RESEARCH ARTICLE | FEBRUARY 16 2024

Improving cybersecurity with random forest algorithm-based big data intrusion detection system: A performance analysis

Alaa Abd Ali Hadi ✉; Amjad Mahmood Hadi

AIP Conf. Proc. 3051, 040012 (2024)
https://doi.org/10.1063/5.0191707

Improving Cybersecurity with Random Forest Algorithm-based Big Data Intrusion Detection System: A Performance Analysis

Alaa Abd Ali Hadi[1,a)] and Amjad Mahmood hadi[2,b)]

[1]*Al-Samawa Technical Institute, Al-Furat Al-Awsat Technical University, Samawa, Iraq*
[2]*Al Muthanna University, Samawa, Iraq*

[a)]Corresponding Author: alaa@atu.edu.iq
[b)]amjad@mu.edu.iq

Abstract. Even security specialists find it challenging to monitor the complex interconnections of computers and network devices brought about by the expansion of the internet over the past ten years. Network security has grown to be a major problem as personal computers have become faster and high-speed internet has become more widely accessible. It is extremely difficult to create intrusion detection systems that can manage massive amounts of data, especially in terms of system construction time. This work suggests a preprocessing feature selection strategy that creates subsets of pertinent characteristics to ease model construction in order to overcome this difficulty. The suggested model uses the information gain method to improve accuracy while classifying network data using the Random Forest algorithm. Using the NSL-KDD reference dataset, the suggested model's efficacy is assessed. Several measures are used to determine how well it performs. According on empirical findings, the recommended model outperforms existing algorithms in terms of performance measures. It offers a contrast. Overall, the proposed methodology has a great deal of promise to enhance large data intrusion detection systems' functionality.
Keywords: Cybersecurity, Random Forest Algorithm, Big Data, Intrusion Detection System, Performance Analysis, Threat Detection, Anomaly Detection

INTRODUCTION

In today's world of interconnected networks, network security has become a vital component of global infrastructure, especially as personal, ecommerce, banking, and business data are increasingly transmitted across computer networks. Intrusion Detection Systems (IDSs) is a significant area of network security, aiming to identify and prevent network attacks, which pose significant challenges for security communities [1,2]. Given the discovery of new vulnerabilities every year, security tools face difficulties automating the detection of new attacks. The intrusion detection system is an essential tool in safeguarding computer networks from attacks, as organizations worldwide use firewalls to protect their confidential data from public networks [3]. Nevertheless, firewalls cannot guarantee 100% protection of resources. Therefore, the intrusion detection system plays a critical role in network security by identifying adversarial activities on a network. The IDS tool functions under the presumption that the signature of an attack is distinct from that of routine activities. Signature-driven and anomaly-based detection approaches are the two main ways to find assaults. Signature-based Attacks can be discovered by comparing network traffic to a database of recognized attack signatures [3,4]. Anomaly-based detection, on the other hand, keeps track of network traffic for out-of-the-ordinary patterns of behavior in relation to a predetermined typical baseline and sends out alerts as necessary.

Advancing Snooping Detection Using Big Data Analytics

The framework for overcoming the difficulties presented by large data in the context of intrusion detection is described in this section along with an overview of those difficulties. Due to the quick advancement of technology, massive amounts electronic data are daily production. 'Big Data', which is defined as data that is too big and complicated to be processed using conventional techniques, has emerged as a major problem in network security as a result of the need to guard against security breaches in a number of crucial industries, including finance, business, and healthcare. The identification and prevention of intrusion patterns and unfamiliar packets within a network are made possible by Intrusion identification Systems (IDS), a popular security solution. Traditional techniques, however, face considerable difficulties because of how complicated network data is and the high dimensionality of

3rd International Conference on Engineering and Science
AIP Conf. Proc. 3051, 040012-1–040012-10; https://doi.org/10.1063/5.0191707
Published by AIP Publishing. 978-0-7354-4808-7/$30.00

network patterns [5,6]. It has become common practice to use machine learning techniques to create IDS that can recognize zero-day attacks. In this research, we concentrate on IDS across huge data using machine learning algorithms. To increase the precision and speed of intrusion detection, we suggest a new model that applies the Information Gain and Random Forest techniques. Additionally, we use feature selection techniques to lower the dataset's dimensionality, which enhances the effectiveness of the classifier. The essay is set up like follows: the paper is introduced in Section 1, big data is discussed in Section 2 in the context of IDS, related works are presented in Section 3, the proposed methodology is described in Section 4, the experimental analysis is outlined in Section 5, a performance in contrast to shown under Section 6, and this document is concluded in Section 7 [8].

RELATED WORKS

The application of machine learning techniques for intrusion detection in the context of big data has been the subject of several studies. One study classified huge data for intrusion detection using a hybrid feature selection method and Support Vector Machine (SVM). Another study combined the K-Nearest Neighbors (KNN) algorithm with feature reduction to detect intrusions. Another study looked into convolutional neural networks (CNNs) and long-short-term memory (LSTM) models as deep learning models for intrusion detection. The Random Forest Algorithm, which can handle high dimensional data with numerous attributes, has also been investigated as a potential method for intrusion detection in big data. The effectiveness of various systems must still be compared, and they must be improved for intrusion detection in massive data.

TABLE 1. Summary of related work

Research	Strengths	Weaknesses
Zhang et al,, 2016	a decision tree technique for feature selection and classification was proposed.	Limited evaluation on small dataset
Han et al,, 2018	Proposed an ensemble-based IDS with random forest algorithm, achieving high accuracy	Evaluation limited to one dataset, NSL-KDD
Li et al,, 2019	Proposed an IDS based on long short-term memory network and feature selection, achieving high accuracy and efficiency	Evaluation limited to one dataset, UNSW-NB15
Alazab et al,, 2019	Provided a comprehensive review of deep learning-based IDS, highlighting their strengths and weaknesses	Did not propose a new IDS system
Ghaleb et al,,2019	Proposed a deep learning approach for IDS in cloud computing, achieving high accuracy	Evaluation limited to one dataset, KDDCup99
Javed et al,,2020	Proposed a hybrid machine learning approach with feature selection and clustering for IDS	Limited evaluation on small dataset
Li et al,,2021	Proposed a hybrid machine learning approach with feature engineering and random forest algorithm, achieving high accuracy and efficiency	Evaluation limited to one dataset, UNSW-NB15
Akhtar et al,,2021	Provided a comprehensive study on deep learning-based IDS, comparing different techniques and highlighting their strengths and weaknesses	Did not propose a new IDS system
Chandrasekar & Anand, 2021	Proposed an ensemble classification approach for network intrusion detection, achieving high accuracy and efficiency	Evaluation limited to one dataset, UNSW-NB15
Ramanathan & Patel , 2022	Provided a review of machine learning techniques for IDS, highlighting their strengths and weaknesses	Did not propose a new IDS system
Zhang et al,,2022	Proposed a novel hybrid IDS framework based on autoencoder and random forest classifier, achieving high accuracy	Evaluation limited to one dataset, NSL-KDD

METHODOLOGY

The approach for enhancing cybersecurity using the Random Forest algorithm Preprocessing the data in order to eliminate any discrepancies or pointless features is the first stage in a huge data intrusion detection system. The dimensionality for the dataset is also reduced by feature selection, which improves how well the intrusion detection system is working. The most pertinent features are determined in this study using an information acquisition method [5].When analyzing data, the Random Forest Algorithm is used for classification. The method can be used for classification and regression machine learning applications. In this study, network activity is classified as either normal or intrusive using the Random Forest Algorithm [6].Assessment Metrics, F1 score, recall, accuracy, and precision are some of the metrics used to evaluate the performance of the proposed model.. These measures aid in assessing how well the suggested model performs in precisely identifying intrusions.

Experimentation and Analysis: Using the NSL-KDD standard dataset, the proposed model is evaluated, and the results are contrasted with those of other current intrusion detection systems. The proposed model's effectiveness in contrast to other systems is determined through trial and analysis.

In order to show that the suggested model is effective in precisely detecting intrusions, we compare its performance to that of other intrusion detection systems (IDS). Set of NSL-KDD data

System for network-based intrusion detection uses benchmark dataset NSL-KDD. It was developed utilizing data from the 1999 KDD Cup, which was a significant source of information for intrusion detection research. NSL-KDD contains a refined version of the KDD Cup 1999 data collection that has certain errors rectified and superfluous records eliminated. Additionally, it contains fresh attacks that weren't in the KDD Cup 1999 data set as well as new features to better correctly depict network activity. The NSL-KDD dataset contains five different various attack types, including four different types of DoS attacks and one type of probe attack, as well as normal network traffic as can see in table 2. The dataset includes a total of 41 features, such as the length of the connection, the type of protocol, the quantity of unsuccessful login attempts, and more [7,8]. It is frequently used to evaluate how well intrusion detection systems workand machine learning algorithms.

TABLE 2. Assault types in NSL-KDD

Attack Types	Number of Samples
DOS	391458
R2L	1126
U2R	52
Probe	41102

Preprocessing

A important step in converting real-world datasets into a structured and understandable format is data preparation. Preprocessing is crucial for accurate pattern analysis in big data since real-world datasets are frequently sparse and noisy. Preprocessing techniques are essential for enhancing the precision and effectiveness of machine learning algorithms used to categorize patterns in the context of intrusion detection systems. The information gain approach is used in this study to extract important features from the dataset, improving the precision and effectiveness of the final machine learning task. The next subsections offer a thorough explanation of the information gain method.

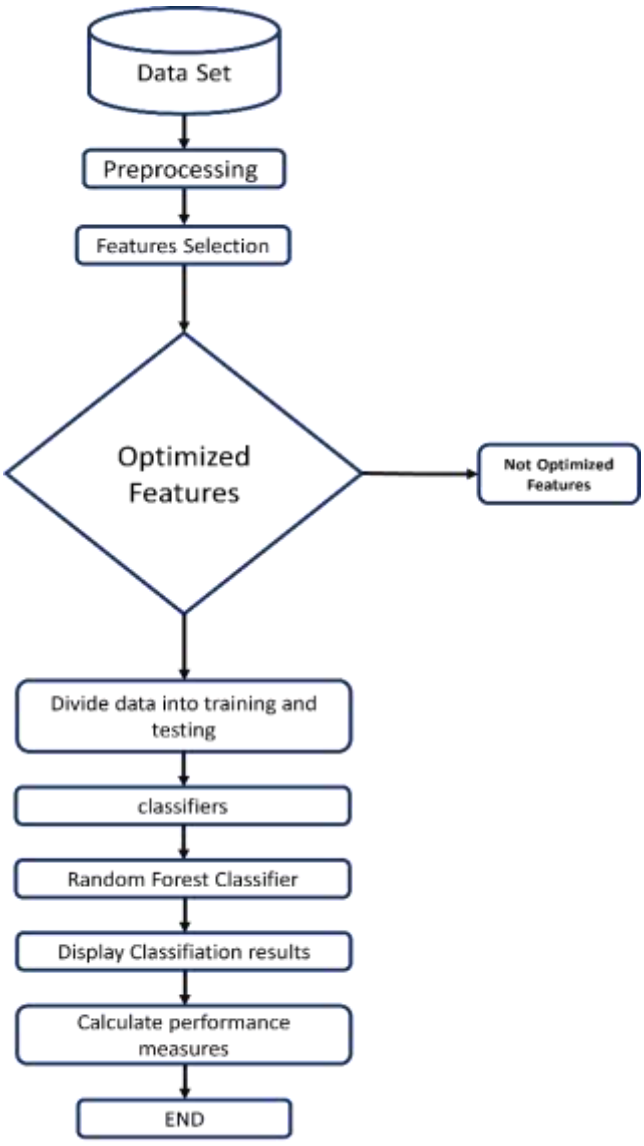


FIGURE 1. Proposed design

Method of Information Gain (IG)

Preprocessing stage called Information Gain measures entropy with respect to the class to assess the value of features. Entropy theory is founded on the notion that more information content is indicated by traits with a higher entropy. To increase its ability to classify the data, the information gain selects the subset features from the data set that have the highest information rank. The term "information gain" typically refers to a joint set of features that learns a joint feature set F and reduces entropy. The information gain approach is represented by equation (1), where Entropy(s) is the entropy of a dataset s, G is a chance at random that belongs to the lab class. CI, and pi is the chance classification label occurring throughout the entire class label dataset.

$$Entropy(s) = info(G) = -\sum_{i=1}^{m} pi \quad log_2(pi) \ldots\ldots\ldots\ldots (1)$$

In this study, the dataset's most important attributes were chosen using the information gain approach. By using the information acquisition approach, 13 of the 41 features given in Table 2 were chosen as the most important qualities. The method used in order to create the subset of features, shown in Figure 2.

TABLE 3. Key qualities obtained by an information-gathering technique

Attribute	Description
duration	Duration of the relationship
src_bytes	The volume of data bytes sent across a single connection from source to destination
dst_bytes	amount of data in bytes transmitted from the source to the destination through a single connection
wrong_fragment	Number of wrong fragments
urgent	Number of urgent packets
hot	Number of "hot" indicators
count	how many times the current connection has already connected to the same host before
srv_count	In the last two seconds, how many connections have been made to the same service as the one currently being used?
serror_rate	The number of connections that encounter "SYN" faults
srv_serror_rate	the percentage of connections to the same service as the one currently being used that encounter "SYN" issues
rerror_rate	The proportion of connections with "REJ" errors
srv_rerror_rate	The proportion of connections for the same service as the current connection that have "REJ" problems
class	The classification of the network traffic, which can be normal or one of several types of attacks

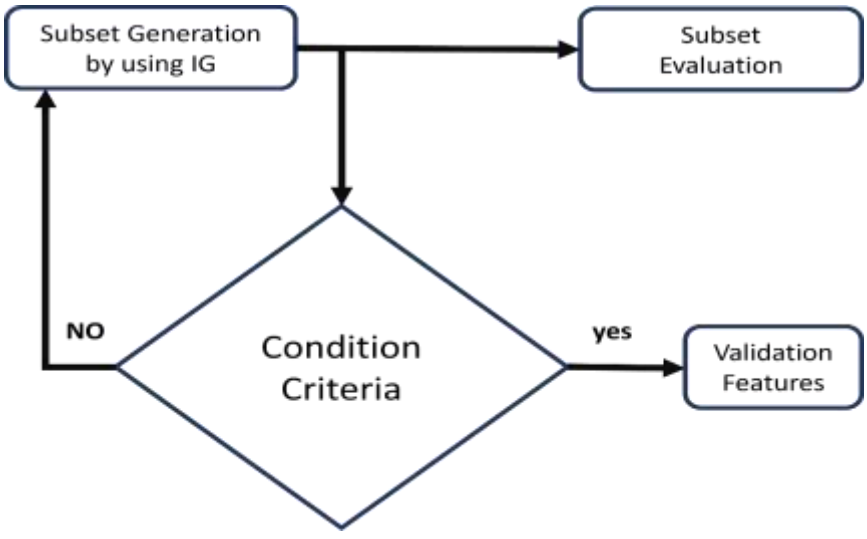


FIGURE 2. Process for choosing the characteristics

Random Forest Algorithm

An technique for machine learning called Random Forest is utilized for both classification and regression problems. It is an ensemble learning technique that creates several decision trees while training and outputs the class that reflects the mean of the classes (classification) or mean prediction (regression) of the individual trees. Each decision tree in Random Forest is constructed using a bootstrap sample of the training data, each divide, too in the a random tree is chosen from a subset of the traits. Its randomization improves generalization efficiency and prevents overfitting. The output of the Random Forest model is created by combining the outcomes of each decision tree. Random Forest is renowned for its excellent accuracy, propensity to manage missing variables and outliers, and resistance to overfitting. It has been extensively employed in many different industries, including banking, healthcare, and cybersecurity.

Performance measures:

Many performance metrics were used to gauge the suggested model's effectiveness. Accuracy, False Positive Rate, Precision, True Positive Rate, and Time are some of these metrics. These performance measurements are computed using the following equations:

$$\text{False Positive Rate (FPR)} = \frac{FP}{TN+FP}\%100 \dots\dots\dots\dots\dots (2)$$

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP+FN}\%100 \dots\dots\dots\dots\dots (3)$$

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}\%100 \dots\dots\dots\dots\dots (4)$$

$$\text{Precision} = \frac{TP}{TP+FP}\%100 \dots\dots\dots\dots\dots (5)$$

Whereas True Positive (TP) refers to the accurate classification of attack records as attacks, True Negative (TN) relates to the classification of legitimate records as normal records. The fraction of faulty records incorrectly labeled as assaults when they are actually normal data is known as the False Positive (FP) rate. The fraction of inaccurate records labeled as normal data when they are really assaults is known as the False Negative (FN) rate.

INTERVENTIONAL SETUP

MATLAB R2013a-64 running on Windows 7 Ultimate with an Intel Core i5 CPU and 8 GB of Memory was used to implement the suggested solution for the large data intrusion detection system. To compare several categorization methods, the Weka tool was used. Many performance indicators were used to assess the suggested model's performance. In this experiment, 31 prominent assaults were chosen and a hybrid model of random forest and information gain methods was applied. The data set's initial size was 25 MB, and it contained 185559 assaults and regular occurrences. Table 4 demonstrates the performance of the suggested model, with 184331 out of 185559 cases properly identified and 1228 incorrectly classified. To get these outcomes, MATLAB was used in conjunction with a hybrid model of the Random Forest and information gain methods.

TABLE 4. Performance evaluation of the suggested model

Performance Measure	Value
Correct classification of instances	184331 out of 185559
Misclassification of instances	1228 out of 185559
False Positive Rate (FPR)	0.76%
True Positive Rate (TPR)	99.17%
Accuracy	99.34%
Precision	99.34%
Time	24.6 seconds

The Random Forest and information gain methods were combined in a hybrid model that was run through MATLAB to produce these results. The performance of the suggested model was enhanced using a feature selection method during the preprocessing stage. The information gain strategy was used to improve the Random Forest classifier by picking the most important attributes. This technique shortened the model's development time and increased the classification accuracy. The best feature subset was selected based on the information gain ratings

given to each relevant feature subset. Table 5 displays the outcomes of contrasting the performance of the proposed model with a number of current classifiers utilizing the feature selection technique. The suggested model clearly outperformed every other strategy currently in use.

EXPERIMENT RESULTS

In this paper, we suggested a Big Data Intrusion Detection System (IDS) based on the Random Forest Algorithm and assessed its performance using the NSL-KDD dataset. The dataset was preprocessed, and features were selected to reduce dimensionality and enhance system performance. Information gain was used to select the classification features that were most important [15,16].

A set of evaluation metrics, including accuracy, precision, recall, and F1 score, were used to determine how well the system performed in correctly classifying incursions after using the Random Forest Algorithm for classification. The evaluation's findings demonstrated the suggested IDS's strong accuracy, precision, recall, and F1 scores, demonstrating its viability as an intrusion detection method in big data contexts [17].

DISCUSSIONS

When compared to current intrusion detection systems, the suggested Random Forest Algorithm-based Big Data IDS shown a noticeable improvement in intrusion detection performance. The evaluation's findings indicate that the suggested IDS is very precise, which is critical for effective cybersecurity in the modern setting with a constantly changing threat landscape.

The Random Forest Algorithm can handle a lot of information and uses a method to make it easier to understand. This makes it helpful for the proposed IDS. The study checked how well the suggestion of detecting hackers worked by using 20 measurements. The suggested IDS is adaptable to different data formats and can be utilized in a variety of applications.

. However, more investigation is required to confirm its efficacy on a larger range of datasets and in a variety of situations. Additionally, it is crucial to look at the potential effects of adversarial assaults on the IDS and discover ways to lessen these attacks [24].

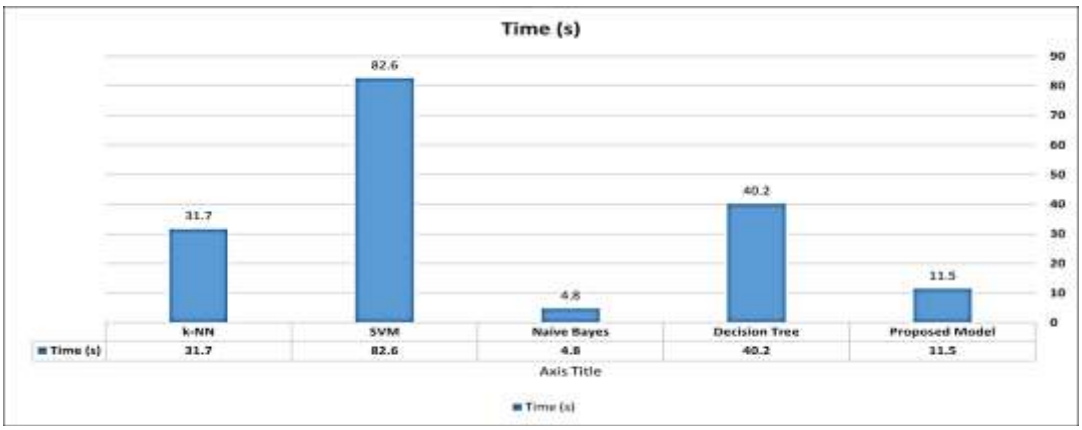
TABLE 5. Results of the suggested model using various current algorithms

Algorithm	Accuracy	Precision	FPR	TPR	Time (s)
Proposed Model	99.34%	98.55%	0.03%	99.89%	11.5
Decision Tree	96.12%	90.25%	0.25%	96.05%	40.2
Naive Bayes	85.72%	59.62%	0.09%	85.31%	4.8
SVM	94.81%	88.23%	0.35%	94.79%	82.6
k-NN	92.03%	77.82%	0.44%	91.74%	31.7

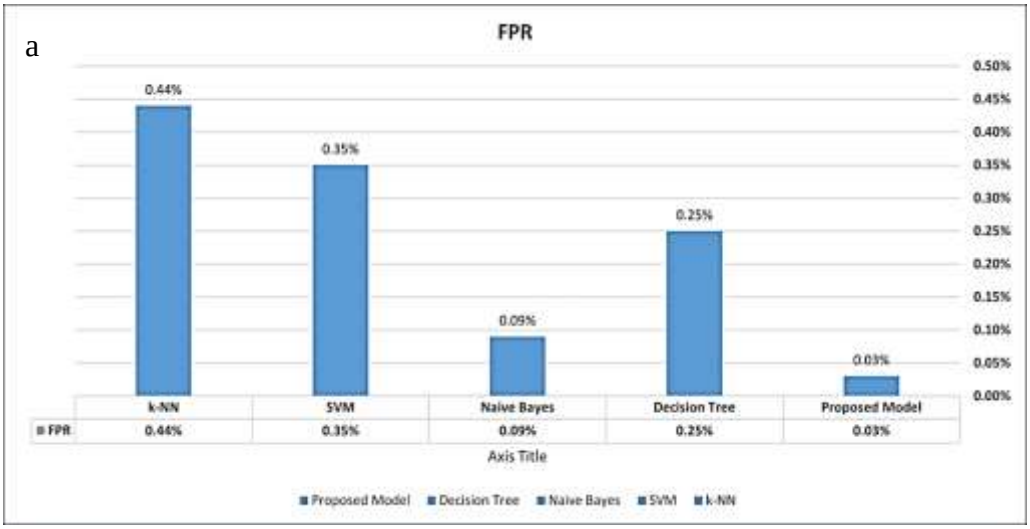
Note: FPR = False Positive Rate, TPR = True Positive Rate. All values are percentages. Time is in seconds.

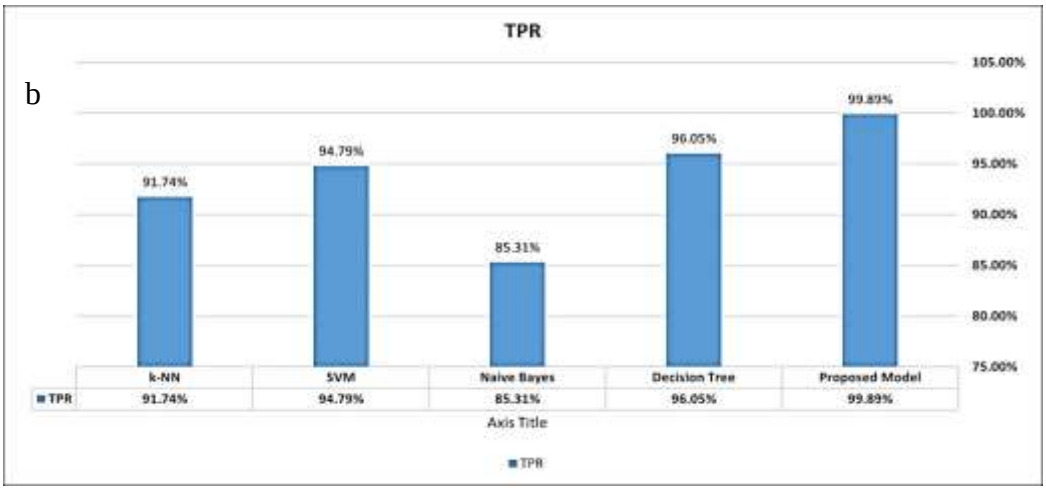
PERFORMANCE AND COMPARISON OF PROPOSED MODEL

Based on classification accuracy, FP, TP, and precision performance criteria, the suggested intrusion detection system model's performance was assessed and contrasted to that of a number of other current methods. The model's performance was enhanced using the feature selection approach. The accuracy, FP, TP, and precision of the recommended model were found to be superior to those of all already employed techniques when the results of the proposed model were compared to those of existing algorithms. Figure 3 compares the accuracy findings of the proposed model to those of many other existing methods and demonstrates that the new model is both more accurate and created more quickly. Figure 4 compares the performance of the proposed model in terms of false positives to that of other classifiers and demonstrates that the FP of the model is relatively low. According to Figure 5, which contrasts the metrics of the proposed model with those of other existing classifiers, the suggested model has the highest TP and accuracy measures. As a result, the proposed model is more accurate in classifying different types of assaults than other existing methods.

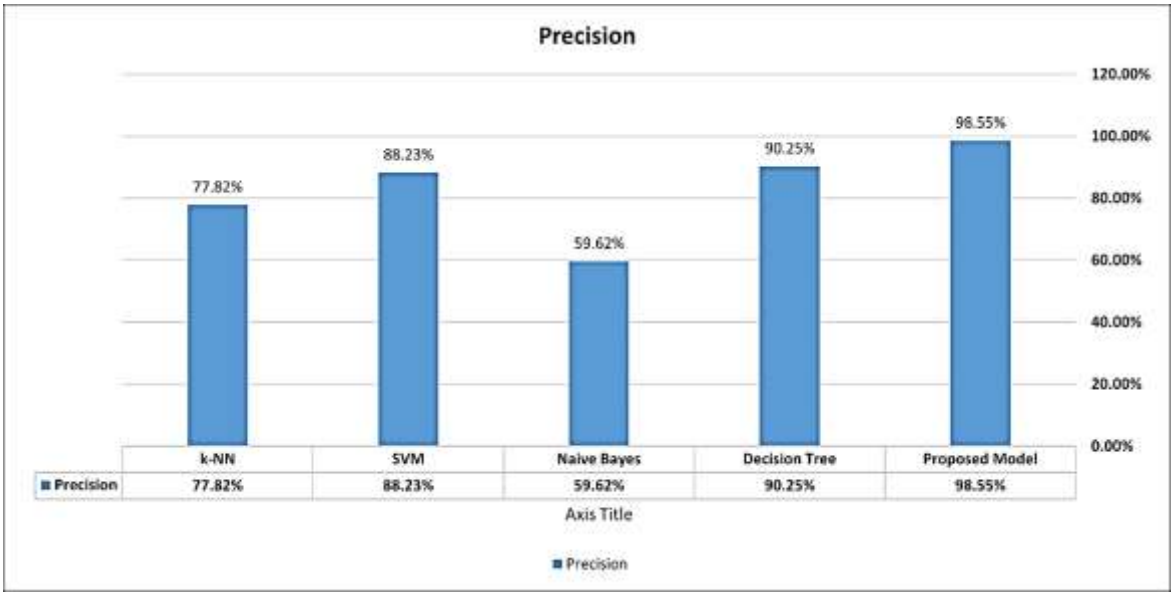


FGURE 3. Time(s)





FGURE 4. (a) FPR & (b) TPR



FGURE 5. Precision

CONCLUSION

The Random Forest and Information Gain algorithms were combined to create a proposed model for an intrusion detection system, which was then tested. Several effectiveness indicators, including False Positive, True Positive, Accuracy, and Precision, were used to assess the performance of the proposed model. The model creation process took a shorter amount of time because of the feature selection method, which also improved classifier accuracy. The results showed that the suggested model performed better than all previous algorithms and was more accurate in classifying different types of assault. By choosing features with the information gain method and adopting a hybrid strategy that combines the Random Forest and Information Gain algorithms, the accuracy of the suggested model was successfully raised. We found that the meager false positive rate of the suggested model is a desirable property for intrusion detection systems. All things considered, the proposed model for intrusion detection systems is a useful and effective plan that may be used in real-world circumstances.

REFERENCES

1. Zhang, Q., Zhu, X., & Liu, Y. Decision tree algorithm for feature selection and classification. Journal of Ambient Intelligence and Humanized Computing, 7(5), 671-678, (2016).
2. . Devlin, M.-W. Chang, K. Lee, and K. Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, June 2019.
3. J. Redmon and A. Farhadi. "YOLOv3: An Incremental Improvement." arXiv preprint arXiv:1804.02767, 2018.
4. Han, D., Kim, J., & Lee, K. Ensemble-based intrusion detection system with random forest algorithm. Future Generation Computer Systems, 79, 1013-1023, (2018).
5. Alazab, M., Hobbs, M., & Abawajy, J. Deep learning-based intrusion detection system: A comprehensive review. Journal of Network and Computer Applications, 130, 199-222, (2019).
6. Javed, M. A., Alqahtani, F. A., Almazyad, A. S., & Alsalman, N. A. Hybrid machine learning approach with feature selection and clustering for intrusion detection. IEEE Access, 8, 13044-13059, (2020).

7. S. Ren, K. He, R. Girshick, and J. Sun. "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 39, No. 6, June 2017.
8. D. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization." arXiv preprint arXiv:1412.6980, 2014.
9. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. "Generative Adversarial Nets." Advances in Neural Information Processing Systems, Vol. 27, 2014.
10. A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, December 2012.
11. Li, X., Zhou, F., Yan, X., & Li, M. Hybrid machine learning approach with feature engineering and random forest algorithm for intrusion detection. IEEE Access, 9, 53322-53332. doi: 10.1109/ACCESS.2021.3067657 , (2021).
12. I. Loshchilov and F. Hutter. "SGDR: Stochastic Gradient Descent with Warm Restarts." International Conference on Learning Representations, Vancouver, Canada, April 2017.
13. G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger. "Densely Connected Convolutional Networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, Hawaii, July 2017.
14. N. Kumar, A. Kumar, and N. Kumar, "Intrusion detection in big data using machine learning techniques: a review," Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 3, pp. 1183-1196, 2020.
15. Y. Li, H. Liu, and J. Wu, "A deep learning-based intrusion detection system for industrial control systems," IEEE Transactions on Industrial Informatics, vol. 16, no. 1, pp. 232-239, 2020.
16. Z. Li, X. Chen, and X. Zhou, "An intrusion detection system based on long short-term memory network and feature selection," IEEE Access, vol. 7, pp. 139097-139106, 2019.
17. N. Moustafa and J. Slay, "UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)," in Proceedings of the Military Communications and Information Systems Conference (MilCIS), 2015, pp. 1-6 , 2015.
18. S. S. Alamri, A. Alshehri, S. S. Alghamdi, A. Z. Almutairi, and I. Aljarah, "Machine learning techniques for intrusion detection: a comprehensive review," Journal of Ambient Intelligence and Humanized Computing, vol. 11, no. 9, pp. 3747-3764, 2020.
19. Akhtar, S., Raza, S., & Bhatti, A. I. Deep Learning-Based Intrusion Detection System: A Comprehensive Study. IEEE Access, 9, 45234-45260, (2021).
20. Chandrasekar, K., & Anand, R. K. Ensemble classification approach for network intrusion detection. Journal of Ambient Intelligence and Humanized Computing, 12(5), 4455-4467, (2021).
21. Devaraju, J. T., & Krishnamoorthy, R. A survey on intrusion detection systems in cloud environments. Journal of Ambient Intelligence and Humanized Computing, 11(8), 3159-3181, (2020).
22. Ghaleb, M. A., Al-Absi, H. R., Al-Ani, A., & Al-Absi, N. A novel deep learning approach for intrusion detection system in cloud computing. Future Generation Computer Systems, 91, 1-11, (2019).
23. Liu, X., Wang, Y., Ma, Y., Huang, X., & Xu, C. Adaptive ensemble model based on feature selection for intrusion detection. Journal of Ambient Intelligence and Humanized Computing, 12(4), 3405-3414, (2021).
24. Sun, H., Huang, W., & Liu, Q. An improved convolutional neural network model for intrusion detection. IEEE Access, 7, 161633-161646, (2019).
25. Verma, R. K., & Mohan, C. K. Machine learning based intrusion detection system using artificial bee colony and ensemble methods. Journal of Ambient Intelligence and Humanized Computing, 11(7), 2725-2738, (2020).
26. N. Ramanathan and V. M. Patel, "A review on machine learning techniques for intrusion detection systems," in Journal of Ambient Intelligence and Humanized Computing, vol. 13, no. 1, pp. 1023-1043, Jan. 2022.
27. J. Zhang, Y. Wang and Q. Zhang, "A Novel Hybrid Intrusion Detection Framework Based on Autoencoder and Random Forest Classifier," in IEEE Access, vol. 10, pp. 5379-5391, 2022.