MACHINE LEARNING CLASSIFIER FOR INTRUSION IDENTIFICATION IN CYBER SECURITY APPLICATIONS

Abstract

Machine learning (ML) methodologies have been applied in many scientific areas due to their special traits, including adaptability, scalability, and the capacity to swiftly adapt to new and unexpected boundaries. Cyber security is a fast growing industry that needs a lot of attention because of the great improvements in social networks, and web technologies, online banking, mobile environments, etc**.** The many ML applications in cyber security are covered and highlighted in this article. This study examines a number of security-related topics, including network intrusion detection, evaluating the security properties of protocols, social network spam detection, energy use monitoring for smart metres, and security issues with machine learning algorithms itself. One of the most crucial steps in data analysis, data cleansing is the first phase of every ML project. In these situations, one-hot encoding produces very high dimensional vector representations, posing issues with memory and computational efficiency for ML models. Data normalisation is the process of proportionally scaling the data such that all values fall inside the desired range. It is discovered that Multi-Support Vector Machine (M-SVM) is the method with the greatest precision and accuracy. An Intrusion Detection System (IDS) is a sort of security tool that scans the system for suspicious activity and analyses network traffic before alerting the system or network administrator. The ID system ML algorithm performs the best with a Receiver Operating Characteristic (ROC) value of 0.961 compared to other ML algorithms.

Keywords: ML, M-SVM, IDS, Data cleaning, Data Normalization.

Authors

R. Sankar
Professor
Department of Electrical and Electronics Engineering
Chennai Institute of Technology
Kundrathur, Chennai, India.
rsankar0324@gmail.com

J. Joylin Mary
Assistant Professor
Department of Electrical and Electronics Engineering
SRM TRP Engineering College
Tiruchirappalli, India.
joylinmary@gmail.com

S. Sivarajan
Assistant Professor
Department of Electrical and Electronics Engineering
Vel Tech Multi Tech Dr. Rangarajan Dr Sakunthala Engineering College
Avadi, Chennai, Tamil Nadu, India.
ssivarajan78@gmail.com

V. Balaji
Associate Professor
Faculty of Electrical and Electronics Engineering
MAI-NEFHI College of Engineering and Technology
Asmara, Eritera.
balajieee79@gmail.com

I. INTRODUCTION

Cyber security practises include safeguarding hardware and software against unauthorised access, changes, destruction, and other threats [1]. Normally, any Cyber Security system should have antivirus software, firewalls and intrusion detection systems. IDS are important because they aid in identifying any unwelcome and undesirable system changes [2]. ML is a sort of Artificial Intelligence method that can automatically extract useful data from enormous datasets [3]. When there is a considerable amount of training data available and the machine learning models are sufficiently generalizable to distinguish attack variations and different assaults, ML-based IDSs can achieve good detection levels. Furthermore, because they do not rely on domain expertise, ML-based IDSs are simple to design and build.

The ML area known as deep learning can produce outstanding results [4]. Deep learning methods are highly helpful since they operate from beginning to finish and can automatically identify feature representations from raw data. For the creation of high-performing, accurate ML applications, the availability of high-quality data is often a need. Data cleaning, also known as data replacement, data modification, and data deletion, the act of correcting or erasing incorrect data from a data file. This includes processing erroneous values, missing values, and data rationality detection [5].The effectiveness of the categorization process is significantly influenced by the type of encoding technique used. A typical method for processing categorical data is one-hot encoding [6]. Categorical variables must be transformed into a format in order for ML models to be more effective at spotting instances of insider data leaking. To prevent incorrect interpretation of the correlations between independent variables, it only draws attention to the variables included in the features. Data normalisation is the process of proportionally scaling the data so that all values fall into the desired range [7].

A non-probabilistic binary classifier called M-SVM divides data into many groups. The binary division SVM accomplishes classification by classifying input data. SVM is a highly helpful tool for categorising undistributed and asymmetrically distributed data, which might include text, pictures, audio, and other types [8]. M-SVMs are based on the premise that two data classes can be separated by a margin on each side of a plane. An M-SVM is used to build a learning model based on supervised learning, which trains the model to categorise training data using prelabeled labels.

Additionally, assault strategies are developing daily, and the complexity of the mysterious offences that must be repelled is rising. In order to determine whether there is abnormal behaviour in the dataset and to offer trustworthy protection assistance for users or terminal equipment [9], IDs may detect and analyse network data .The development of deep learning offers a fresh approach to the Internet of Thing’s IDS study. In order to achieve real-time network status monitoring and the Internet of Things' current state is used as a network data set, and it is imported into the ML model for training and learning. Based on learnt normal and attack behaviour, machine learning-based IDS offers a learning-based approach to find attack classes [10].

This study makes a contribution to the ML classifier for intrusion identification in cyber security applications. Here is a method of cleaning data that incorporates traditional

cleaning, one-hot encoding, and normalisation of the data. ID System evaluates the data that is processed in cyber security in order to classify IDSs. Even with only a limited number of training data, M-SVM was able to produce the highest overall classification accuracy. Machine learning models offer adequate generalizability to identify cyber security threats if there is enough training data available for them.

The structure of this work is as follows: In the second section, similar works are suggested. In third section of current work, the suggested system is explored in greater detail. The study's results are summarized in Section fourth.

II. RELATED WORKS

Yakub Kayode Saheed et al (2021) illustrated how supervised ML may be utilised in the IoMT context to create an effective and efficient IDs for categorising and predicting unforeseen cyber threats. Network data is preprocessed and normalised. The standard data for IDs is used to do an SML. Extensive testing revealed that the proposed SML model exceeds previous methodologies with an accuracy of 99.76%. The effectiveness of the proposed approach in detecting IoMT attacks made with blockchain technology is not evaluated. To better serve the IDS by using real-time datasets, considerable research on ML algorithms must be improved.

Xianwei Gao *et al* (2019) described current IDS difficulties and recent advances while recommending an adaptive ensemble learning strategy. Comparing the ensemble model to previous research publications has shown that the ensemble model considerably improves detection accuracy. Additionally, data analysis reveals that the quality of the data features is a crucial factor in influencing the efficacy of detection. For better results, future feature selection and IDs data preparation efforts could be enhanced.

Mohamed Goudjil *et al* (2018) suggested a brand-new active learning technique for classifying texts. The basic goal of active learning is to intelligently choose which samples should be labelled in order to decrease the labelling effort without sacrificing classification accuracy. Using a collection of multi-class SVM classifiers to produce posterior probabilities, the suggested technique chooses a batch of informative samples, which are then manually labelled by a subject matter expert. According to experimental findings, the suggested active learning strategy greatly decreases the labelling effort while also improving classification accuracy.

Mohamed Amine Ferrag *et al* (2022) evaluated IDS for Agriculture 4.0 cyber security. Particularly the performance evaluation metrics for a IDSs for Agriculture 4.0 and contemporary cyber security issues. According to the machine learning strategy used, classify IDS thoroughly for each emerging technology. Lastly, we go over problems and possible research directions for Agriculture 4.0 cyber security ID. However the performance of the IDS have to be improved.

Darshana Upadhyay et al (2020) developed a framework for an ID) for smart grids that combines feature engineering-based preprocessing with ML classifiers. The bulk of ML algorithms, however, focus on fine-tuning the hyper-parameters in order to increase the detection rate. After utilising a GBFS module to extract the most interesting traits from the

power grid dataset, and tested a number of decision-tree based machine learning techniques. There has to be extensive study on ML algorithms to advance our ability to use real-time datasets to better serve the IDS.

III. PROPOSED SYSTEM

The proposed IDs for Cyber Security is shown in Figure 1. Here is a method of cleaning data that incorporates traditional cleaning, one-hot encoding, and normalisation of the data. ID System evaluates the data that is processed in cyber security in order to classify IDSs. ML models can generalise well enough to identify cyber security, and with enough training data, ML-based IDSs can attain respectable detection levels.

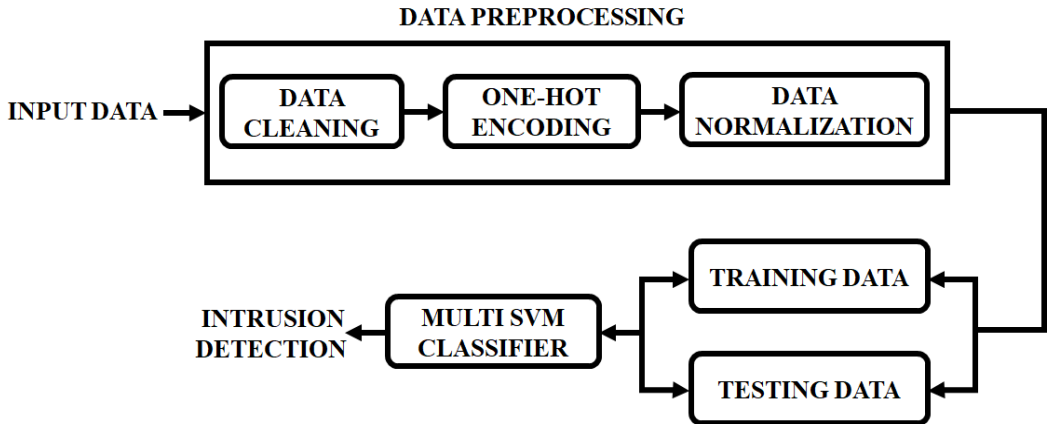

Figure 1: Proposed Intrusion Detection for Cyber security

1. **Data Cleaning:** Every ML project starts with data purification, one of the most important processes in data analysis. It is an important phase in ensuring that the dataset is devoid of erroneous or misleading information. It might be done either manually using tools for data wrangling or automatically with a computer software. Data preparation for analysis entails a variety of activities known as data cleansing. Due to its significance across many different industries, a growing number of people are interested in developing efficient and effective data cleansing systems. However, data is seldom correct in practise due to erroneous inputs from manual data curation or unintended mistakes from robotic data collection or generation operations.

   Real-world datasets frequently contain inconsistencies and gaps brought on by malfunctioning sensors or human mistake, for instance, which might have an effect on the machine learning systems based on those datasets. Structured data at scale that has to satisfy integrity requirements, denial restrictions, and functional dependencies is traditionally cleaned using schema. Recently, efforts have been made to enhance data validation methods and machine learning pipeline accuracy. The urgent problems of model fairness and model robustness against adversarial data are not, however, solved by these strategies.

Inconsistencies and mistakes existing in the training data might prevent algorithms from finding patterns, therefore data cleaning to assure consistency of the training data is a crucial step for preserving the model performance. Ineffective data management can waste resources, reduce productivity, and waste marketing revenue. The aforementioned procedure is used to segment pictures that have been distorted by high density impulse noise in order to achieve efficient filtering.

Data cleaning, also known as data replacement, data modification, and data deletion, include processing invalid values, missing values, and data rationality detection. It also involves correcting or removing incorrect data from the data file. The primary method of data cleansing is as follows, as illustrated in Figure 2:

- **Step 1:** Find the missing data Occasionally, throughout the data gathering process, there may be flaws or human mistakes that leave one or more data elements with a blank value. The positioning is achieved by changing positioning condition to a null value.

- **Step 2:** Because the sample data set utilised in this study includes few missing values, the strategy of instantly deleting this row of data is employed to deal with the missing values.

- **Step 3:** After removing missing numbers, the data will still contain some glaring errors. Because the information in these two columns actually should be kept in numerical form, the entire row of data involved in the problem will be deleted.

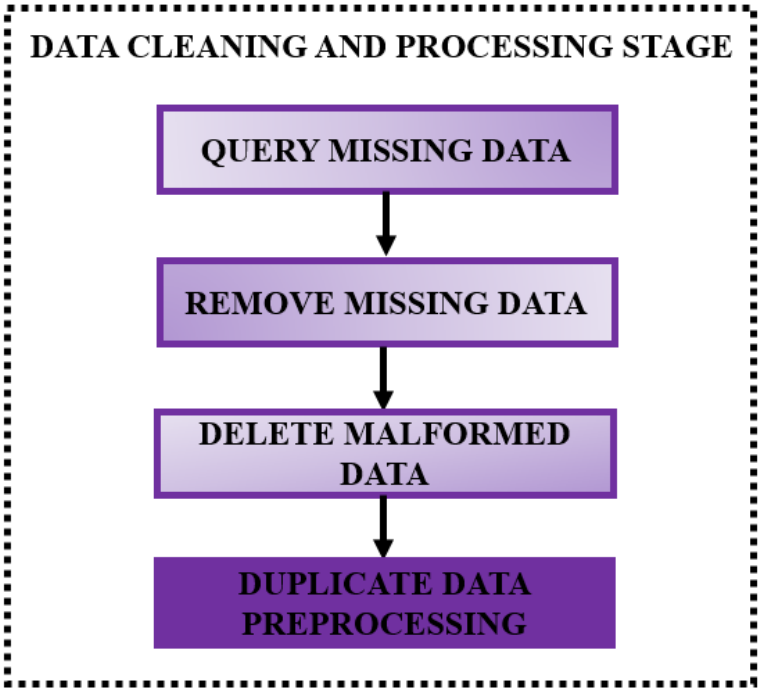

Figure 2: Data Cleaning Process

Data cleaning involves procedures like deleting duplicate entries in addition to activities like identifying missing values and content issues. Any logical mistakes or excessive data repetition will have an impact on the model's performance. After searching, no evidence of logical mistake is discovered, and the repetition of the data is low. As a result, no remedial action is done. However Encoding the classified data is necessary to prevent data leakage and for further classification.

2. **One Hot Encoding:** The performance of the categorization process is significantly influenced by the type of encoding technique used. Categorical data management frequently involves the use of one-hot encoding. In order for ML models to be more effective at finding instances of insider data leaks, categorical variables must be transformed into a format. It only draws attention to the variables that make up features in order to prevent incorrect interpretation of correlations between independent variables. One-hot encoding is a useful encoding technique for classification problems. The method of label encoding that was previously used is simpler than the one-hot encoding approach, however ordering problems might occur since some ML algorithms may not understand certain integer values. The one-hot encoding method is used to solve such ordering problems. Each category value is translated into a new column in one-hot encoding, and the label values are changed to either a digital version of (1 or 0).

Textual data cannot be directly processed by ML systems. Data must contain numbers. Because of this, the email text was encoded as one-hot vectors for this study's data preprocessing. A typical method for representing strings with a finite number of values is one-hot encoding, which employs a sparse vector with one member set to 1 and all other elements set to 0. High cardinality will result in high dimensional feature vectors when using one-hot encoding. However, one-hot encoding is a popular encoding technique since it is straightforward. One-hot encoding works well for tweets or phrases with few repeating parts and is typically used with models with high smoothing qualities. In neural networks, which need input to be in the discrete range of [0,1] or [-1,1], one-hot encoding is frequently utilised. One-hot vectors are 1 N matrices (vectors) that are entirely made up of 0s, with the exception of a single 1 in a cell that is used to specifically identify a word. Categorical data may be represented more expressively using one hot encoding. As illustrated in Table 1, an example word range of [good, good, terrible] would be expressed in 3 such encodings [0, 0, 1].

Table 1: One-hot encoding

1	0
1	0
0	1

There are a lot of discrete data in the data that the Internet of things collects. Since its data is in string format, it is regarded as digitising the labels in string format. The aforementioned information cannot be utilised in the model without first being translated from labels to numbers. The solution to this issue is One-Hot Encoding. The process entails encoding n states. Each state is represented by a single digit, hence n digits are required to represent n states. When a certain state of the result is present, the state's corresponding number is 1 and the other digits are 0. It is also clear that each feature will

become m binary features after one-hot encoding if it has m potential values. Additionally, only one of these characteristics may be active at once, and they are all mutually exclusive. Data will consequently become scarce. In addition to addressing the issue of data characteristics, one-hot encoding extends the experimental data set's with appropriate dimensions.

3. **Data Normalization:** Data normalisation is the process of proportionally scaling the data such that all values fall inside the desired range. The two primary benefits of normalising cleansing of data Quick encoding Normalisation of data after preprocessing data set Practise data Test results Develop a network model for IDs.

The standard deviation is not employed in order to more accurately imitate the data condition of the real incursion scenario. The maximum value will also need to be revised as new data is uploaded because it will change. In order to achieve the normalisation of the data's standard deviation, this experiment uses the standard deviation normalisation approach. Formula (1) displays the computation for normalising standard deviation. where x_{norm} is the normalized data set.

$$\chi_{norm} = \frac{\chi - \mu}{\sigma} \quad (1)$$

When, χ indicate the mean value, the standard deviation is, and the original sample data set for the Internet of Things is x. In order to have a good normalisation impact when applying standard deviation normalisation, the data distribution should typically be near to the Gaussian statistic.

4. **Multiclass SVM:** Binary classifications were the original focus of the traditional SVM classifiers. However, some real-world categorization issues have more than two classifications. For instance, in order to properly gauge or forecast the severity of the possible squeezing effect, it is necessary to classify the effect into a number of groups based on the size of the normalised convergence.

The "one-against-one" and "one-against-all" techniques are two often employed methods for building multiclass SVMs. This method is known as the "one-against-one" strategy. We employ a voting technique for the classification, whereby votes may be made for any of the samples and each binary classification is treated as a voting. After all voting is completed, a point is assigned to the class that received the most votes. The one-against-all method, on the other hand, involves creating as many binary classifiers as there are classes, and then using each trained classifier to distinguish one class from the others. We select the classifier with the highest decision function value to forecast a new instance. Depending on the output value, we decide whether to utilise the left or right M-SVM starting at the root node for an input data sample. The final classification of the input data sample would then be made based on the node's output value.

The relevant features for the categorization are chosen using the feature selection approach. A data collection containing only the most important elements improves the model's acceptance and accuracy. The classifiers are then trained using the list of important characteristics. Following classifier training, the test dataset is examined using

the same set of characteristics to determine whether each instance is normal data or attack data.

5. **Intrusion Detection:** An intrusion for an IDS is an effort to unlawfully or uninvitedly access computer system data or tamper with system operation. An intrusion detection system (IDS) is a computer security technology that aims to detect a wide range of security breaches, from insider abuse and system penetration efforts to outside incursion attempts. IDSs' main responsibilities involve tracking servers and networks, analysing computer system activities, producing alerts, and responding to suspicious behaviour. IDSs are frequently located near the protected network so that they can monitor related hosts and networks. IDS classification techniques fall into two categories: approaches based on data sources and methods based on detection.

IDSs fall under two categories among detection-based techniques: abuse detection and anomaly detection. IDSs can be categorised into network-based and host-based approaches among data source-based techniques. These two IDS categorization categories are combined in this study, with the data source serving as the primary classification factor and the detection technique serving as a supporting classification component.

IV. RESULTS AND DISCUSSION

1. **Attack Categories:** The work was conducted out using Python. The server's attack types are shown in Figure 3, with the Normal category having no attacks while the others do. The assault category graph's determined attack percentage is shown in the pie chart. The survey is compared in Figure 3 in terms of attacks. The ratio of assaults is represented by the number 1, which is equivalent to 55.06%, while the absence of attacks is represented by the number 0, which is equal to 44.94%.

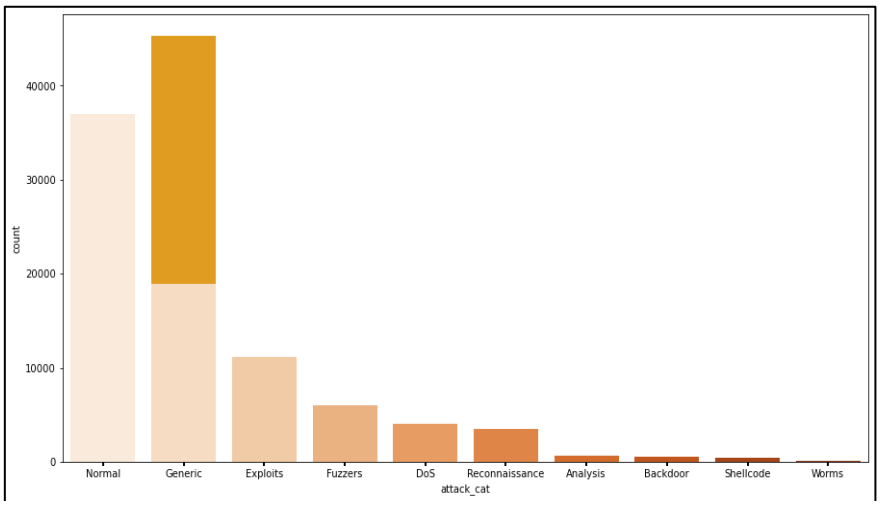

Figure 3: Attack Categories

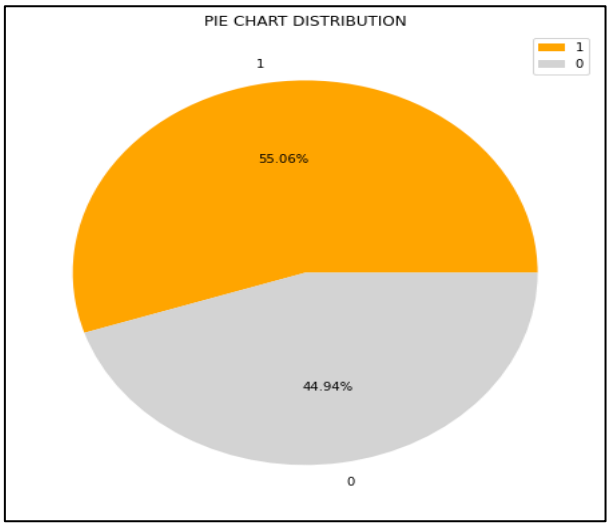

Figure 4: Classes

2. Protocols and Features

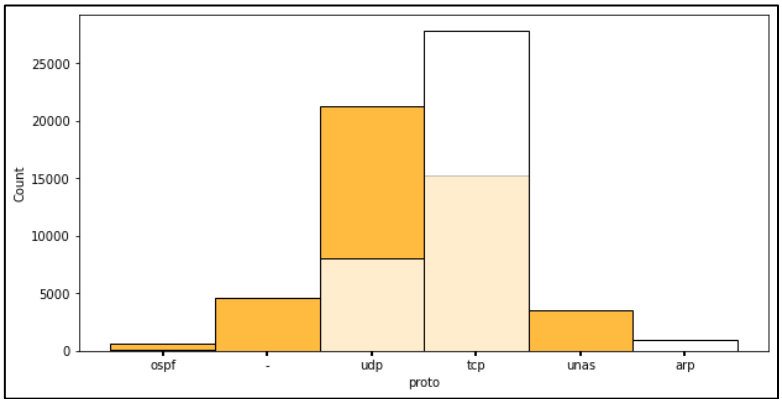

Figure 5: Protocols

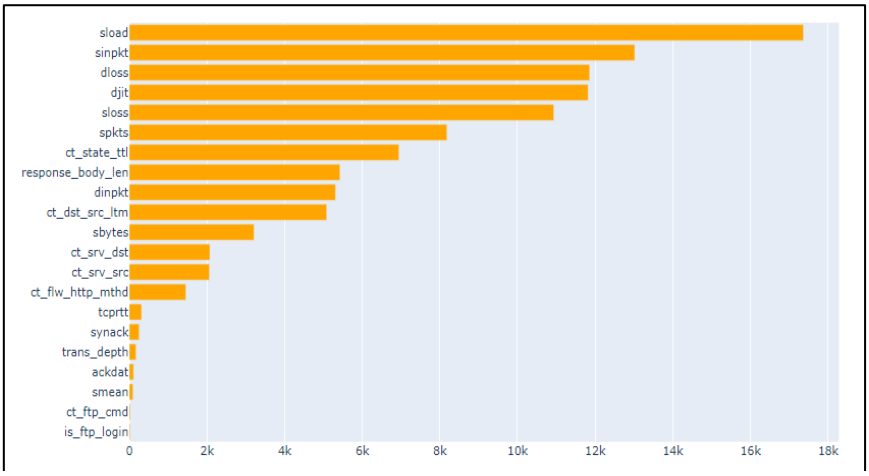

Figure 6: Features

The server and protocols are derived from IOT. The protocol indicates that the maximum number of counts for the tcp protocol is over 25,000. Figure 5. Shows a graph of proto-counts. The top 20 features out of 45 features in the dataset are displayed in Figure 6.The maximal feature of the Sload is around 17k.It can be seen from the comparisons that the algorithms' performance also depends on the size of the dataset and the applications used.

3. **M SVM –ROC Curve:** The true positive rate (TPR) is shown against false positive rate (FPR) using a visual called the ROC curve. The area under the ROC curve between the range (0 to 1) is taken into account while calculating the ROC value. The performance of the ML classifier increases with TPR.

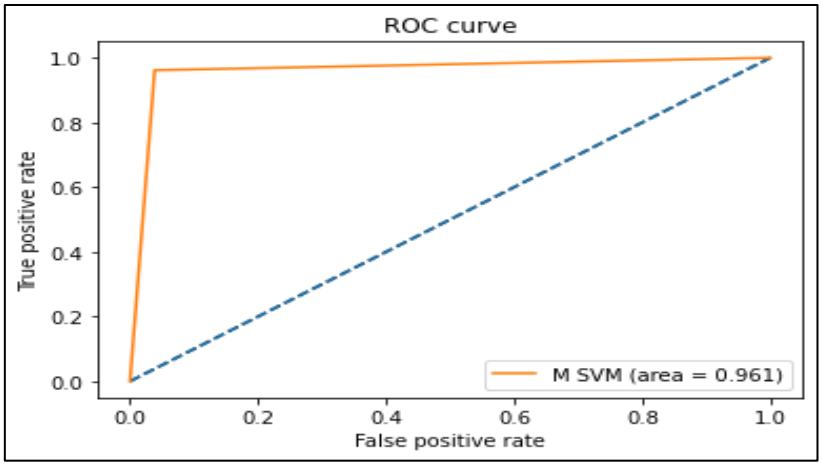

Figure 7: M SVM

The ROC is calculated in order to obtain the number that indicates how effective the model is in classifying in comparison to a one-hot encoding technique. The area is greater and the ROC value is higher the more top-left the curve is. Figure 5 displays the average ROC curve values while using one-hot encoding. It demonstrates that, when compared to other ML algorithms, the ID system ML method performs the best, with a ROC value of 0.961.

4. **Confusion Matrix:** To graphically show the algorithm's categorization performance, a tabular diagram with True label on the vertical axis and predicted label on the horizontal axis is provided.

Figures 8 show the evaluation parameters of normal (0) and abnormal (1) traffic in the datasets obtained from the confusion matrix which shows the applying of CM for evaluating the ML models in detecting insider dataset. The False Positive Set (1-1) displays the dataset with attacks, the set (0-1/1-0) displays partial attack, and the True Positive Set (0-0) has no attacks.

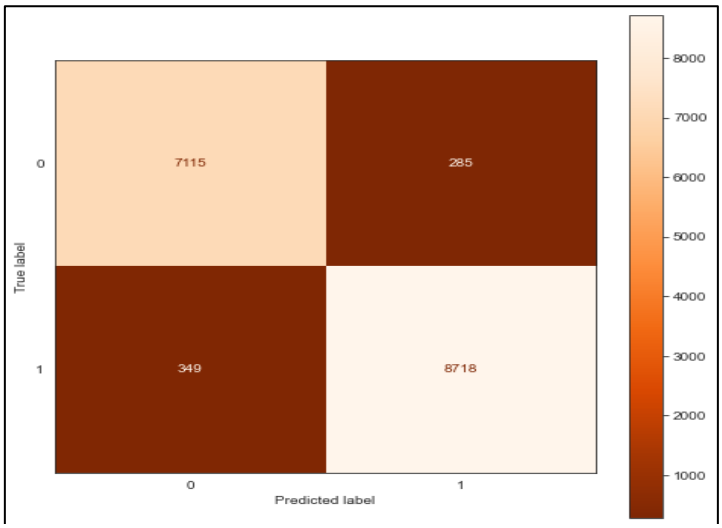

Figure 8: Confusion Matrix

V. CONCLUSION

The security and privacy of users have been severely compromised by the rising frequency of network and host machine breaches. A promising subject in artificial intelligence and cyber security is use of ML approaches in IDS. A first framework to effectively combine data sanitization, encoding, and cleaning. One-hot encoding was used for the data preprocessing in ML models. The suggested M-SVM classifier resulted in some improvements since it produced more accuracy and permitted attack prediction. IDSs are intended to detect attacks, thus it's important to pick the best data source based on those features.ID System evaluates the data that is processed in cyber security in order to classify IDSs. Even with only a limited number of training data, M-SVM was able to produce the highest overall classification accuracy. With enough training data, ML-based IDSs may reach respectable detection levels and have adequate generalizability to identify cyber security. The study shows that the strategy is effective in identifying an attack. The ID system ML algorithm performs the best with an ROC value of 0.961 compared to other ML algorithms.

REFERENCES

[1] Z. Zhang, H. A. Hamadi, E. Damiani, C. Y. Yeun and F. Taher, "Explainable Artificial Intelligence Applications in Cyber Security: State-of-the-Art in Research," in IEEE Access, vol. 10, pp. 93104-93139, 2022.

[2] I. A. Khan, D. Pi, M. Z. Abbas, U. Zia, Y. Hussain and H. Soliman, "Federated-SRUs: A Federated-Simple-Recurrent-Units-Based IDS for Accurate Detection of Cyber Attacks Against IoT-Augmented Industrial Control Systems," in IEEE Internet of Things Journal, vol. 10, no. 10, pp. 8467-8476, 15 May15, 2023.

[3] Asif, Muhammad, Sagheer Abbas, M. A. Khan, Areej Fatima, Muhammad Adnan Khan, and Sang-Woong Lee. "MapReduce based intelligent model for intrusion detection using machine learning technique." *Journal of King Saud University-Computer and Information Sciences* (2021).

[4] Otoum, Yazan, Dandan Liu, and Amiya Nayak. "DL-IDS: a deep learning–based intrusion detection framework for securing IoT." *Transactions on Emerging Telecommunications Technologies* 33, no. 3: e3803, 2022.

[5] Vinayakumar, Ravi, Mamoun Alazab, K. P. Soman, Prabaharan Poornachandran, Ameer Al-Nemrat, and Sitalakshmi Venkatraman. "Deep learning approach for intelligent intrusion detection system." *Ieee Access* 7 (2019): 41525-41550.

[6] Dahouda, Mwamba Kasongo, and Inwhee Joe. "A deep-learned embedding technique for categorical features encoding." *IEEE Access* 9 (2021): 114381-114391.

[7] H. Chen, J. Chen and J. Ding, "Data Evaluation and Enhancement for Quality Improvement of Machine Learning," in IEEE Transactions on Reliability, vol. 70, no. 2, pp. 831-847, June 2021.

[8] Ahmad, Iftikhar, Mohammad Basheri, Muhammad Javed Iqbal, and Aneel Rahim. "Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection." *IEEE access* 6: 33789-33795, 2018.

[9] M. Wang, K. Zheng, Y. Yang and X. Wang, "An Explainable Machine Learning Framework for Intrusion Detection Systems," in IEEE Access, vol. 8, pp. 73127-73141, 2020.

[10] Bertoli, Gustavo De Carvalho, Lourenço Alves Pereira Júnior, Osamu Saotome, Aldri L. Dos Santos, Filipe Alves Neto Verri, Cesar Augusto Cavalheiro Marcondes, Sidnei Barbieri, Moises S. Rodrigues, and José M. Parente De Oliveira. "An end-to-end framework for machine learning-based network intrusion detection system." *IEEE Access* 9: 106790-106805, 2021.

[11] Y. K. Saheed and M. O. Arowolo, "Efficient Cyber Attack Detection on the Internet of Medical Things-Smart Environment Based on Deep Recurrent Neural Network and Machine Learning Algorithms," in IEEE Access, vol. 9, pp. 161546-161554, 2021.

[12] X. Gao, C. Shan, C. Hu, Z. Niu and Z. Liu, "An Adaptive Ensemble Machine Learning Model for Intrusion Detection," in IEEE Access, vol. 7, pp. 82512-82521, 2019.

[13] Goudjil, Mohamed, Mouloud Koudil, Mouldi Bedda, and Noureddine Ghoggali. "A novel active learning method using SVM for text classification." *International Journal of Automation and Computing* 15 (2018): 290-298.

[14] M. A. Ferrag, L. Shu, O. Friha and X. Yang, "Cyber Security Intrusion Detection for Agriculture 4.0: Machine Learning-Based Solutions, Datasets, and Future Directions," in IEEE/CAA Journal of Automatica Sinica, vol. 9, no. 3, pp. 407-436, March 2022.

[15] Z. Qu, W. Chen, S. -Y. Wang, T. -M. Yi and L. Liu, "A Crack Detection Algorithm for Concrete Pavement Based on Attention Mechanism and Multi-Features Fusion," in IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 8, pp. 11710-11719, Aug. 2022, doi: 10.1109/TITS.2021.3106647.

