IJITCE

ISSN 2347- 3657

International Journal of Information Technology & Computer Engineering

Email : ijitce.editor@gmail.com or editor@ijitce.com

International Journal of
Information Technology & Computer Engineering

ISSN 2347–3657
Volume 12 , Issue 1, Mar 2024

Optimizing prediction of security data using feature selection and ensembling

S.Sandhya
Assistant Professor,
Department of Computer Science
G. Narayanamma Institute of
Technology and Science
Hyderabad, Telangana
s.sandhya@gnits.ac.in

Donthala Harshitha
Department of Computer Science
(Data Science)
G. Narayanamma Institute of
Technology and Science
Hyderabad, Telangana
dharshitha2413@gmail.com

Gangidi Anisha
Department of Computer Science
(Data Science)
G. Narayanamma Institute of
Technology and Science
Hyderabad, Telangana
gangidi10@gmail.com

Akula Varshini
Department of Computer Science
(Data Science)
G. Narayanamma Institute of
Technology and Science
Hyderabad, Telangana
akulavarshini@gmail.com

Sambu Sarayu
Department of Computer Science
(Data Science)
G. Narayanamma Institute of
Technology and Science
Hyderabad, Telangana
sarayu0803@gmail.com

Abstract— **Network security is becoming increasingly difficult in today's hyperconnected environment and network traffic and infrastructure must be protected since attacks on businesses are increasing. Anomaly based Intrusion Detection System models identify anomalies as a deviation from the expected behavior. With Machine Learning, the system can learn patterns of normal behavior across environments and applications and it offers the ability to find complex correlations in large amounts of data for detecting attacks. Machine learning algorithms working on large datasets with multi attack types increase computational time and create a problem in decision making.**

In this work, an Intrusion Detection System model with ensemble feature selection technique is developed to reduce large-scale datasets and improve feature selection and accuracy prediction of the model using ensemble machine learning algorithms

Keywords— Network Security, Ensemble Learning, Feature Selection, Machine Learning, Intrusion Detection System

I. INTRODUCTION

The purpose of cybersecurity [1] is to create a strong security posture that defends against a wide range of threats while also ensuring the confidentiality, integrity, and availability of systems and data. Cyberattacks can have disastrous consequences, resulting in financial losses, brand damage, and lost privacy. Effective cybersecurity measures are crucial for protecting key infrastructure, intellectual property, personal data, and national security. The selection of features is critical in improving the effectiveness of cybersecurity systems. It entails recognizing and picking the most important qualities or features from a bigger collection of options.

Feature selection [2,3] enhances model performance by lowering dimensionality and focusing on the most useful attributes. Feature selection [4] plays an important role in increasing model performance, lowering complexity, improving interpretability, and adjusting to evolving threats. Integrating feature selection methodologies with cybersecurity practices empowers organizations to build robust defenses and ensure the integrity and security of their digital assets.

Ensemble learning is one of the most powerful machine learning [5] strategies for solving a specific computational intelligence problem by combining the output of two or more models/weak learners. A Random Forest algorithm, for example, is a collection of several decision trees [6] merged. The aim of ensemble feature selection is to integrate numerous feature selection approaches, considering their strengths, to provide an optimal best subset.

Machine learning is important in network security because it provides sophisticated approaches and tools for detecting and mitigating various cyber threats. Machine learning algorithms, with their real-time reaction and adaptive learning process, contribute significantly to cybersecurity [7] by lowering the computational time necessary to evaluate vast amounts of data with duplicate or meaningless features.

The key to increase the performance of training a model with a large dataset that contains a variety of attack types. However, problems like large dataset dimensionality represent a danger to most of these solutions, making it difficult to implement them on older systems and respond in real time. Our paper discusses ensemble feature selection to get the subset of best features and using these features to train machine learning algorithms. These algorithms also use ensemble technique to give out the best results.

II. Literature Review

One of the important datasets in cybersecurity world is NSL-KDD [8] which was developed from KDD Cup 1999 dataset to remedy some of the original dataset’s flaws and limitations. This dataset contains network traffic data which has different kinds of assaults and typical activity. It includes attacks such as Denial of Service (DoS), Probe, Remote to Local (R2L), and User to Root (U2R). The collection also contains elements that were taken directly from network packets, including content-related variables, statistical characteristics, and basic header information.

There were many methods used on NSL-KDD dataset to demonstrate the performance.

1. *Mutual Information and Information Gain:* Mutual information and information gain are statistical metrics used to analyse the significance of attributes in relation to the goal variable. These metrics quantify the quantity of information produced by each feature and can be used to help with feature selection in cybersecurity applications.

2. *Recursive Feature Elimination (RFE):* RFE is an iterative method that begins with all features and eliminates the ones that are least important in each iteration. It uses a machine learning model to rank features based on their coefficients or relevance scores, then chooses the top-ranking features for the final subset.

3. *L1 Regularization (Lasso):* L1 regularization [9] is often used in cybersecurity initiatives for feature selection. L1 regularization favours sparse solutions by including a penalty term depending on the absolute values of the coefficients, resulting in the selection of features with non-zero coefficients.

4. *Tree-based Feature significance:* Tree-based models, such as random forests [10] and gradient boosting algorithms, assign feature significance scores depending on each feature's contribution to the model's prediction performance. Features with higher significance scores are deemed more significant and are eligible for inclusion in the final feature subset.

5. *Correlation Analysis:* Correlation analysis aids in the identification of correlations between features and the target variable. Features with a high correlation or mutual information with the target variable may be deemed significant and included in the feature subset.

6. *Hybrid Methodologies:* Hybrid feature selection strategies integrate different methods to capitalize on their strengths and increase selection performance. Combining filter-based techniques (e.g., information gain) with wrapper-based techniques (e.g., RFE), for example, might result in more robust and accurate feature subsets in cybersecurity initiatives.

7. *Bagging and boosting:* These methods were also utilized with specific features in the NSL KDD datasets that resembled an Internet of Things (IoT) sensor node attack.

III. Challenges In Existing System

Before you begin to format your paper, first write and Machine learning (ML) techniques in network security presents several challenges. Some of the key challenges include:

1. *Inadequate and Representative Training Data:* To learn successfully, ML models require huge volumes of high-quality training data. However, getting labelled datasets for network security is frequently difficult due to the scarcity of real-world attack data and the requirement for expert expertise to appropriately label the data. Furthermore, network settings and attack strategies develop over

time, necessitating regular updates to the training data.

2. *Uneven Data Distribution:* In network security, assaults occur far less frequently than typical network traffic, resulting in uneven datasets. Machine learning models based on skewed data may exhibit biased behavior and struggle to identify infrequent or unique assaults. To solve this issue, techniques such as oversampling, under sampling, and the use of cost-sensitive learning algorithms are used.

3. *Adversarial Attacks:* Adversarial attacks are purposeful attempts to alter machine learning models by exploiting flaws. In the context of network security, attackers can create malicious network traffic or change existing traffic to avoid detection by machine learning-based security systems. Adversarial attacks can cause model poisoning, evasion, or data poisoning, posing substantial hurdles to the efficacy and resilience of machine learning-based security solutions.

4. *Interpretability and Explainability:* ML models used in network security are frequently regarded as black boxes, which means it can be difficult to comprehend how and why they produce specific predictions or choices. Interpretability and explainability are critical in security applications because understanding the reasons behind a model's decisions is critical for effectively responding to attacks and identifying false positives or false negatives.

5. *Generalization to New and Evolving Threats:* As network security threats evolve, ML models trained on historical data may struggle to generalize to new and unknown attack methodologies. To respond to emerging risks, ML models may require ongoing retraining and updating. Techniques such as transfer learning and continuous learning can help increase the model's ability to handle new threats.

6. *Constraints on Resources:* ML models, particularly complicated deep learning architectures, can need large computing resources, memory, and processing power. Due to restricted hardware capabilities and real-time processing needs, deploying such resource-intensive models on network security devices such as firewalls or intrusion detection systems can be difficult.

7. *Privacy and Ethical Concerns:* ML models in network security frequently demand access to sensitive network traffic data, generating privacy and data protection concerns. To preserve privacy while yet allowing effective model training, proper data anonymization and encryption techniques must be used.

To address these issues, a mix of domain expertise, data collecting tactics, feature engineering, model selection, and ongoing monitoring and adaptation are required.

Our model uses Ensemble learning method to select the best features from NSL-KDD [11,12] and trains a model which again is an ensemble of different machine learning techniques like XG Boost [13] , Random Forest, KNN Classifier, Decision Tree, Naïve Bayes and Logistic Regression. Feature Selection methods used in the ensemble are Select K-Best, Univariate F test, XG Boost, Information Gain and Correlated Feature Elimination [14].

IV. METHODOLOGY

A. *Data Pre-Processing*

The dataset used in this study is the NSL-KDD dataset, and it is the successor to the KDD'99 dataset. The NSL-KDD dataset was created in order to solve flaws in its predecessor. The KDD'99 dataset, based on the Intrusion Detection System (IDS) [15] evaluation program, included a substantial amount of fake data. A study of this dataset indicated that nearly 78% of records on the train dataset were duplicated, while 75% of records on the test dataset were duplicated. This redundancy gave unneeded bias to records that were more broadly available in the dataset compared to records that were not present in a substantial enough quantity to offer enough information for the machine learning model to train on.

NSL-KDD addressed this issue by excluding duplicate records from both the test and train datasets. NSL-KDD is made up of test and training datasets. The training dataset contains 125,973 records, whereas the test dataset contains 22,544 records, each of which contains 42 attributes that can be used for prediction. Protocol_type, service, and flag are the three category attributes. Variables in these category attributes were one-shot encoded before being used as training input.

One-hot encoding is a common technique used to convert categorical variables into numerical representations that can be effectively used in machine learning models. The One-hot encoding approach converts categorical variables to integers and provides a one-hot encoding approach converts the numerical data into categorical variables. The one-hot encoding is done by utilizing a unit vector for each category with 0's and 1's. One-hot encoding was applied to the categorical variables in the NSL-KDD dataset using a similar method. TCP (Transmission Control Protocol), UDP (User Datagram Protocol), and ICMP (Internet Control Message Protocol) are all included in the feature protocol_type. It can be expressed as [1, 0, 0], [0, 1, 0], and [0, 0, 1] when one-hot encoded. Protocol_type, service, and flag each had 3, 70, and 11 variables in the training dataset, which were one-hot-encoded before to training. On this dataset, previous research also employed one-hot encoding. Where protocol type, flag and service were changed into numerical values. One-hot encoding assumes that both the training and test datasets have an equal number of variables in a feature column. This was not the case for the feature service because the number of variables in the test and training data were different. There were 64 variables in the test dataset and 70 in the training dataset.

To solve this issue, fictitious records with missing fields were produced in the test dataset before performing one-hot-encoding. The application of one-hot-encoding on the previously indicated categorical variables, combined with the existing continuous attributes, resulted in 123 training features from the previously existing 42 features.

The results are observed and performing one hot encoding and scaling is done for those columns. Numbers are allocated to each attack type/virus which categorizes the data into different types of attacks like DOS attacks, U2R attacks, R2L attacks, probe attacks, etc.

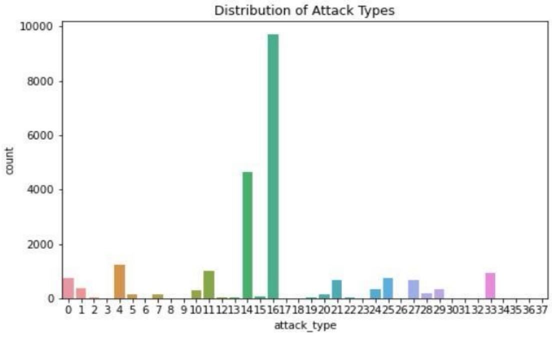


Figure 1: Distribution of Attack Types

B. Feature Selection techniques Application Using Ensemble Learning

Cybersecurity datasets typically contain many features, not all of which are equally informative or contribute significantly to the prediction task. Using all available features, on the other hand, can result in the inclusion of irrelevant or duplicated information, which can have a negative impact on the performance of machine learning models.

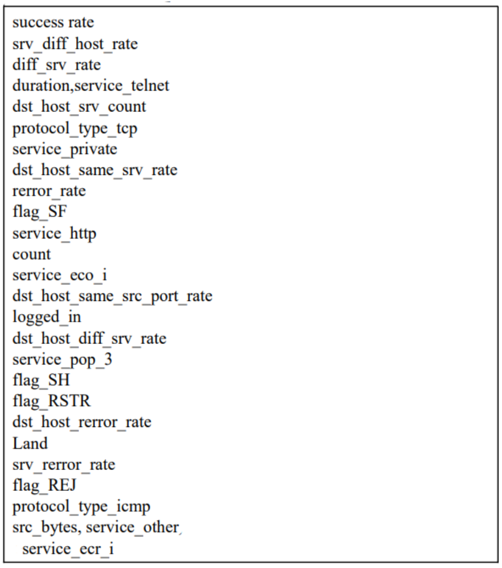
The feature selection procedure decreases the dimensionality of the data by picking the most useful features, which might have numerous advantages. First, it increases the computational efficiency of machine learning algorithms by requiring them to process fewer features. Second, it helps to alleviate the "curse of dimensionality," which refers to the difficulties and restrictions that come with dealing with high-dimensional data. The model can make better use of available resources and concentrate on the most important aspects by focusing on relevant aspects of cybersecurity problem.

The best features are assembled using XG boost, Univariate F Test, Select K Best, Correlated Feature Elimination, Information Gain. The output of various feature selection strategies was able to cut the number of features by more than 50% while greatly increasing predictive performance. The Feature Selection procedure also recommended the best algorithms from among all the machine learning algorithms employed, which are replicable and suitable for deployment. Ensembling is performed on feature selection techniques using probability voting to provide best features by considering the best 30 attributes which resulted from the five techniques mentioned above. The outputs of these ensemble learning were documented and utilized later in the process to compare the outcomes before and after the feature selection strategies were used.

Using the Select K-Best feature selection method, 30 best features are provided from the given dataset, which is represented by the representative 'k' as 30, and again 30 best features are also provided using the XG Boost, Univariate F Test, and Information Gain approaches. Using the Correlated Feature Elimination approach, on the other hand, we limit the threshold by 0.8 measure, and that threshold yields 10 best features that are required for further processing.

Obtaining the 30 best features from the above five feature selection techniques:

1. Select a few traits from each approach and combine them into one collection.
2. Vote to choose the most popular features.
   feature_votes = {feature: 0 for feature in all_selected_features}
3. Add up the votes for every feature.
4. Depending on the most popular votes, choose the top K features.
   k = 30
   best_features = sorted (feature_votes, key=lambda x: feature_votes[x], reverse=True) [:k]
5. Printing the best features out there.



success rate
srv_diff_host_rate
diff_srv_rate
duration,service_telnet
dst_host_srv_count
protocol_type_tcp
service_private
dst_host_same_srv_rate
rerror_rate
flag_SF
service_http
count
service_eco_i
dst_host_same_src_port_rate
logged_in
dst_host_diff_srv_rate
service_pop_3
flag_SH
flag_RSTR
dst_host_rerror_rate
Land
srv_rerror_rate
flag_REJ
protocol_type_icmp
src_bytes, service_other,
service_ecr_i

Table 1: Best Features Selected from the Ensemble of Feature Selection Techniques

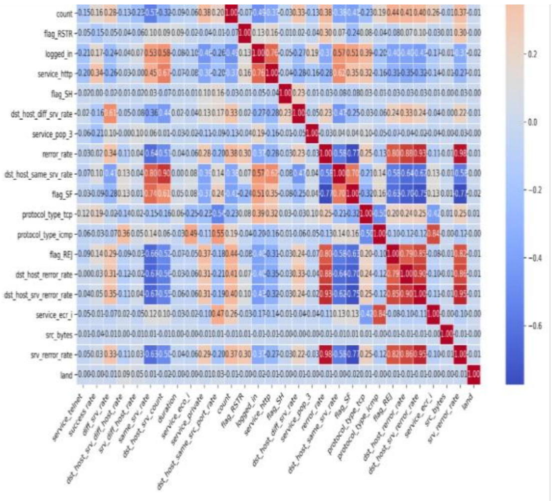


Figure 2: Correlation Heatmap of Selected Features

C. *Training Machine Learning Algorithms Using Ensemble Learning*

A machine learning technique called Ensembling [16] combines several separate models (commonly referred to as "base models" or "weak learners") to produce a single prediction. Ensembling is done to increase the model's generalizability and overall predictive performance in comparison to utilizing a single model. The concept of "wisdom of the crowd" or "collective intelligence" underlies assembling. The ensemble can take use of the strengths of several models and make up for their deficiencies by mixing the predictions of various models.

As a result, forecasts may become more reliable and accurate, and complicated relationships in the data may be handled more effectively.

After retrieving all the best features from the feature selection procedures, the most 30 common characteristics which are generated by Majority Voting method are used to produce a new dataset. After the new dataset is formed, it is separated into training and testing data and normalized once more. Logistic Regression, Naive Bayes, KNN, Decision Tree, AdaBoost, Random Forest, and SVM are all reapplied to the new dataset. The performance metrics are then scrutinized. The soft voting approach is used to ensemble machine learning algorithms, and the performance metrics are obtained as a result.

International Journal of Information Technology & Computer Engineering

ISSN 2347–3657
Volume 12 , Issue 1, Mar 2024

RESULTS AND DISCUSSION

The project mainly builds an ensembling model to compare the results of the NSL-KDD Test+ dataset used in cyber security by taking the best features through feature selection and applying machine learning algorithms to these selected features. This ensembling machine learning model uses feature selection techniques that ultimately enhance the overall resilience and effectiveness of security measures. The target variable in this dataset is the attack type, which is used for the working of the project. Firstly, the raw data is used to check the performance metrics for different machine learning algorithms.

The dataset before feature selection [17] includes the one hot encoding step that converts the categorical values like protocol_type, service, and flag to numerical. These categorical values, along with the different attack types in the dataset, are used to perform label encoding, which brings up the total number of columns in the dataset from 42 to 78. This is when the feature selection techniques are performed on the dataset containing the attack types in order to get the best features.

Using the best features obtained by the different feature selection techniques, an ensemble model is built. Building a model for ensembling first requires the process of majority voting, which considers the most common and best features from the feature selection techniques. These features are then formed into a dataset, which is used to analyze the performance metrics of different machine learning algorithms. As we can see from the table, the results before and after selection have improved a lot. The metrics before and after feature selection for Logistic Regression show a drastic change of around 0.075, the Naive Bayes show an improvement of around 0.209, and the Ada Boost shows an increment of 0.011.

On the other hand, the KNN Classifier, Decision tree, and Random Forest [18,19] have a slight change. After performing all these metrics on the machine learning model, soft voting is performed that considers the probability scores of these machine learning algorithms and finally builds a model.

Machine Learning Algorithms	Results before Feature Selection	Results from Best Features Dataset
Logistic Regression	0.670	0.745
Naïve Bayes	0.556	0.765
K Neighbors Classification	0.965	0.958
Decision tree classification	0.985	0.978
Ada boost classification	0.465	0.476
Random forest	0.990	0.987

Table 2: Machine Learning Algorithm results from before Feature Selection and after Best Features Dataset

The ensemble learning model produced an accuracy of 99%.

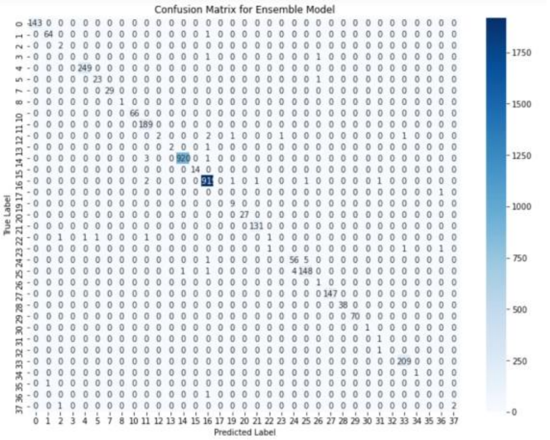


Figure 3: Confusion Matrix Graph for Ensemble Model

CONCLUSION AND FUTURE WORK

In conclusion, Feature Selection is a technique used to improve machine learning prediction in cybersecurity by selecting relevant features for analysis. By adapting to changing data patterns and incorporating real-time insights, feature selection helps identify key variables contributing to cybersecurity threats and attacks. Traditional approaches often rely on static feature sets, which may not capture the evolving nature of cyber-attacks, leading to suboptimal predictions and increased false positives. The feature selector leverages algorithms and techniques to continually learn and improve predictive capabilities over time. This approach improves accuracy, faster processing times, and increased resilience against evolving threats, enabling organizations to strengthen defenses, enhance incident response capabilities, and reduce the risk of data breaches or system compromises.

Future research aims to extend the Feature Selection algorithm for reinforcement learning and apply self-learning techniques to improve its performance. This will enable the model to be applied to datasets from other domains and enhance cybersecurity. These advancements will contribute to more accurate threat detection, improved incident response, and strengthened cybersecurity defenses in an evolving threat landscape.

REFERENCES

[1] R. Chivukula, T. Jaya Lakshmi, L. Ranganadha Reddy Kandula and K. Alla, "A Study of Cyber Security Issues and Challenges," 2021 IEEE Bombay Section Signature Conference (IBSSC), Gwalior, India, 2021, pp. 1-5, doi: 10.1109/IBSSC53889.2021.9673270.

[2] Ahsan M, Gomes R, Chowdhury MM, Nygard KE. Enhancing Machine Learning Prediction in Cybersecurity Using Dynamic Feature Selector. *Journal of Cybersecurity and Privacy*. 2021; 1(1):199-218. https://doi.org/10.3390/jcp1010011

[3] Girish Chandrashekar and Ferat Sahin. 2014. A survey on feature selection methods. Comput. Electr. Eng. 40, 1 (January, 2014), 16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024

[4] R. Thomas and D. Pavithran, "A Survey of Intrusion Detection Models based on NSL-KDD Data Set," 2018 Fifth HCT Information Technology Trends (ITT), Dubai, United Arab Emirates, 2018, pp. 286-291, doi: 10.1109/CTIT.2018.8649498.

[5] I. Abrar, Z. Ayub, F. Masoodi and A. M. Bamhdi, "A Machine Learning Approach for Intrusion Detection System on NSL-KDD Dataset," 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2020, pp. 919-924, doi: 10.1109/ICOSEC49089.2020.9215232.

[6] Panigrahi R, Borah S, Bhoi AK, Ijaz MF, Pramanik M, Kumar Y, Jhaveri RH. A Consolidated Decision Tree-Based Intrusion Detection System for Binary and Multiclass Imbalanced Datasets. Mathematics. 2021; 9(7):751. https://doi.org/10.3390/math9070751

[7] Abdulla Hussain, Azlinah Mohamed, and Suriyati Razali. 2020. A Review on Cybersecurity: Challenges & Emerging Threats. In Proceedings of the 3rd International Conference on Networking, Information Systems & Security (NISS2020). Association for Computing Machinery, New York, NY, USA, Article 28, 1–7. https://doi.org/10.1145/3386723.3387847

[8] Dhanabal, L. and S. P. Shantharajah. “A Study on NSL-KDD Dataset for Intrusion Detection System Based on Classification Algorithms.” (2015).

[9] J. A. O'Reilly and W. Chanmittakul, "L1 regularization-based selection of EEG spectral power and ECG features for classification of cognitive state," 2021 9th International Electrical Engineering Congress (iEECON), Pattaya, Thailand, 2021, pp. 365-368, doi: 10.1109/iEECON51072.2021.9440359.

[10] P. Bolourchi, M. Moradi, H. Demirel and S. Uysal, "Random Forest Feature Selection for SAR-ATR," 2018 UKSim-AMSS 20th International Conference on Computer Modelling and Simulation (UKSim), Cambridge, UK, 2018, pp. 90-95, doi: 10.1109/UKSim.2018.00028.

[11] Revathi, S., & Malathi, A. (2013). A Detailed Analysis on NSL-KDD Dataset Using Various Machine Learning Techniques for Intrusion Detection. International journal of engineering research and technology, 2.

[12] A. Kothari, P. Vashishtha, P. Singh, M. Diwakar and N. K. Pandey, "Ensemble Methods on NSL-KDD," 2021 5th International Conference on Information Systems and Computer Networks (ISCON), Mathura, India, 2021, pp. 1-7, doi: 10.1109/ISCON52037.2021.9702439.

[13] Manju, N., Harish, B.S., & Prajwal, V. (2019). Ensemble Feature Selection and Classification of Internet Traffic using XGBoost Classifier. International Journal of Computer Network and Information Security.

[14] Herve Nkiama, Syed Zainudeen Mohd Said and Muhammad Saidu, “A Subset Feature Elimination Mechanism for Intrusion Detection System”

International Journal of Advanced Computer Science and Applications(IJACSA), 7(4), 2016. http://dx.doi.org/10.14569/IJACSA.2016.070419

[15] Ngoc Tu Pham, Ernest Foo, Suriadi Suriadi, Helen Jeffrey, and Hassan Fareed M Lahza. 2018. Improving performance of intrusion detection system using ensemble methods and feature selection. In Proceedings of the Australasian Computer Science Week Multiconference (ACSW '18). Association for Computing Machinery, New York, NY, USA, Article 2, 1–6. https://doi.org/10.1145/3167918.3167951

[16] Sarkar, A., Sharma, H.S. & Singh, M.M. A supervised machine learning-based solution for efficient network intrusion detection using ensemble learning based on hyperparameter optimization. Int. j. inf. tecnol. 15, 423–434 (2023). https://doi.org/10.1007/s41870-022-01115-4

[17] Powell, A., Bates, D., Wyk, C.V., & Abreu, D.D. (2019). A cross-comparison of feature selection algorithms on multiple cyber security data-sets. Fundamentals of Artificial Intelligence Research.

[18] Darst, B.F., Malecki, K.C. & Engelman, C.D. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet 19 (Suppl 1), 65 (2018). https://doi.org/10.1186/s12863-018-0633-8

[19] P. Bolourchi, M. Moradi, H. Demirel and S. Uysal, "Random Forest Feature Selection for SAR-ATR," 2018 UKSim-AMSS 20th International Conference on Computer Modelling and Simulation (UKSim), Cambridge, UK, 2018, pp. 90-95, doi: 10.1109/UKSim.2018.00028.