IJIRIS: International Journal of Innovative Research in Information Security
Volume 10, Issue 04, May 2024
E-ISSN: 2349-7017
P-ISSN: 2349-7009
https://www.ijiris.com/archives

Managing and Evaluating Data Integrity in machine Learning-Based Cyber Intrusion Detection

Shammi L
Assistant Professor/ Dept of CSE
East Point College of Engineering and Technology,
Bengaluru, Karnataka, INDIA
shammi.l@eastpoint.ac.in
Pandey Saurabhraj Krishnakant
Department of Computer Science Engineering
East Point College of Engineering and Technology,
Bengaluru, Karnataka, INDIA
1ep20cs063.cse@eastpoint.ac.in
Varsha YS
Department of Computer Science Engineering
East Point College of Engineering and Technology,
Bengaluru, Karnataka, INDIA
1ep20cs105.cse@eastpoint.ac.in
Deepanshu
Department of Computer Science Engineering
East Point College of Engineering and Technology,
Bengaluru, Karnataka, INDIA
1ep20cs022.cse@eastpoint.ac.in
Raunak Kumar
Department of Computer Science Engineering
East Point College of Engineering and Technology,
Bengaluru, Karnataka, INDIA
1ep20cs07.cse@eastpoint.ac.in

Publication History
Manuscript Reference No: IJIRIS/RS/Vol.10/Issue04/MYIS10094
Research Article | Open Access | Double-Blind Peer-Reviewed | Article ID: IJIRIS/RS/Vol.10/Issue04/MYIS10094
Received: 07, April 2024|Revised:16, April 2024 | Accepted: 27, April 2024 Published Online: 07, May 2024 Volume 2024
Article ID MYIS10094 http://www.ijiris.com/volumes/Vol10/iss-04/15.MYIS10094.pdf
Article Citation: Shammi,Pandey,Varsha,Deepanshu,Raunak(2024). Managing and Evaluating Data Integrity in machine Learning-Based Cyber Intrusion Detection. International Journal of Innovative Research in Information Security, Volume 10, Issue 03, Pages 203-209
doi:> https://doi.org/10.26562/ijiris.2024.v1004.15
BibTex key: Shammi@2024Managing

Copyright: ©2024 This is an open access article distributed under the terms of the Creative Commons Attribution License; which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract: This study investigated the impact of data quality on the performance of intrusion detection systems. Experiments were conducted using intrusion datasets and machine learning models. The pre-trained models were less affected by data duplications and overlaps compared to classic ML models. Removing overlaps and duplicates from training data improved the pre-trained models' performance in most cases, but had adverse effects in datasets with highly similar sequences. The study also proposed a framework for model selection and data quality assurance for building high- quality intrusion detection systems. Additionally, we focus on optimizing nine hyperparameters within a 1D-CNN model, using two well-established evolutionary computation methods—genetic algorithm(GA) and particle swarm optimization(PSO).The performances of these methods are assessed using three majordatasets—UNSW-NB15,CIC-IDS2017, and NSL-KDD. The key performance metrics considered in this study include the accuracy, loss, precision, recall, and F1-score. The results demonstrate considerable improvements in all metrics across all datasets, for both GA- and PSO-optimized models, when compared to those of the original non optimized 1D-CNN model.

Keywords: Intrusion detection, Data preprocessing, Reinforcement Learning, Support Vector Machines (SVMs), artificial neural networks (ANNs)

I. INTRODUCTION

Machine learning is becoming a powerful tool in data curation, the process of keeping your data clean and organized. By applying machine learning algorithms, data can be automatically identified, filtered for inconsistencies, and even labeled with relevant information. This helps ensure the data used to train machine learning models itself is accurate and unbiased, leading to better overall model performance and generalizability.

Managing and evaluating data integrity in machine learning-based cyber intrusion detection is a critical process aimed at ensuring the trustworthiness, accuracy, and security of the data utilized in identifying and mitigating cyber threats. As organizations increasingly rely on machine learning algorithms to detect and prevent intrusions, it becomes imperative to maintain the integrity of the data feeding into these systems. In managing data integrity, several pivotal tasks are involved. Data preprocessing stands as an initial step, encompassing activities such as cleansing and restructuring raw data to conform to a format compatible with machine learning models. This preparation enhances the quality and consistency of the data, thus optimizing the performance of subsequent analyses. Additionally, feature selection plays a crucial role in focusing on pertinent attributes within the data that contribute significantly to the detection of intrusions. By identifying and prioritizing these features, the model's efficacy in discerning malicious activities is enhanced. Moreover, addressing dataset imbalances is imperative to rectify disparities in the distribution of data samples across different classes. This ensures that the model does not exhibit biases towards any particular class, thus preserving its objectivity and effectiveness. Furthermore, implementing robust security measures is paramount to safeguarding data integrity against tampering and unauthorized access. Encryption, access controls, and data monitoring mechanisms are among the strategies employed to fortify the security of the data utilized in intrusion detection systems. These measures serve to mitigate the risks associated with data manipulation and unauthorized intrusion attempts, there by bolstering the overall integrity of the system.

In the realm of evaluating data integrity, a comprehensive assessment of machine learning models' performance is conducted to ascertain their reliability and effectiveness in detecting cyber intrusions. This entails the utilization of various performance metrics, including accuracy, precision, recall, and F1-score, to gauge the model's ability to accurately classify and identify potential threats. Validation and test datasets are employed to validate the model's performance and generalization capabilities, ensuring its robustness across diverse scenarios. Techniques such as cross-validation further enhance the model's reliability by validating its performance across multiple iterations and subsets of the data. Managing data integrity involves various tasks, including preprocessing, feature selection, and dataset balancing. Preprocessing involves cleaning and transforming raw data to prepare it for analysis, ensuring consistency and reliability. Feature selection focuses on identifying the most relevant attributes that contribute to detecting intrusions, while dataset balancing addresses any imbalances in the distribution of data samples across different classes, preventing biases in the model.

II. LITERATURE REVIEW

Using fitness evaluations as guidance, the PSO algorithm effectively reached the optimal set of hyperparameters. Consequently, the intrusion-detection model was improved, showcasing customized effectiveness for smart farming applications. When studying hyper parameter optimization for nods, it is important to take into account the many methodologies and optimization techniques that have been previously suggested. This extensive research, which includes a range of strategies and models, is crucial for establishing a solid basis to enhance our comprehension and direct our investigation towards creating more productive and successful methods in the field of 1D-cnns.

Hyperparameters directly control the performance of deep learning algorithms and have a major impact on the model's ability to discover patterns [5]. Hence, the choice of hyperparameters is crucial for the successful implementation of deep learning projects. Improper choice of hyperparameters can result in insufficient learning outcomes. For example, if the learning rate is too high, it can result in the model overlooking significant patterns in the data or failing to reach a stable solution. On the other hand, a low learning rate might lead to the model disregarding important information, resulting in a slower learning process and a decline in overall performance.

The PROA is an exceptional approach that relies on the observed patterns of wealth distribution in society. It primarily focuses on two distinct categories of individuals—affluent individuals, whose wealth exceeds the average wealth, and less affluent or impoverished individuals, whose wealth is below the average wealth. Within the algorithm's framework, the affluent class consistently endeavors to expand the disparity between themselves and the less affluent class. In the PROA, solutions in the less affluent population move towards the global optimum solutions in the search area by acquiring knowledge from the more affluent solutions determined in the wealthier population. This methodology introduced a novel technique for optimizing hyperparameters, emphasizing the similarities between social wealth behaviors and solution search tactics. Therefore, by implementing the PROA, the authors asserted that they had notably enhanced the detection rate of their CNN-ALSTM model.

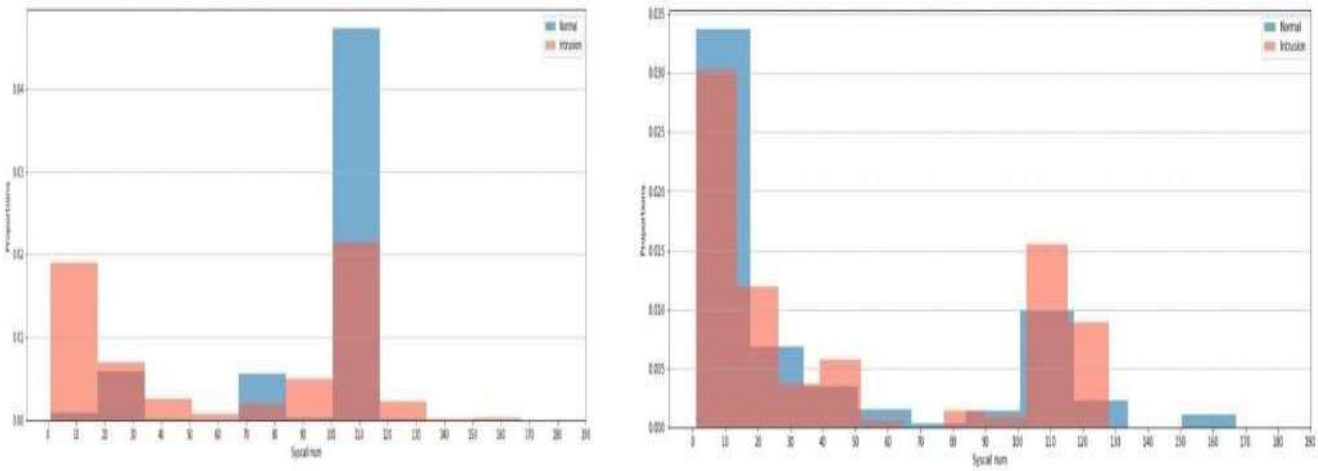
In their work, authors in reference [11] employed the tree-structured Parzen estimator (TPE) technique to perform Bayesian hyperparameter optimization, introducing a novel approach. This procedure consisted of multiple sequential steps. At first, the search space for hyperparameters was determined by specifying the hyperparameters to be optimized and their possible ranges of values. Afterwards, the experiment chose the categorical cross-entropy loss function. Subsequently, the TPE algorithm was implemented, and the potential hyperparameters were methodically sampled throughout the search space until the predetermined criteria were met. Ultimately, the most effective model parameters were established. The Bayesian optimization strategy provided a methodical and well-founded approach to tweaking hyperparameters, effectively balancing the exploration and exploitation aspects within the search space. The key hyperparameters of all base CNN models in the proposed framework were optimized using Particle Swarm Optimization (PSO) in [8].

The hyperparameters optimized in this study included the proportion of frozen layers, the learning rate, the dropout rate, the batch size, the number of epochs, and the early-stop patience. Hyperparameters have a direct impact on the structure, performance, and efficiency of CNN models. However, as CNN models with default hyperparameter settings already achieved almost perfect accuracy on the Car-Hacking dataset, the hyperparameter tuning process was only carried out for the CICIDS2017dataset. With the intention of forging a new direction in this domain, this study endeavors to establish innovative approaches. Given the ability of Convolutional Neural Networks (CNN's) to effectively handle intricate datasets and the significance of choosing suitable hyperparameters to get the best model performance, our focus is on optimizing hyperparameters in 1D-CNN models. Our objective is to utilize the advantages of current approaches while addressing their limitations, in order to advance hyperparameter optimization and ultimately enhance the effectiveness and efficiency of IDSS.

III. DATASET DESCRIPTION

In order to compare the difference between the two classes, we graph an overlaid histogram of normal data and intrusion data from each dataset in Figure 2 We use all of the UNM (except Send mail), MIT Live Lpr, and ADFA-LD datasets. Each dataset has two histogram versions. However, we only include the histograms of the processed and unprocessed UNM Synthetic Send mail dataset. The histograms of the rest of the datasets along with their descriptive analyses are included in our got hub repository. The first histogram version displays the dispersion of original traces that have not been cleaned nor processed yet, so it can show the actual difference between normal and intrusion sequences. After cleaning duplicated sequences that exist in both classes, the overlaid histograms of most datasets have changed. Significantly. Therefore, the second version exhibits a more distinguished difference between the two classes than the original data. The goal is to differentiate normal sequences from intrusion sequences as much as possible so that the algorithms can learn to distinguish them from one another. Therefore, the more distinctive the two classes are, the better the candidate algorithms can perform.

Figure 2a shows that there is a slight difference between normal data and intrusion data from the Synthetic Send mail dataset. However, most of them overlap each other. After cleaning duplicated sequences, the dispersion of normal system calls and intrusion system calls have changed in Figure 25 Normal system calls have expanded from two wide ranges (68 to 83 and 101 to118) to multiple specific ranges. Particularly, sequences that contain system call numbers ranging between 13 and 15, 64 and 77,100 and 102, 128 and 134, and 150 and 177 are most likely normal. This increases the homogeneity of normal data and reduces false negatives. Additionally, after being processed, intrusion system calls have expanded to more specific ranges. For example, sequences that contain system call numbers from 17to 27, 77 to 84, and 102 to 128 are most likely intrusion. This narrows down possible intrusion likelihoods and reduces falsepositives in overlapped system calls.



(a) The Original UNM Synthetic Send mail Dataset (b)The Processed UNM Synthetic Send mail Datasets

Fig 1:- Overlaid Histograms of the Original and the Processed System Calls in Normal or Intrusion Data of the UNM Synthetic Send mail Dataset.

IV. WORKFLOW

The workflow of data curation and quality evaluation involves a systematic process aimed at ensuring the reliability, accuracy, and usability of the dataset for analysis and decision-making purposes. Initially, the dataset is collected from various sources and undergoes thorough documentation to provide insights into its contents, format, and provenance. Data preprocessing techniques are then applied to clean and transform the raw data, addressing issues such as missing values, outliers, and inconsistencies. Next, feature engineering methods are employed to extract relevant information and enhance the dataset's representativeness and predictive power. Subsequently, the dataset is subjected to comprehensive quality evaluation across multiple dimensions, including completeness, accuracy, consistency, timeliness, and relevance. Quality evaluation metrics are calculated for individual features as well as for the dataset as a whole, providing insights into the overall integrity and reliability of the data.

IJIRIS © 2014-24, AM Publications - All Rights Reserved

Finally, the results of the quality evaluation are documented and communicated to stakeholders, along with recommendations for improving data quality and usability. Overall, the workflow of data curation and quality evaluation ensures that the dataset is well-prepared and reliable for analysis, enabling researchers and practitioners to derive meaningful insights and make informed decisions based on high quality data.

V. METHODOLOGY

To assess the user's experience including their perspective regarding data sharing, their attitudes toward data curation, and the level of satisfaction with the data curation service provided at each of the authors' institutions, the survey instrument was drafted consisting of multiple choice and open-ended questions. The results of this survey were intended to evaluate and improve data repository service; the IRB determined this research does not involve human subjects as defined by DHHS and FDA regulations. The survey was implemented in Qualtrics and distributed to data depositors at six Data Curation Network institutions, including: Cornell University's commons; Duke University's Research Data Repository; Johns Hopkins University's Data Archive; University of Illinois Urbana-Champaign's Illinois Data Bank, University of Michigan's Deep Blue Data, and the University of Minnesota's Data Repository for UofM (DRUM).

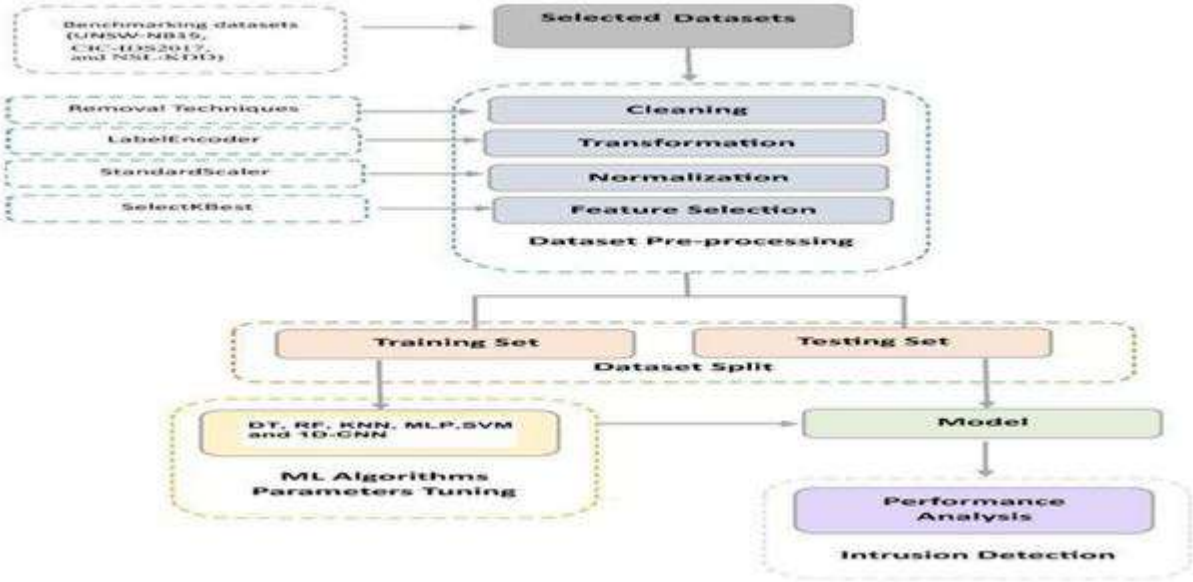
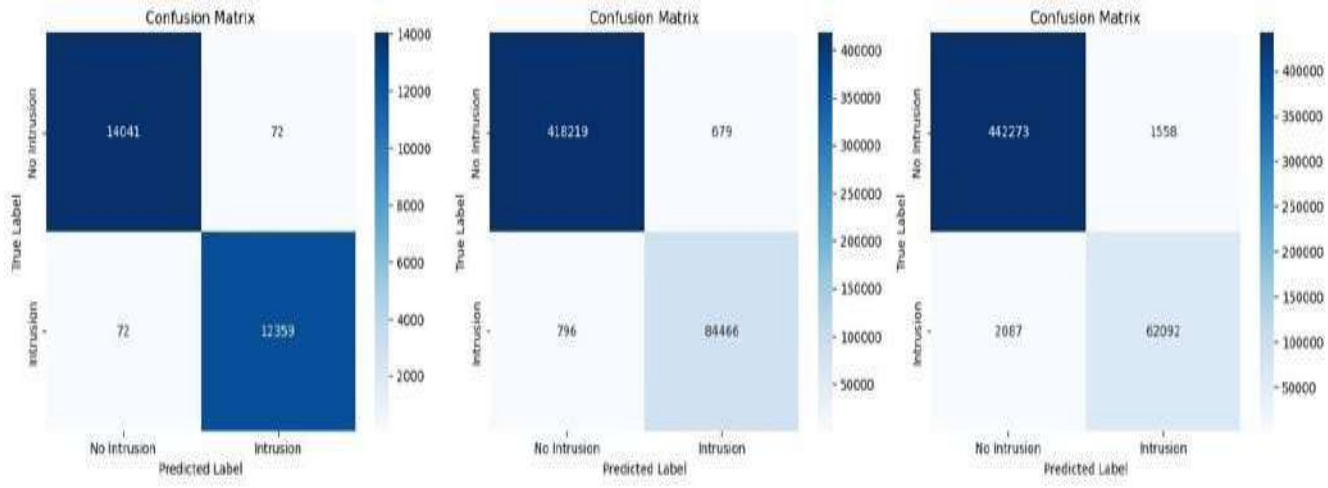


Fig2:- Data Preparation work flow for ML based IDS

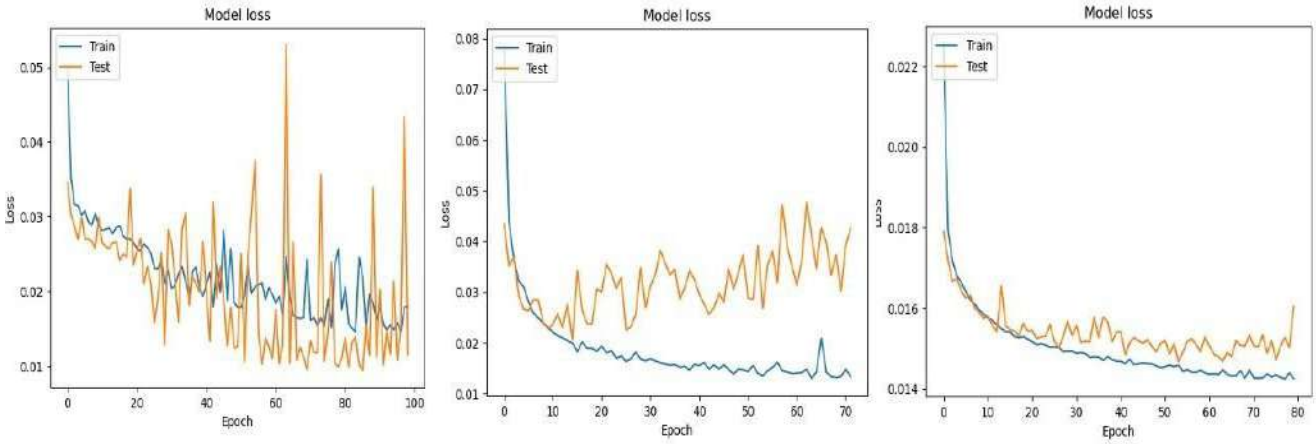
The study also proposed a framework for model selection and data quality assurance for building high- quality intrusion detection systems. Additionally, we focus on optimizing nine hyperparameters within a 1D-CNN model, using two well-established evolutionary computation methods—genetic algorithm (GA) and particle swarm optimization(PSO).The performances of these methods are assessed using any major datasets—UNSW-NB15, CIC- IDS2017, and NSL-KDD. The key performance metrics considered in this study include the accuracy, loss, precision, recall, and F1-score. The results demonstrate considerable improvements in all metrics across all datasets, for both GA- and PSO-optimized models, when compared to those of the original non optimized1D-CNN model.

VI. RESULTS



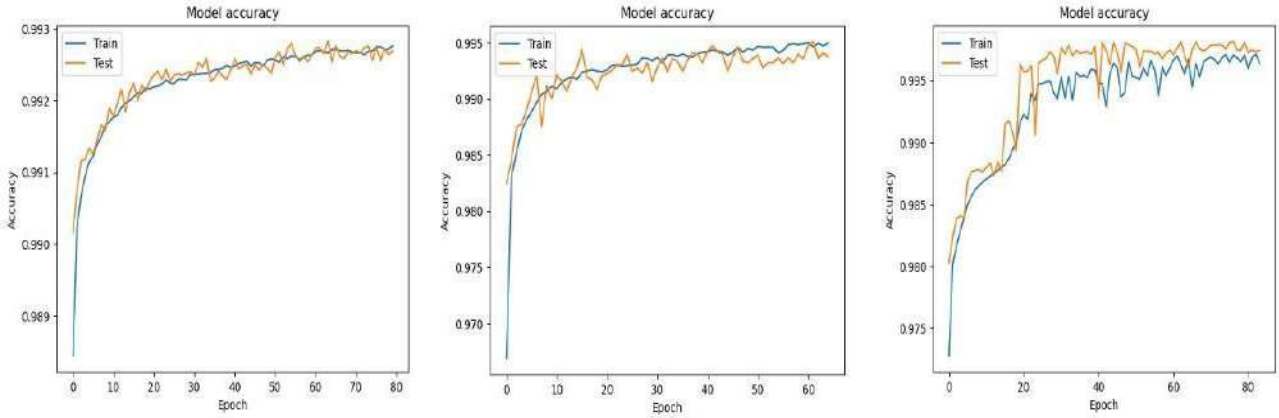
(a) NSL-KDD **(b)CIC–IDS2007** **(c)UNSW-NB15**

Fig2.1:-Confusion matrix using GA



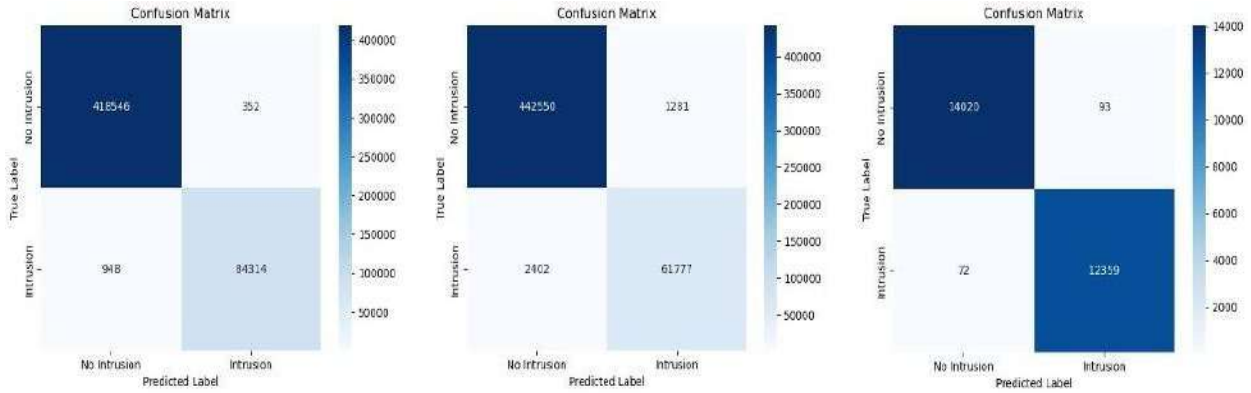
(a) CIC –IDS2007 **(b) NSL-KDD** **(c) UNSW-NB15**

Fig 2.2:-Model loss using GA



(a) UNSW-NB15 **(b) NSL-KDD** **(c) CIC – IDS2017**

Fig2.3:-Model accuracy using PSO



(a) CIC–IDS2017 **(b) UNSW-NB15** **(c) NSL – KDD**

Fig 2.4:-Confusion matrix using PSO

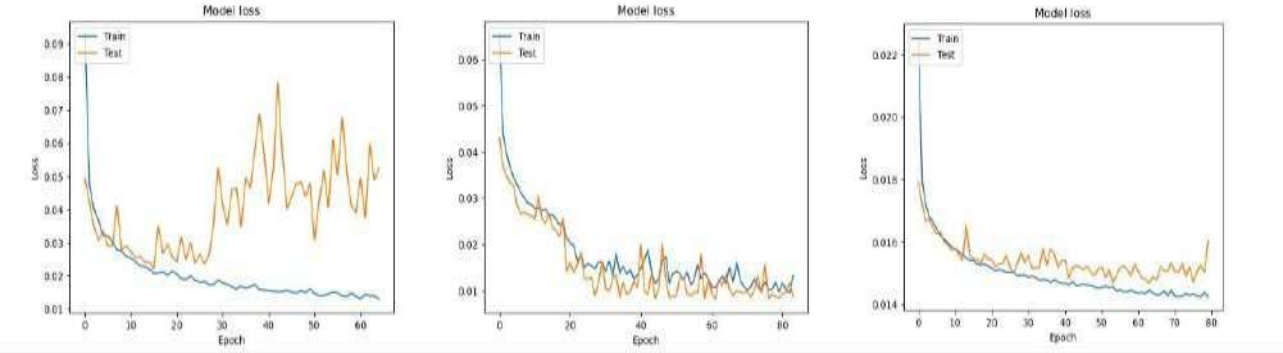
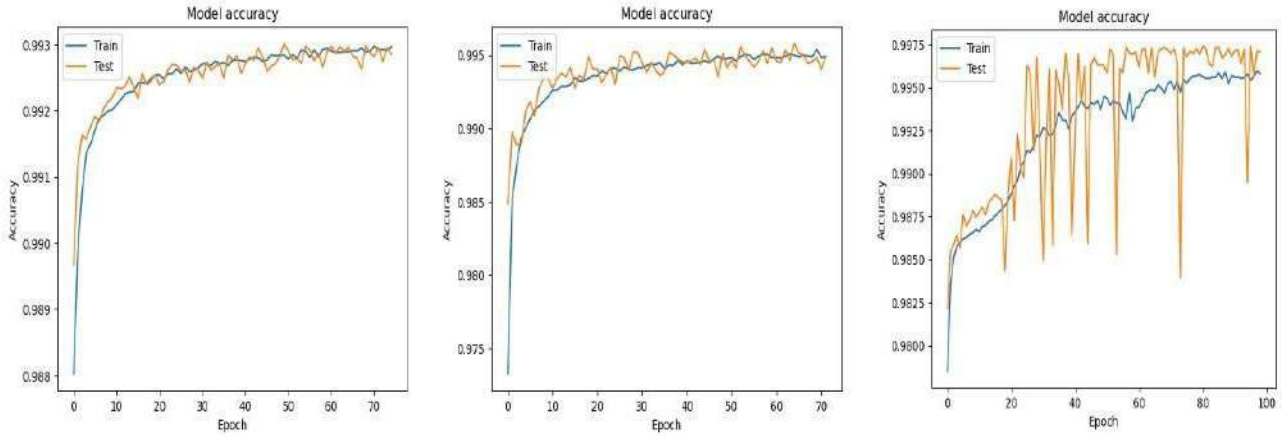


Fig 2.5:-Model loss using PSO



(a) UNSW-NB15 **(b) NSL–KDD** **(c) CIC –IDS2007**

Fig2.6:-Model accuracy using GA

VI.CONCLUSION

In conclusion, the field of "Managing Data and Quality Evaluation for Machine Learning-Based Cyber Intrusion Detection" plays a vital role in enhancing the efficacy and reliability of cyber security measures. Cyber threats continue to evolve and pose a significant risk to organizations, making the development of accurate and efficient intrusion detection systems crucial. In a world where cyber threats are ever-present, the work on data curation and quality evaluation for machine learning-based cyber intrusion detection is a proactive and necessary step to protect sensitive data, critical infrastructure, and digital assets. Item powers organizations to build robust and adaptive intrusion detection systems that can effectively safeguard against a wide range of cyber threats and vulnerabilities. As the field of cybersecurity continues to evolve, ongoing research and development in data curation and quality evaluation will remain essential for staying one step ahead of cyber adversaries.

REFERENCES

[1]. KaziAbuTaher,Billal Mohammed Yasin Jisan, Md.Mahbubur Rahman,"Network Intrusion Detection using Supervised Machine Learning Technique with Feature Selection",2019; IEEE
[2]. AKM Mashuqur Rahman Mazumder, Niton Mohammed Kamruzzaman, NasrinAkter,Nafija Arbe and MdMahbubur Rahman," Network Intrusion Detection Using Hybrid Machine Learning Model",2021,IEEE
[3]. J.Olamantanmi Mebawondu, Olufunso D.Alowolodu, Jacob O.Mebawondu, Adebayo O.Adetunmbi," Network intrusion detection system using supervised learning paradigm",2020,Elservier.
[4]. RuchiTuli,"PacketSniffingandSniffingDetection",2020,IJIET.
[5]. Camila Pontes, Manuela Souza, João Gondim, Matt Bishop and Marcelo Marotta, "A new method for flow-based network intrusion detection using the inverse Potts model", 2021,IEEE
[6]. Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo."every one wants to do the model work, not the data work": Data cascades in high-stakes ai. In Proceedings of the2021 CHI Conference on Human Factors in Computing Systems,CHI'21, NewYork, NY, USA,2021.
[7]. JoyBuolamwini and Timnit Gebru.Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency, pages 77–91,2018.
[8]. Haihua Chen, Jiangping Chen, and Junhua Ding. Data evaluation and enhancement for quality improvement of machine learning. IEEE Transactions on Reliability, pages1–17, 2021.
[9]. AbhijeetSahu, ZeyuMao, KatherineDavis, and AnaEGoulart. Data processing and model selection for machine learning based network intrusion detection. In 2020 IEEE International Workshop Technical Committee on Communications Quality and Reliability (CQR),pages1–6. IEEE,2020
[10]. JiankunHu, XinghuoYu, DongQiu, and Hsiao-HwaChen. A simple and efficient hidden markov model scheme for host-based anomaly intrusion detection. IEEE network,23(1):42–47,20