Hybrid Machine Learning Techniques For Detecting Intrusion Detection Systems And Analysing The Trust-Based WSN

Mrs. K. R. Prabha[1], Dr. B. Srinivasan[2]
[1]Dept of Computer Science
[2]Associate Professor, Dept of Computer Science
[1, 2] Gobi Arts & Science College, Gobichettipalayam, Tamil Nadu, India

Abstract- *Machine Learning can deliver real-time solutions that optimise network resource use, prolonging network lifetime. It can process autonomously without being programmed externally, making the process simpler, more efficient, less expensive, and more reliable. ML algorithms can process complex data more quickly and precisely. Machine Learning is being used to enhance the Wireless Sensor Network environment. Wireless Sensor Networks (WSN) comprise several decentralised and distributed networks by design. WSNs comprise sensor nodes and sink self-organizing nodes and self-healing. We proposed the HMIDS method to detect intrusion and analyze the trust-based WSN using hybrid machine learning algorithms. Since real datasets are inaccessible, most IoT intrusion detection research is predicated on the benchmarked KDD cup or simulated datasets. WSNs have grown dramatically in recent years due to electronics and wireless communication technology advancements. Yet, significant issues persist, such as low computational capability, limited memory, and limited energy supplies. To need source-based privacy measures, infrastructure must be physically susceptible. WSNs are used to monitor changing environments, and Machine Learning approaches are essential for sensor networks to adapt to this situation and avoid wasteful redesign. An analysis of machine learning methods for WSNs indicates a wide range of applications where security is highly valued. To secure data from hackers and other attackers, the WSN's system should be capable of erasing instructions if hackers or other attackers try to steal data.*

Keywords- Intrusion Detection System, Wireless Sensor Network, Hybrid Algorithm, K-means, SVM,

I. INTRODUCTION

WSNs comprise many sensor nodes capable of detecting real-world events, analysing data, and communicating with other nodes [1]. These resource-constrained networks' distinctive properties, such as deployment density, unattended operations in a harsh or hostile environment, multi-hop communication, decentralised administration, and self-healing and self-organizing network needs, provide a serious security challenge [2].

Wireless Sensor Networks (WSNs) are recognised as candidates for processing with low power consumption and contemptible qualities, which are used in various industries to collect data on human activities and behaviour, monitor diverse natural activities, and so on [3]. Security is a critical issue with WSNs. Active and Passive are the two major classifications. They are undetectable in passive attacks and tap the link above to store data; alternatively, they remove the internet's performance element [4-7]. Passive attack types include broken nodes, tampering, and traffic. The primary network assaults target the network itself in an active attack, and the source of attacks may be observable [8]. Because of these attacks, services may be disrupted or compromised for some time [9]. Jamming, hole attacks, Denial-of-Service (DoS), Sybil-type attacks, and flooding are all types of attacks. The network's activity is completed, whether passively or aggressively. "Intrusion Protection" If the instructions are not followed, "Intrusion Detection" will take place [10]. Intrusion Detection Systems (IDSs) offer information to other systems: intruder detection and location, intrusion instance, intrusion type, and intrusion location [11]. As more information about an incursion is discovered, it may help minimise and resolve the source of the attacks As a result, detecting intrusion systems benefits network security [12].

IDSs are classified into three types depending on their detection methods: anomaly detection, abuse detection, and specification-based detection [13]. Statistical models of the system's normal characteristics are utilised for anomaly identification. A certain amount of difference is seen as an attack. Although this strategy is excellent at detecting unanticipated threats, regularly updating standard profiles puts an additional load on resource-constrained WSNs [14]. Usage-based (signature or rule-based) detection consistently identifies previously defined/known threats. Detecting newly-emerging attacks is quite unlikely, even though it can discover

assaults that already exist in the signature database [15]. defines the difference between anomaly and misuse detections: "anomaly detection systems aim to detect the result of improper behaviour, while misuse detection systems attempt to identify known bad behaviour." Specification-based or hybrid intrusion detection systems combine the advantages of anomaly and abuse detection approaches [16-17].

The rest of this paper is organised as follows. Section 2 has several writers that cover a wide range of IDS detection topics. Section 3 depicts the HMIDS model. Section 4 describes the investigation's findings. Part 5 finishes with a discussion of the findings and future directions.

II. BACKGROUND STUDY

Alkasassbeh and Almseidin [1], Three categorization algorithms were employed to address the low accuracy often experienced by intrusion detection systems (IDS) that combine artificial neural networks with fuzzy clustering to identify rare attacks. They improved the accuracy and lowered the complexity of each training set by splitting the diverse collection of training data into homogenous subsets of training data. J48 trees, Multilayer Perceptron (MLP), and Bayes network approaches were used in the proposed research, with J48 trees providing the maximum accuracy. A significant weakness in their work was their inability to employ feature selection to delete any disconnected, redundant, and undesirable characteristics.

A real-time hybrid intrusion detection approach was proposed by Dutt I. et al. [5]. The abuse technique was used to detect well-known attacks, while the anomaly strategy was used to detect new attacks. The ability of the anomaly detection technique to identify assault invasion patterns that avoided abuse detection resulted in a high detection rate in this investigation. The model's accuracy achieved a significant number of 92.65% on the last day of the trial. As the model learns and trains the system daily, the proportion of false negatives decreases drastically. When the model was applied to massive data sets, the issue of a slow detection rate persists. Kazi Abu Taher et al. [8] Some prior initiatives could not apply feature selection to their datasets to exclude irrelevant, undesired, or redundant properties. Many ML models and techniques were investigated using the NSL-KDD dataset, and feature selection was accomplished using the wrapper methodology. The accuracy was comparable compared to previous research utilising the same dataset. Because of the model's high false positive rate and the work's sole focus on signature-based attacks, new attacks go undetected, a fundamental problem of zero-day detection that has yet to be fixed.

Marzia Z. and Chung-Horng L. [10] The results of many supervised and unsupervised machine learning methods were integrated using a voting classifier. The research increases the accuracy and performance of current intrusion detection systems. They chose the Kyoto2006+ dataset, which seems more promising than the KDDCup '99 dataset, which appears to be the most employable. This allows them to reach a certain level of accuracy. However, the findings' recall was relatively poor in a few cases, suggesting a substantial false-negative rate (FPR)

Verma et al. [15] indicate that anomaly-based intrusion detection may be enhanced, especially regarding false positives. Extreme gradient boosting (XGBoost) and Adaptive boosting (AdaBoost) learning techniques were used on the NSL-KDD dataset. While the accuracy was 84.253, the performance should be improved by employing hybrid or ensemble machine learning classifiers.

Vinoth Y. K and Kamatchi K. [16] This work adds to managing imbalanced data by identifying the most practical features to be taught to detect intrusion and alert system administrators to whether the intrusion was normal or abnormal. Even though the models perform well on NSL-KDD, an experiment on the most current datasets was required.

Zhou et al. [17] A novel intrusion detection system was presented that blends ensemble classification with feature selection, resulting in increased efficiency and high-precision intrusion detection. The study used three separate datasets, including the well-known NSL-KDD dataset and two freshly published datasets, CIC-IDS2017 and AWID. A CFS-BA-based technique was employed for feature selection. The ensemble-based strategy enhances multiclass classification performance on unbalanced data sets. In the AWID dataset, the model had the highest accuracy, with 99.90% accuracy.

III. MATERIALS AND METHODS

3.1 Feature Selection

To improve the efficiency of data mining algorithms and build more effective hybrid intrusion detection models. In that case, the first step you need to do is to carefully pick the characteristics that will be included in each model. Classifiers often receive data in a high-dimensional feature space, but only certain features help determine which classes they belong to. Some information is unhelpful because it is irrelevant, redundant, or loud. The presence of irrelevant and redundant traits might hamper the learning process. Reducing the number of characteristics, eliminating irrelevant, noisy, or redundant

features, and the effects on applications such as speeding up a data mining algorithm, improving learning accuracy, and producing a more intelligible model are all positive outcomes. The next step in developing an efficient detection system is to extract the set of characteristics or features that are the most effective so far.

The goal of feature selection in network intrusion detection is to maximise the detection rate while minimising the false alarm rate. WEKA 3.7, a machine learning tool, was used to compute the feature selection subsets for the SVM classifier to assess classification performance on each feature set.

3.2 Dataset Normalization

Dataset normalisation is a kind of preprocessing that is helpful in categorization. The learning process may be sped up, and the efficiency of an intrusion detection system can be boosted by normalising the input data when the dataset is too large. Linear changes A dataset X may be transformed into the specified interval using Min-Max Normalization. This strategy proportionately changes the data from (Xmin, Xmax) to (Newmin, Newmax) (Newmin, Newmax). A plus of this method is that it preserves all relationships between data values. The minimum and maximum values in (1) are as follows:

$$X_{new} = \frac{(X - X_{min})}{(X_{max} - X_{min})} \text{ ----(1)}$$

3.3 K-Means

K-Means is an unsupervised technique for fixing the intrusion detection system's training datasets problem. When a sample has been randomly divided across K clusters, the centres of those clusters are found, and the sample is redistributed to the clusters that contain their nearest centres. The process continues until the clusters' centres no longer shift noticeably. Scores are based on how far off individual elements are from the cluster centres, which is always the case when assigning elements to clusters. To organise a set of n x j coords where j = (1,...,n), we need to divide them up into a set of c x I groups where I = (1,...,n) (1,...,c). The function is defined in terms of the Euclidean distance between a group J vector Xj and the centre of its corresponding cluster Ci, as shown in the following Equation (2)

$$j = \sum_{i=0}^{c} ji \sum_{i=1}^{c} [\sum_{k,x \in Gi}^{c} \left| |XK - Ci| \right|^2] \text{ ------- (2)}$$

Using the K-means clustering technique, we determined and created two clusters for each output class. Algorithmically, each cluster's structure is updated as it iteratively processes the training data. During a cluster update, elements are removed from one group and added to another. The centred values shift when clusters are refreshed. This update is in line with what we currently have in our clusters. The K-Means algorithm's clustering is complete when no changes are made to any cluster.

3.4 Support Vector Machine

An approach of an organisation that relies on categories An intrusion detection system (IDS) will classify every network activity as either benign or malicious. DDoS, U2R, R2L, and Probing are the four main categories of network attacks. The results of the above processing are divided into two categories, "normal" and "assault," and then specific attack types are allocated to each category. Support vector machines (SVMs) use a set of training inputs called support vectors to classify data by outlining a hyperplane in the feature space. SVMs provide a generic approach to fitting the hyperplane surface to the data using a kernel function. While doing traditional supervised learning, we are provided with n training samples (xi, yi), I = 1, 2,..., n, where xi X represents the input vector and yi Y, yi +1, -1 represents the output vector.

Accuracy (A) was used as the evaluation metric (3). The sum of all correctly categorised connections, both non-invasive and invasive, is illustrated in Equation (3). Based on Equation (4), Detection Rate (DR) is the fraction of assaults correctly detected as attacks relative to the attack sample size and false alarm rate. Eq. (5) shows that the False Alarm Rate (FAR) is the percentage of false positives expressed as a fraction of the total number of standard samples.

$$A = (TP + TN) / (TP + TN + FP + FN) \text{ ------- (3)}$$

$$DR = (TP)/(TP + FP) \text{ -------- (4)}$$

$$FAR = (FP)/(FP + TN) \text{ ------- (5)}$$

IV. RESULTS AND DISCUSSION

False negative (FN): This represents the no. of discovered normal traffic flow; however, it is undoubtedly abnormal.

Here, the following factors, like False Positive Rate (FPR), Detection Rate (DTR) and Accuracy (ACC), are used as assessment measurements.

$$FPR = \frac{False\ positive}{False\ positive + true\ negative} * 100$$

$$DTR = \frac{True\ positive}{True\ positive + False\ negative} * 100$$

$$ACC = \frac{True\ positive + true\ negative}{TP + TN + FP + FN} * 100$$

TABLE I.COMPARISON OF EVALUATION METRICS FOR (SMO, KMEAN, SVM,K-MEAN+SVM)

Algorithms	Accuracy	DTR	FPR
K-mean	71.45	52.73	3.27
SVM	74.45	61.26	9.7
SVM+K-mean	98.34	95.38	1.23

Figure 1 and Table I compare the SVM, Kmean, and Hybrid (SVM + Kmean) algorithms with 37 chosen features using the three evaluation metrics described before. Based on the findings, it is evident that the Hybrid (SVM + K-mean) algorithm classifies normal and aberrant WSN traffic with high DTR and low FPR.

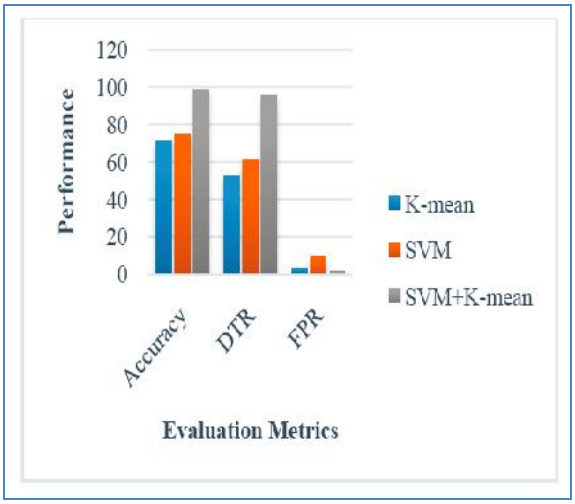

Figure 1 Comparison of evaluation metrics for (SMO, Kmean, SVM, Kmean+SVM)

V. CONCLUSION

In an overview of current research to provide knowledge of multi-operator architecture for intrusion detection, This research explored the necessity to build a mechanism to develop a new benchmarking data set for WSN that includes updated attack and ordinary network traffic that is not included in the KDD CUP data set. The proposed structure offers the whole PDML file, inferring the entire PCAP to content, in a manner suited for Weka's built-in loading. The created WSN dataset for IDS is analysed using a hybrid machine learning (K-mean+SVM) technique, yielding a 95% attack detection rate. The dataset will be updated in the future to help research the development of future IPv6 network attack detection solutions, and further Machine Learning and Deep Learning algorithms will be implemented to increase IDS accuracy.

REFERENCES

[1] Alkasassbeh and Almseidin. (2018). Machine Learning Methods for Network Intrusions. International Confrernce on Computing, Communication (ICCCNT). Arxiv

[2] Bhavani T. T, Kameswara M. R and Manohar A. R. (2020). Network Intrusion Detection System using Random Forest and Decision Tree Machine Learning Techniques. International Conference on Sustainable Technologies for Computational Intelligence (ICSTCI). (pp. 637-643). Springer.

[3] Bolon –C.V. (2012) Feature Selection and Classification in Multiple class datasets-An application to KDD Cup 99 dataset. https://doi.org/10.1016/j.eswa.2010.11.028

[4] Deyban P. Miguel A. A, David P. A, and Eugenio S. (2017). Intrusion detection in computer networks using hybrid machine learning techniques. XLIII Latin American Computer Conference (CLEI). (pp. 1-10). IEEE

[5] Dutt I. et al. (2018). Real Time Hybrid Intrusion Detection System. International Conference on Communication, Devices and Networking (ICCDN). (pp. 885-894). Springer.

[6] Farah N. H et al. (2015). Application of Machine Learning Approaches in Intrusions Detection Systems. International Journal of Advanced Research in Artificial Intelligence. IJARAI. (9-18).

[7] Iqbal and Aftab. (2019). A Feed-Forward ANN and Pattern Recognition ANN Model for Network Intrusion Detection. International Journal of Computer Network and Information Security, 4. Researchgate (19-25)

[8] Kazi A., Billal M. and Mahbubur R. (2019). Network Intrusion Detection using Supervised Machine Learning Technique with feature selection. International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). (pp. 643-646). IEEE.

[9] Maniriho et al. (2020). Detecting Intrusions in Computer Network Traffic with Machine Learning Approaches. International Journal of Intelligent Engineering and Systems. INASS. (433-445)

[10] Marzia Z. and Chung-Horng L.(2018). Evaluation of Machine Learning Techniques for Network Intrusion Detection. IEEE. (pp. 1- 5)

[11] N. F. Haq et al. (2015). An Ensemble framework for anomaly detection using hybridized feature selection approach (HFSA). Intelligent System Conference. (pp. 989-995). IEEE.

[12] Ponthapalli R. et al. (2020). Implementation of Machine Learning Algorithms for Detection of Network Intrusion. International Journal of Computer Science Trends and Technology (IJCST). (163-169).

[13] Rajagopal S., Poornima P. K. and Katiganere S. H. (2020). A Stacking Ensemble for Network Intrusion Detection using Heterogeneous Datasets. Journal of Security and Communication Networks. Hindawi. (1-9).

[14] S. Thapa and A.D Mailewa (2020). The Role of Intrusion Detection/Prevension Systems in Modern Computer Networks: A Review. Conference: Midwest Instruction and Computing Symposium (MICS). Wisconsin, USA. Volume: 53. (pp. 1-14).

[15] Verma P, Shadab K, Shayan A. and Sunil B. (2018). Network Intrusion Detection using Clustering and Gradient Boosting. International Conference on Computing, Communication and Networking Technologies (ICCCNT). (pp. 1-7). IEEE.

[16] Vinoth Y. and Kamatchi K. (2020). Anomaly Based Network Intrusion Detection using Ensemble Machine Learning Technique. International Journal of Research in Engineering, Science and Management. IJRESM. (290-296).

[17] Yuyang Z., Guang C., Shanqing J. and Mian D. (2019). Building an Efficient Intrusion Detection System Based on Feature Selection and Ensemble Classifier. Computer Networks. Doi: https://doi.org/10.1016/j.comnet.2020.107247