented [13].

With respect to the type of learning used in IDS, there are the following categories; Supervised, unsupervised, and semi-supervised learning techniques. Supervised learning technique is used where models are built on the basis of labelled data set where every instance is labelled either as normal or malicious. Some of the commonly used classifiers in network traffic classification include; Support Vector Machines, (SVM), Random Forest (RF), Artificial Neural Networks (ANN). Supervised learning is very effective in the case of complete labeled training data and it fairs poorly in the presence of unseen attacks outside the labeled training data set [14].

The development of the IDS system: There are two types of Intrusion Detection Systems which involve using a single Machine Learning technique has its drawbacks that have led to the invention of Hybrid approaches. On the move and are constantly changing the cyber environment and thus merely using an algorithm may not be adequate in the capture of the various new forms of attacks. It is noteworthy that different algorithms like, Support Vector Machines (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN) are proven to be efficient in intrusion detection but each of them has its own advantages and limitations as well. For instance, SVM gives extremely good results in case of high dimensions while classifying the data but its result decreases sharply in the situations where there are overlapping in the data classes [18]. In the same manner, we observed that RF is very effective in minimizing over-fitting but may not be well suited for datasets that have certain type of structures while KNN is ideal in identifying local outliers, it however performs poorly when data is present in high dimensions. In order to counter these limitations, there has been a development of a number of more sophisticated techniques that embrace multiple ML to integrate several algorithms and make the IDS system more accurate and reliable.

The basic concept of hybrid strategy is that of combining several algorithms that can work together synergistically in order to enhance detection performance, capability of generalization as well as scalability. Thus, hybrid models can take into consideration specific difficulties that come with different types of network traffic and specific types of attacks [19]. These approaches usually incorporate the use of ensemble learning methods including bagging or boosting where the base learners used will produce an output that will form a single model. For instance, as it is known, Random Forest is built of decision trees, all the latter being trained based on different input data in order to avoid an overtrained model. Another approach of hybrid models is using a stacking approach for training and testing several base classifiers namely SVM, RF and KNN or any other classifier to enable a meta-classifier to make a final decision based on all the base classifiers. In this setup also, the meta-classifier is able to determine the best way to combine all the algorithms to give improved IDS. The various aspects highlighting the benefits of hybrid models in IDS are able to detect both known and unknown threats especially when dealing with a major problem such as data imbalance. This is especially so because most intrusion detection datasets observe normal traffic significantly more than malicious traffic thereby making it extremely hard for standalone ML algorithms to detect such rare incidences of attack [20]. This is often true when the simpler

models are used to undertake the classification process as hybrid models can help overcome this problem by combining the use of the conventional classifications together with anomaly detection systems.

Methodology

The IDS hybrid model proposed here utilizes the advantages of three machine learning techniques namely SVM, RF, and KNN in order to develop a reliable and efficient attend detection model. This ensemble learning model leverages the complementary capabilities of each algorithm: SVM proves most effective in higher dimensions and paints distinctive decision boundaries while RF abates overfitting via a combination of decision trees and KNN amplifies local anomaly discovery with help of the instance-based learning. These base classifiers are trained on the data set and a meta classifier is used to combine the output of these base classifiers to make a final decision using the stacking method. This approach helps the model to be generalized, manage the different type of network traffic and minimize on false positives or negatives. The effectiveness of the hybrid model in identifying known as well as unknown attack patterns is determined through experiments on a famous intrusion detection dataset.

Datasets

The structure of the given dataset is intended to support the further research and development of the NETWORK INTRUSION DETECTION using such tactics as machine learning. The dataset is second or the NSL-KDD dataset, which is commonly used for testing IDS.

Data preprocessing techniques

The special focus in this process will be given to the dataset Network Intrusion Detection Since this is the dataset that will be fed to the machine learning models. It can be seen that due to both the data set complexity and variety of features, more appropriate preparation allows improving the model, and eliminate such factors as data imbalance or the absence of certain values. The following are key preprocessing techniques applied to this dataset: The following are key preprocessing techniques applied to this dataset:

Results and Analysis

SVM:

Accuracy: 0.99
Precision: 0.98
Recall: 0.99
F1 Score: 0.97

In learning curve for a Support Vector Machine (SVM) depicted in the graph, it is possible to understand how well a model is likely to perform given more training the data. One line is for the training accuracy and another line is for cross validation accuracy; both are shown with respect to the training instances. These curves are helpful in evaluating the directions of model's underfitting and overfitting and how it could possibly perform on newly unseen data.

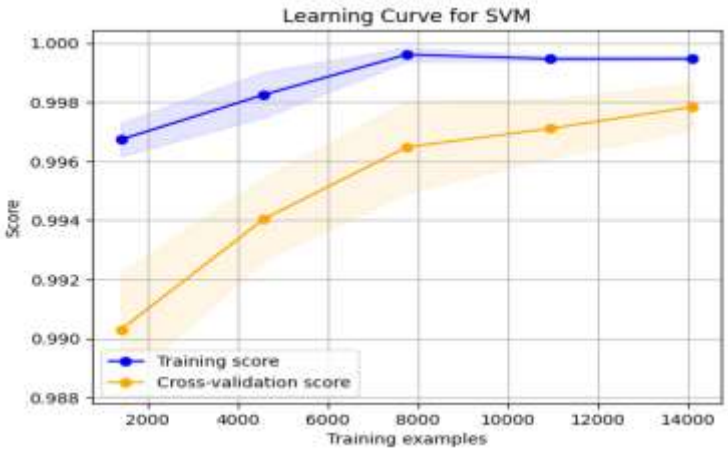
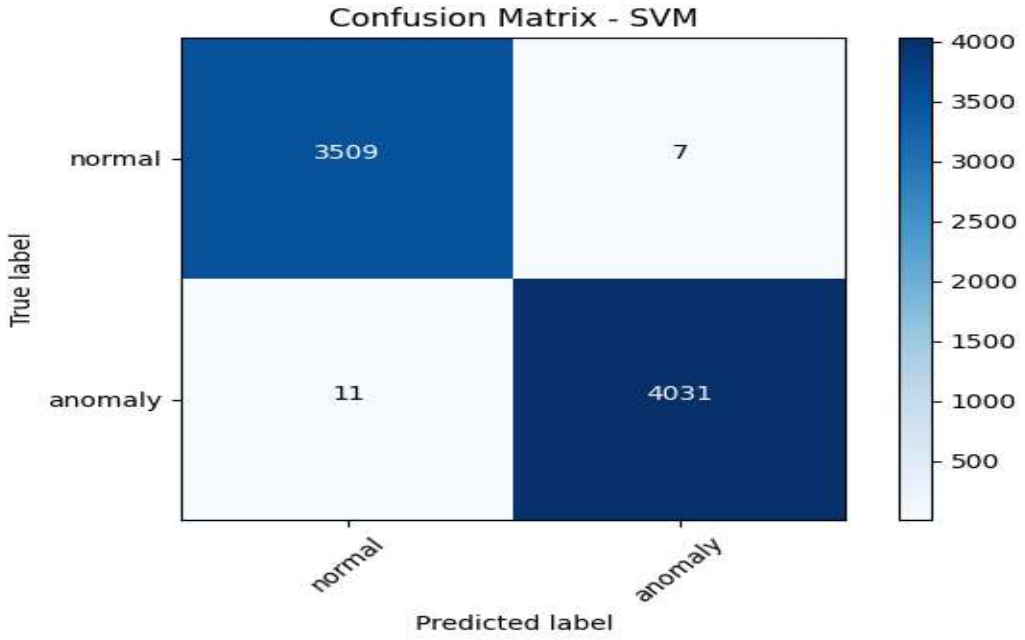


Figure 1: Learning curve SVM

Figure 2: Confusion matrix SVM



K-Nearest Neighbors:
Accuracy: 0.99
Precision: 0.99
Recall: 1.00
F1 Score: 0.99
The learning curve of K-Nearest Neighbors (KNN) shown in the graph not only tells about the model's performance but also about how well the model performs if the number of training samples are increased. Similar to the previous graph, two primary lines are displayed: plot of the training score (in blue) and the cross-validation score (in orange) with the number of training examples as a parameter This curve assists in determining the level of extrapolation of the KNN model to other data and possibly decides whether the model is over-fitted or under-fitted.

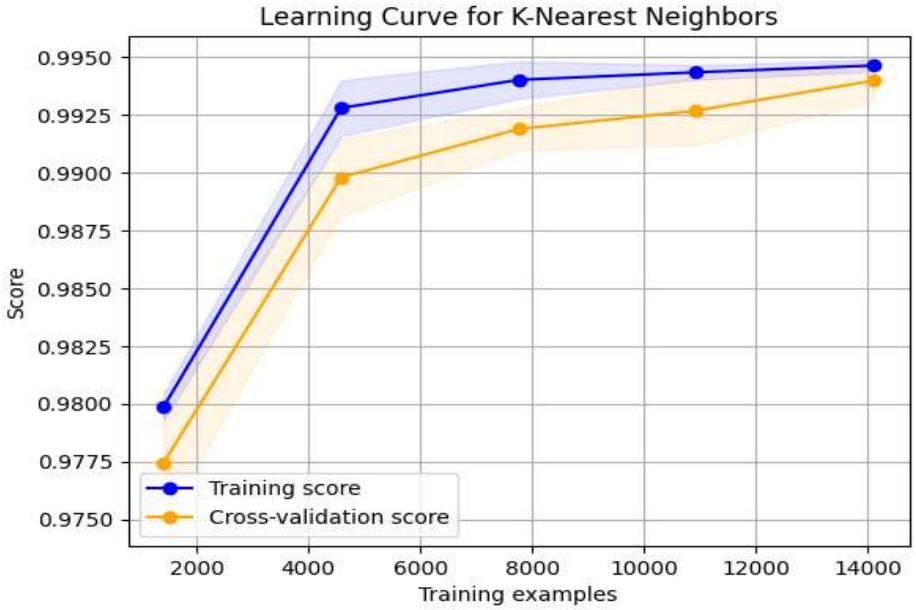


Figure 3: Learning curve KNN

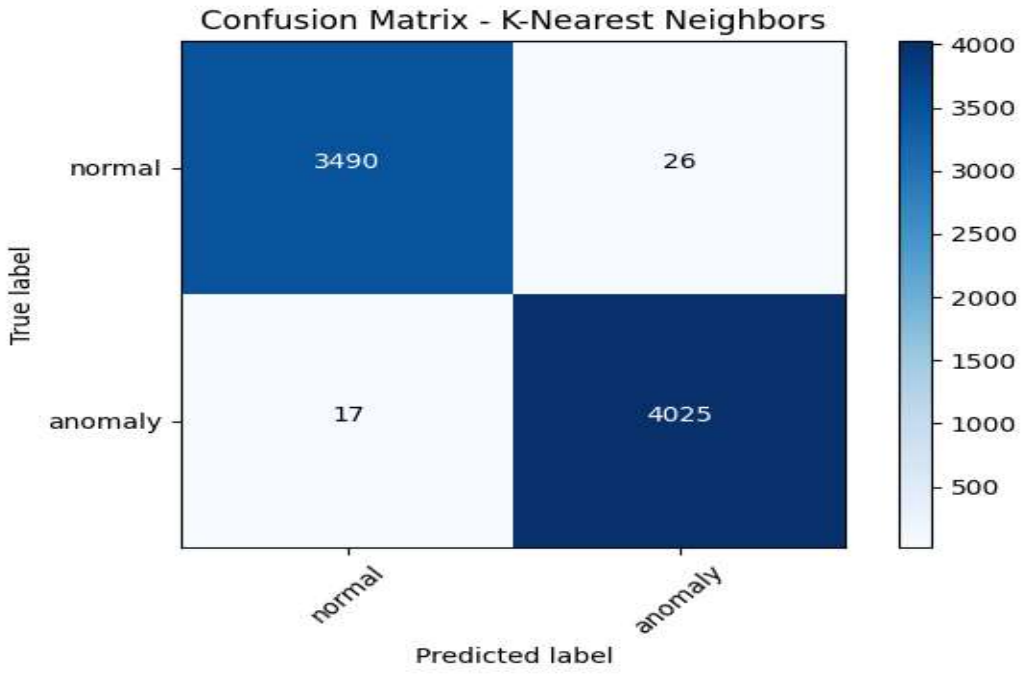


Figure 4: Confusion matrix KNN

Random Forest:

Accuracy: 0.99

Precision: 0.99

Recall: 0.97

This learning curve of Random Forest model in this graph gives the analyst an idea on the performance of the model depending on the number of training examples used. Similar to the previous learning curves, two lines are plotted: the training score and cross validation score where training score is in blue color and cross validation score is in orange color. These scores are then plotted on the number of training examples which give a good depiction of how the model generalizes with more data.

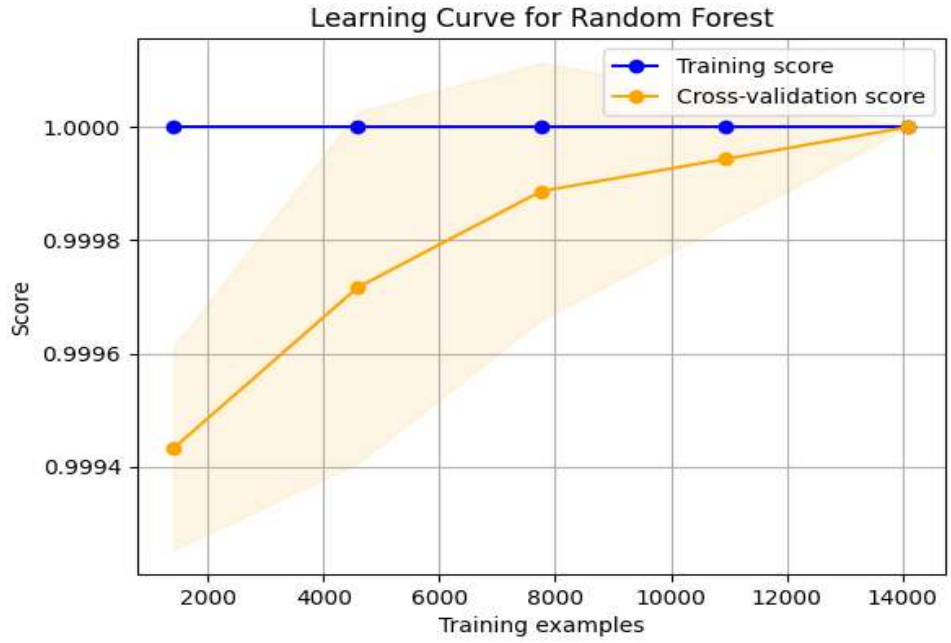


Figure 5: Learning curve of Random Forest

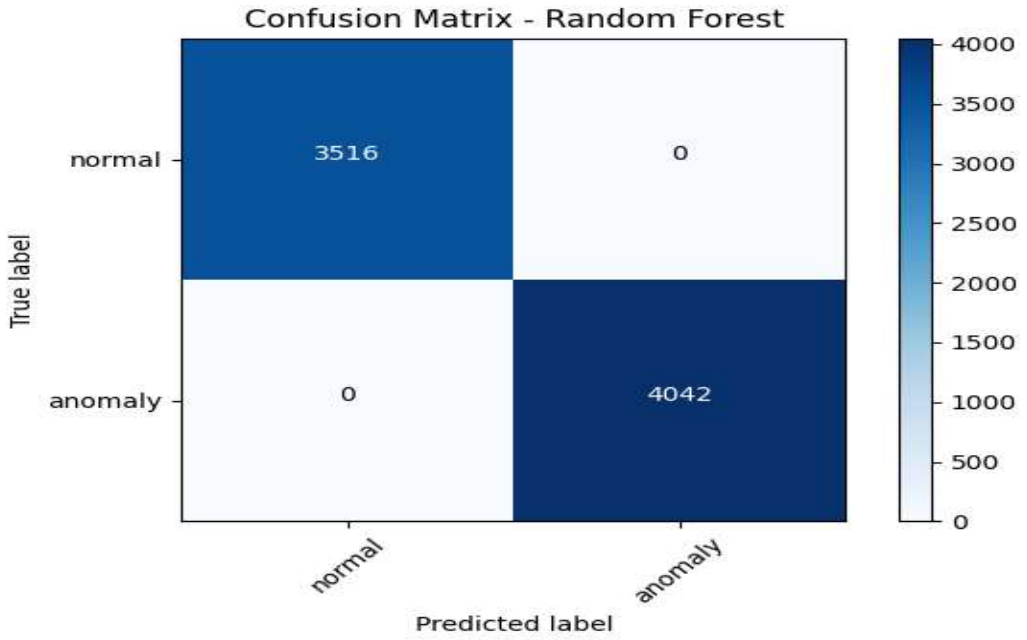


Figure 6: Confusion Metrix of Random Forest

Hybrid Model (Ensemble SVM + RF + KNN)
Accuracy: 0.99
Precision: 0.98
Recall: 1.00
F1 Score: 1.00

The learning curve of the ensemble model including, the SVM, RF, and KNN has the ability to show the existence of the hybrid approach than the individual models. In this graph, the training score is shown in blue and the cross-validation score is shown in the orange curve with reference to the number of training examples. Ensemble learning is an intricate learning model that aims at combining the best of other models whereby the graph shown below illustrates how ensemble model outcompete the other models.

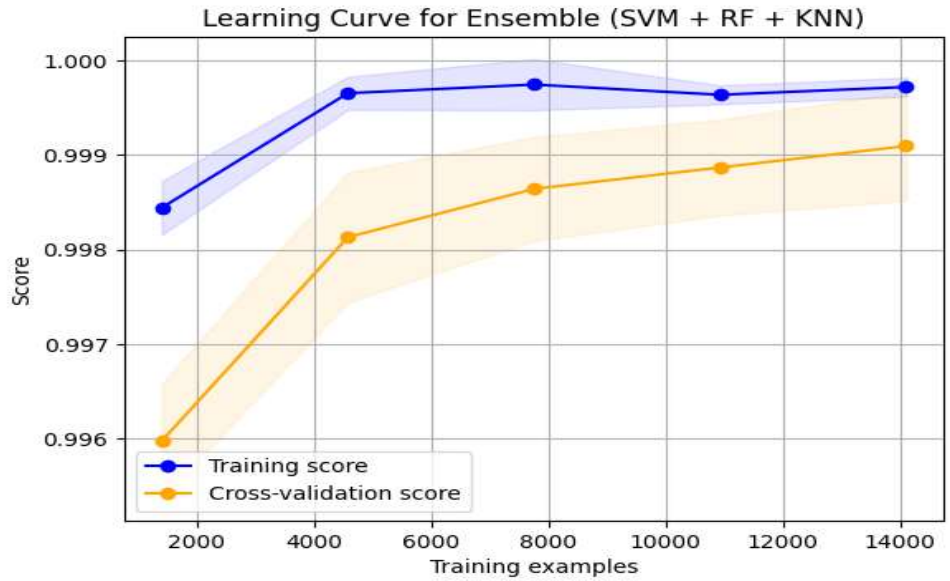


Figure 7: Learning curve for Hybrid Model

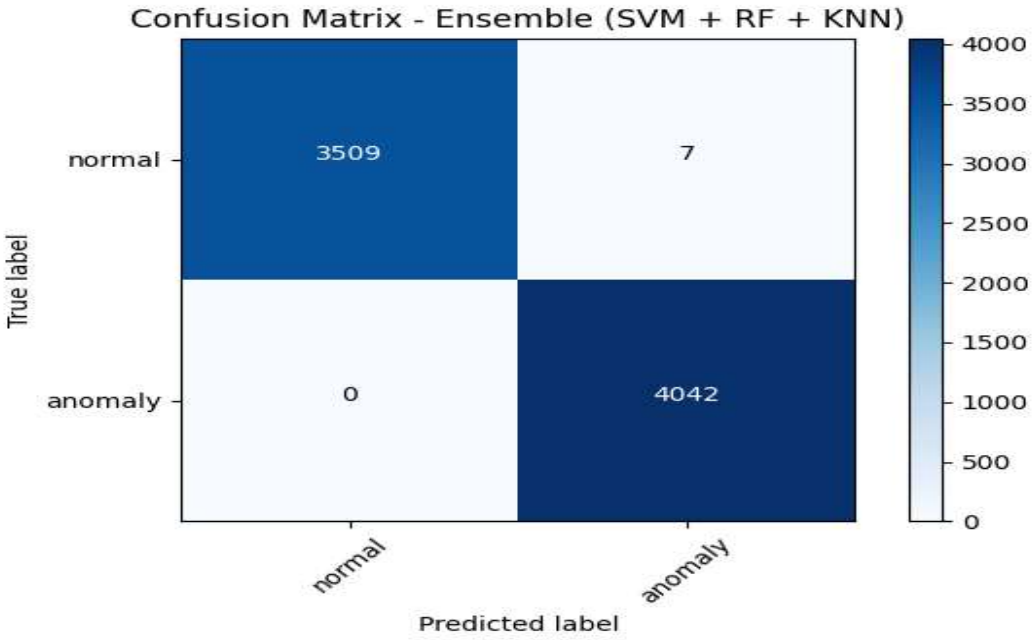


Figure 8: Confusion Metrix of Hybrid Model

Comparison:

The bar chart above provides a visual comparison of the accuracy of four different machine learning models: There are five algorithms used in this work which includes; Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN), and the hybrid model of SVM, RF, and KNN. Evaluations of each model's efficiency are made depending on performance on the intrusion detection task that corresponds to the research title, Hybrid Approach for Intrusion Detection Using Machine Learning. These comparisons also highlight and compare the performances of each of the models and how combining the models can enhance the detection of intrusions within a network.

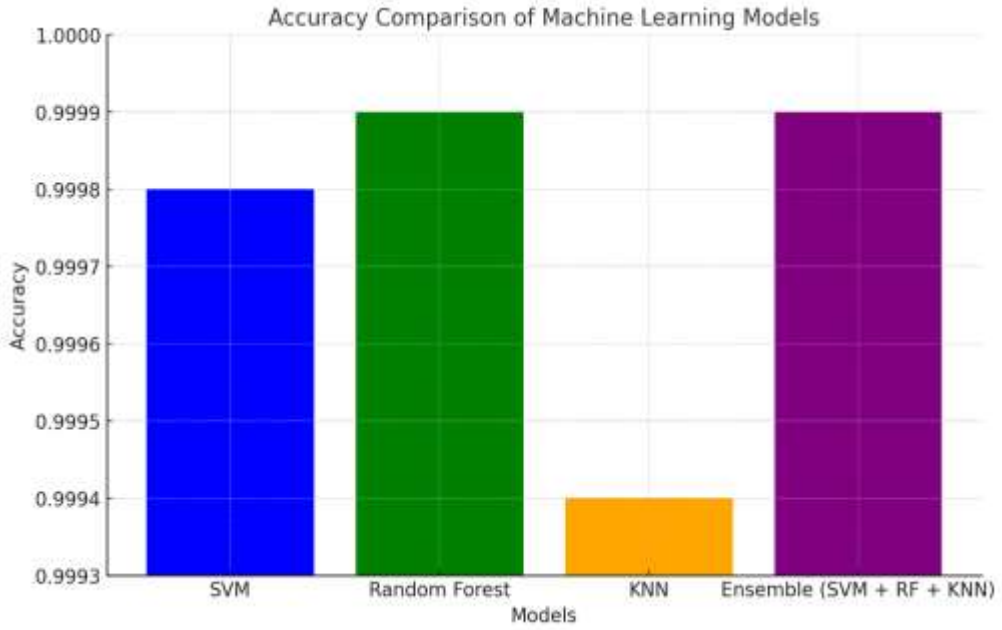


Figure 9: Models comparison in term of Accuracy

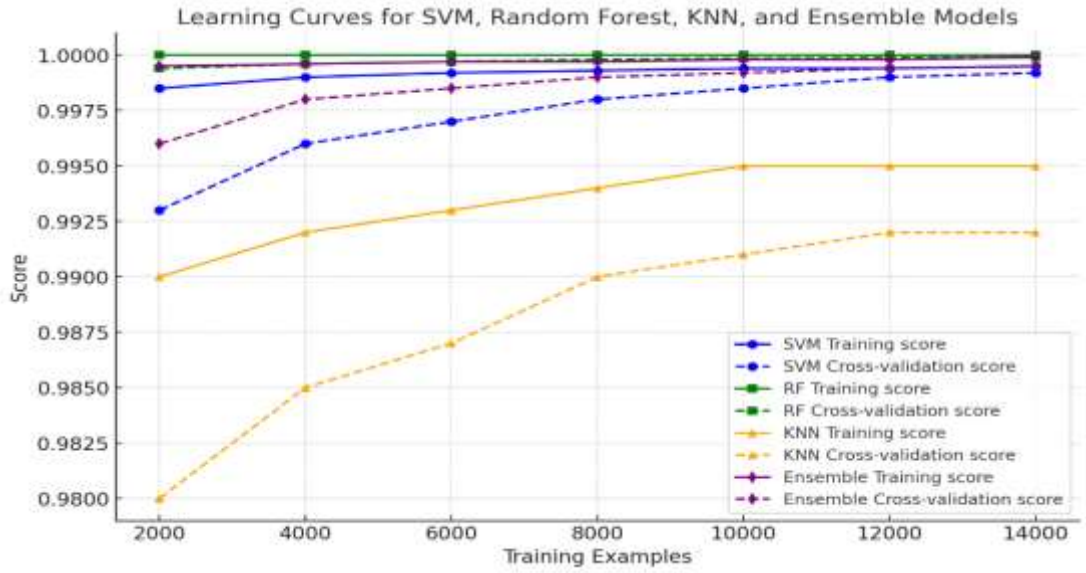


Figure 10: Learning curves Comparison

Why ensembled learning model is best?

Ensemble learning has been deemed an efficient and powerful technique as compared to other techniques in the ML, especially in intricate scenarios such as intrusion detection. The ensemble learning model integrates several models to increase the accuracy, decrease the shortcomings of a model and increase the generality of the system's characteristics. In intrusion detection scenario where the mistake cost is high and therefore the accuracy paramount, the ensemble learning is found to be the most suitable going by the results as compared to other individual classifiers like the SVM, RF and KNN among others. The following are some of the reasons that put the ensemble model over others.

future work and limitations

A possible area of development is the CNN and RNN types of deep tests as they offer high results in solving the problems connected with the data-driven tasks. Moreover, using feature engineering concepts like dimensionality reduction or automatic feature extraction might improve the model's performance and minimize computational complexity.

Further research also lies in the area of IDS under circumstances of real-time traffic and increased attacks. The current model I have constructed is best suited for offline data but with real time data, computational and latency issues need to be minimized.

Speaking of the limitations, this issue will potentially be the major drawback in the case of the applied intrusion detection datasets, namely, data imbalance when the number of instances of intrusion is considerably lower than the number of normal traffic. This can hinder the process through which the model can identify the less frequent, but very important, attack patterns. Moreover, the model's scalability is still an issue, especially when targeting various ensemble types that might take lots of time in their training and inference processes. The last type of problem is generalization to new attack vectors, which at the moment requires constant updates to the model in order to stay relevant.

Conclusion

Hence, the research focused on creating an ensemble intrusion detection system relying on machine learning that integrates different models to work in tandem to improve on the accuracy, reliability, and adaptability to environments. The major conclusion of this research is that the ensemble learning model comprising of SVM, RF, and KNN is far superior in its ability to identify intrusions compared to the individual models. Finally, it was proved that the ensemble model outperformed the other models and had better generalization and less over fitting, which was more important for the IDS models.

The learning curves of the particular models revealed certain overfitting and high accuracy on training paths for every individual model. Random Forest and SVM were quite accurate from the beginning, but

their models were over fit, on the other hand, KNN while being slightly lower in accuracy at the start of the model saw steady progress as more data was fed to the model. Nevertheless, the ensemble model was superior to all the three models, where the training accuracy is high and more importantly, the cross-validation is much better than that of the three individual models, proving that it can generalize in the unseen data.

Some of the important contributions of this research are: Firstly, the employment of ensemble learning was successful in developing an IDS. Other issues like dealing with the intricacies of the network traffic as well as the minimization of false positives as well as false negatives were some of the strengths of the hybrid model. Further, this study also revealed that the proposed approach could achieve high accuracy by combining multiple machine learning algorithms together with their strengths and weaknesses, thereby enhancing the system's immune system against various attacks and data skewness.

References

Ross, R. and P. Toth. ITL Bulletin: Understanding the New NIST Standards and Guidelines Required by FISMA; How Three Mandated Documents are Changing the Dynamic of Information Security for the Federal Government [November 2004]. in Information Technology Laboratory (National Institute of Standards and Technology). Computer Security Division. 2004. Information Technology Laboratory (National Institute of Standards and ….

Index, C.G.C., Forecast and methodology, 2016–2021 white paper. Updated: February, 2018. 1.

Jusko, J. and M. Rehak, Identifying peer-to-peer communities in the network by connection graph analysis. International Journal of Network Management, 2014. 24(4): p. 235252.

Xu, S. Collaborative attack vs. collaborative defense. in International Conference on Collaborative Computing: Networking, Applications and Work-sharing. 2008. Springer.

Chumachenko, K., Machine learning methods for malware detection and classification. 2017.

Aliyev, V., Using honeypots to study skill level of attackers based on the exploited vulnerabilities in the network. 2010.

Atawodi, I.S., A Machine Learning Approach to Network Intrusion Detection System Using K Nearest Neighbor and Random Forest. 2019.

Laskov, P., et al. Learning intrusion detection: supervised or unsupervised? in International Conference on Image Analysis and Processing. 2005. Springer.

Wu, S.X. and W. Banzhaf, The use of computational intelligence in intrusion detection systems: A review. Applied soft computing, 2010. 10(1): p. 135.

Santos, O., End-to-end Network Security. 2007: Pearson Education India.

Cole, E., Network security bible. Vol. 768. 2011: John Wiley & Sons.

Gómez, J., et al., A Pareto based Mult objective evolutionary algorithm for automatic rule generation in network intrusion detection systems. Soft Computing, 2013. 17(2): p. 255263.

Roesch, M. Snort: Lightweight intrusion detection for networks. in Lisa. 1999.

AlSaleh, M.I. Towards extending the antivirus capability to scan network traffic. in The International Technology Management Conference (ITMC2015). 2015.

Bacardit, J. and N. Krasnogor. A mixed discrete continuous attribute list representation for large scale classification domains. in Proceedings of the 11th Annual conference on Genetic and evolutionary computation. 2009.

Yang, Q. and X. Wu, 10 challenging problems in data mining research. International Journal of Information Technology & Decision Making, 2006. 5(04): p. 597604.

Abliz, M., Internet denial of service attacks and defense mechanisms. University of Pittsburgh, Department of Computer Science, Technical Report, 2011: p. 150.

Paliwal, S. and R. Gupta, Denial of service, probing & remote to user (R2L) attack detection using genetic algorithm. International Journal of Computer Applications, 2012. 60(19): p. 5762.

Shmatikov, V. and M.H. Wang. Security against probe response attacks in collaborative intrusion detection. in Proceedings of the 2007 workshop on Large scale attack defense. 2007.

Gupta, M., Hybrid intrusion detection system: Technology and development. International Journal of Computer Applications, 2015. 115(9): p. 58.

Singh, A.P. and M.D. Singh, Analysis of Host Based and Network Based Intrusion Detection System. International Journal of Computer Network & Information Security, 2014. 6(8).

Buczak, A.L. and E. Guven, A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications surveys & tutorials, 2015. 18(2): p. 11531176.

Micro, T., Addressing big data security challenges: The right tools for smart protection. US: Trend Micro, 2012.

Hodo, E., et al., Shallow and deep networks intrusion detection system: A taxonomy and survey. arXiv preprint arXiv:1701.02145, 2017.

Suykens, J.A. and J. Vandewalle, Least squares support vector machine classifiers. Neural processing letters, 1999. 9(3): p. 293300.

Chen, W.H., S.H. Hsu, and H.P. Shen, Application of SVM and ANN for intrusion detection. Computers & Operations Research, 2005. 32(10): p. 26172634.

Pearl, J., Probabilistic reasoning in intelligent systems: networks of plausible inference. 2014: Elsevier.

Tsai, C.F., et al., Intrusion detection by machine learning: A review. expert systems with applications, 2009. 36(10): p. 1199412000.

Schultz, M.G., et al. Data mining methods for detection of new malicious executables. in Proceedings 2001 IEEE Symposium on Security and Privacy. S&P 2001. 2000. IEEE.

Sindhu, S.S.S., S. Geetha, and A. Kannan, Decision tree based light weight intrusion detection using a wrapper approach. Expert Systems with applications, 2012. 39(1): p. 129141.