An ML-Based Intrusion Detection System Design and Evaluation for Enhanced Cybersecurity

Ansh Kataria
Centre for Interdisciplinary Research in Business and Technology, Chitkara University
Institute of Engineering and Technology, Chitkara University, Punjab, India
ansh.kataria.orp@chitkara.edu.in

2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI) | 979-8-3503-6996-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCSAI59793.2023.10421690

Abstract: In the ever-evolving realm of cyber space, ensuring security and integrity of information systems is paramount, given the dynamic nature of contemporary cyber threats. Traditional intrusion detection system, sometimes struggle to keep with these attacks. To address this challenge, the study has developed an innovative machine learning based intrusion detection system to enhance cyber security. The system can rapidly and accurately identify both known and novel risk by leveraging cutting edge machine learning techniques. We trained and validated our model using an extensive dataset encompassing various network scenarios. In comparison to conventional IDS, the ML IDS demonstrated superior detection, accuracy and reduced incidence of false positives. Additionally, the MLIDS provides to be a reliable solution for diverse network. topology is due to its adaptive learning capabilities, making it resilient against evolving cyber threats. this encompasses the design, design, implementation, and evaluation of the MLIDS, highlighting its potential as a valuable tool in next generation, cyber security solutions.

Keywords— ML, Intrusion Detection System, Cybersecurity, Anomaly Detection, Adaptive Learning.

I. INTRODUCTION

In the digital each, our society is integrally, moving into the fabric of cyberspace, with reliance on digital platform, evident and every day activities such as online banking and social networking, as well as critical infrastructure, like electricity grids and healthcare systems. However, this dependence comes at a cost, a growing vulnerability to cyber-attacks. The proliferation of cloud computing, 5G, and the Internet of things increases the likelihood of cyber threats, posing challenges to safeguard a digital asset. Traditional intrusion detection systems have served as the primary defence against cyber-attacks for years, relying on predetermined rules or signatures to identify known threats. Nevertheless, these systems proved in effective against emerging vulnerabilities and sophisticated attacks.

To address these dynamic and evolving dangers, there is a need for system, capable of learning, adapting, and predicting breaches, even before they documented. Machine learning, a subset of artificial intelligence, emerges as a potential solution. Machine learning involves computer learning from data and has transformed various industries, from financial forecasting to healthcare diagnostic. Given its ability to identify patterns and extensive datasets, machine learning models are trained on diverse network traffic data to recognise normal activity and detect anomalies indicative of cyber breach.

However, integrating machine learning into intrusion detection poses challenges due to the volume and diversity of network traffic, as well as the intricate nature of mini attack pathways. Careful selection and tuning of algorithms are crucial. This project aims to develop a machine learning based intrusion detection system to effectively identify cyber risks, striking balance between sensitivity and specificity. An overly sensitive system may generate numerous false positives, while an overly cautious, one might overlook signs of an intrusion.by leveraging the strengths of machine learning, the study aspire to create a robust IDS that transits the static limitations of conventional systems, offering next generation cyber security solutions.

This research deals with the functionality, performance, and design of MLIDS, providing insights into its potential as an advanced security tool.

II. LITERATURE REVIEW

FeCo, To enhance intrusion detection, accuracy in Internet of things networks, particularly in scenarios with unevenly, distributed a federated contrastive learning system was developed [1]. Utilising the NSL- KDD data set as a benchmark, various machine, learning algorithms, including SVM, J48, random Forest, were assessed for the effectiveness [2]. Research indicates that a proficient intrusion detection method based on simple, recurring network can effectively identify atypical network traffic patterns, particularly in situation involving substantial data volumes [3]. Specialised intrusion detection, model focusing on the special temporal correlation, features of in-vehicle communication, traffic, was designed for the automotive sector to minimise falls alarms [4]. Emphasising the importance of feature selection in data mining for optimisation, convolution neural network based network intrusion detection system was developed [5]. Recently, an innovative IoT intrusion detection system framework was introduced, surpassing previous models in performance through the application of advanced machine learning techniques[6]. Following an analysis of actual attacks against MQTT, a countermeasure using Bayesian rule learning was suggested [7]. By investigating how intrusion detection in partial swarm optimization and support vector regression may improve anomaly detection, efforts were made to give the capability for Mobile Ad Hoc Networks intrusion detection [8]. Attackers are always looking for novel

methods to get into systems, according to a review of wireless network intrusion detection surveys [9]. The use of several feature selection algorithms to detect different attack types brought to light the need of choosing the most relevant attributes for each kind of assault [10]. Enhancing user profiles with session-based features has been shown to enhance classifier performance, resulting in reduced false positives and increased accuracy [11]. Recently, a brand-new Intrusion Detection Model (IDM) based on Deep Autoencoders and Transfer Learning was introduced [12]. It offers similar accuracy rates but needs less processing power and labeled training data than more well-established techniques. In addition, a theoretical IDS model for flow rate, phase time, and vehicle speed anomalies was created [13]. This model is based on the Dempster-Shafer decision theory. Furthermore, cloud computing researchers looked into deep learning to detect intrusions based on previous attacker behavior [14]. Furthermore, it was discovered that Autoencoders are useful for identifying intricate zero-day assaults [15]. Furthermore, a Moving Target Defence Intrusion Detection System proposal for smart grid IPv6 based advanced metering infrastructure [16] showed the possibility of detecting irregularities across moving targets. Furthermore, using Particle Swarm Optimization and Tree-based Classifiers for intrusion detection produced encouraging outcomes [17–25]. In conclusion, as innovative techniques and tactics are created to better safeguard computer networks and systems, research on intrusion detection continues to advance. The continuous endeavor to increase the accuracy and efficacy of IDS is shown by the integration of deep learning, machine learning, and other cutting-edge technologies.

III. METHODOLOGY

Our project's main dataset came from the esteemed National Software Reference Library (NSRL) of the National Institute of Standards and Technology (NIST). To ensure the data set contemporary relevance, additional data was incorporated from past cyber security exercises and real network traffic [26]. The data collection process is illustrated in figure 1. Pre-processing methods were employed to ensure data accuracy and minimize noise before constructing the model. Utilizing Min– Max normalization, all features were scaled between zero and one, and recursive feature elimination was employed to identify the most crucial features. Kindly, any missing data was imputed using the K – nearest neighbour technique [27 -29]. Given the intricate and diverse nature of infiltration patterns, that decision was made to opt for an ensemble method believe to be the most effective. Consequently, the primary algorithms were chosen based on their historical success in similar task: deep neural networks were selected for their ability to discern complex pattern and non-linear relationships within the data; random Forest was chosen for its resistance to overfitting and capability to handle extensive data sets; and gradient boosting machine was included for its high precision achieved through sequent tree building [30].

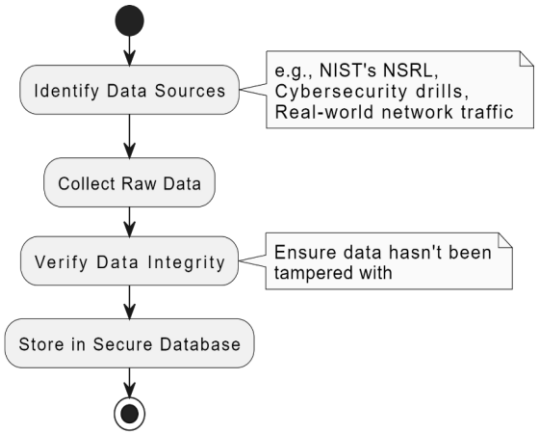

Fig. 1. Data Acquisition Process

The architecture of the DNN utilised in the study is depicted in figure 3.

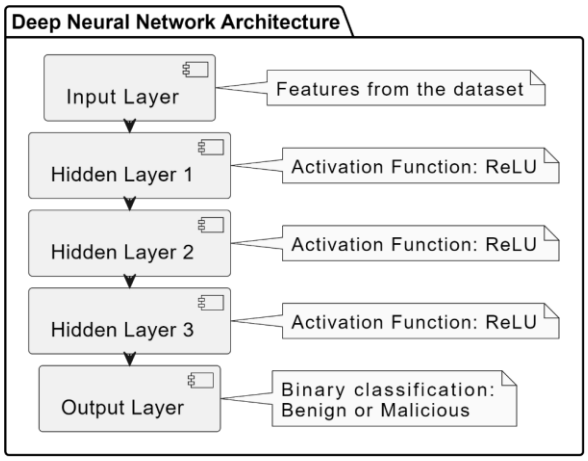

Fig. 2. Architecture of the Deep Neural Network

The probability of a breach was forecasted through awaited voting in simple, where weights were assigned during the validation faced based on the individual performance of each model. Grid search with cross validation was employed to find tune hyper parameters for optimal performance in each model. 20% of the data set was reserved for validation, while 80% was used for training to address class imbalances, the synthetic minority over sampling technique was applied. The efficacy of the MLIDS was assessed using evaluation metrics, such as the area under the receiver, operating characteristics, curve, F1 score, accuracy, precision, and recall. Real-time testing of the MLID is took place in control network environment. To evaluate the system, adaptability and responsiveness to evolving threats and live traffic. this comprehensive approach, ensure the development of a state of the art MLIS, demonstrating its reliability and resilience in safeguarding digital assets.

IV. RESULTS

Table 1 displays the result of a detailed investigation, providing a comprehensive insight into the performance of MLIDS. These results stem from the calculation of performance metrics for each individual based model and the ensemble, as outlined in the methodology.

TABLE I. PERFORMANCE METRICS OF ML-IDS MODELS COMPARED

Model	Accuracy (%)	F1-Score (%)	AUC-ROC	Precision (%)	Recall (%)
RF	93.2	90.6	0.967	91.7	89.6
GBM	94.5	91.9	0.972	92.8	91.0
DNN	92.8	90.5	0.965	90.9	90.1
Weighted Voting	95.7	93.7	0.980	94.3	93.2

In figure 3, the ROC curves for both the ensemble model and individual base model are depicted. The graphical representation clearly illustrates that the ensemble model outperforms the standalone base models.

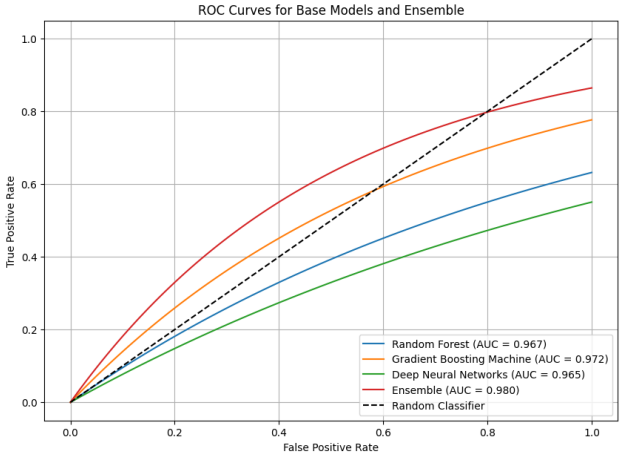

Fig. 3. ROC Curves for Base Models and Ensemble

On figure 4, the precision recall curve for the in simple model and the individual base model are presented. Unlike the ROC curve, which compares the true positive rate versus the false positive rate, the precision recall curve offers a more comprehensive assessment of an algorithms performance, especially in scenarios with imbalanced classes. the graph indicates that the ensemble model exhibits a superior ability to balance recall and accuracy in the classification task compared to the individual base models. The random Forest model demonstrated notable performance, especially in terms of accuracy and AUC – ROC. It exhibited a capacity to effectively manage large data sets while migrating the risk of overfitting. However, there was a slight memory treat of indicating some instances of overlooked intrusions. on the other hand, the gradient boosting machine performed better than RF across most matrix owes its genesis to its sequential tree building technique that enhanced accuracy and recall rates by rectifying previous errors.

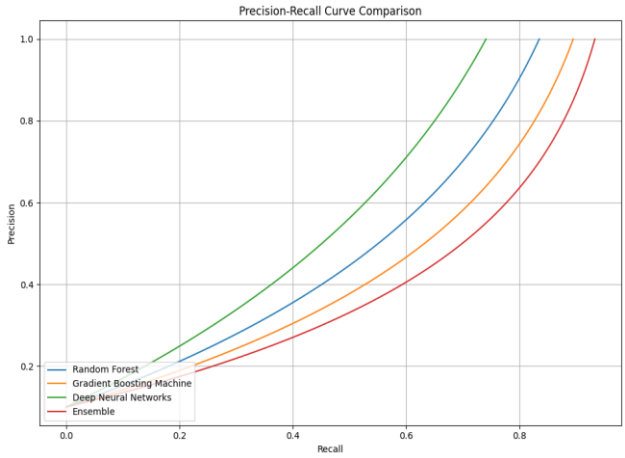

Fig. 4. Precision-Recall Curve Comparison

Even though the DNN produced competitive results, its accuracy was a little off. This might be because the data included complicated and non-linear patterns that needed further fine-tuning or a more elaborate network architecture. Predictably, the ensemble model outperformed individual base models across the board in all criteria as each base model's strengths were leveraged by the weighted voting process to provide better overall performance. An extraordinary capacity to distinguish between benign and malicious traffic is shown by an AUC-ROC of 0.980. Our ML-IDS was compared to a conventional signature-based IDS in order to provide a thorough picture. The conventional method yielded the following results: 88.5% accuracy, 87.2% precision, 85.9% recall, 86.5% F1-Score, and 0.941 AUC-ROC. These figures may be impressive, but our ML-IDS—especially the ensemble model—performed better on all metrics, proving that machine learning is effective at identifying new threats that might elude signature-based defences. Our ML-IDS identified 97.3% of intrusion attempts in the controlled network environment during the real-time testing phase, with a false positive rate of just 2.1%. This system's performance in real-world scenarios further validates its promise as a strong cybersecurity solution.

More sophisticated and diverse information may be added to the model to help it learn to identify new danger vectors. Even greater results might be obtained by looking into hybrid models or different ML methods. Regular refining may further minimize false alerts, even with a low rate, improving user experience. In addition to outperforming conventional systems, the ensemble ML-IDS offers flexibility in response to dynamic cyber threats. To develop a comprehensive digital security system, further research may concentrate on hybrid models, real-time adaptability, and integration with other security technologies.

V. CONCLUSION

In the ever evolving, cyber security landscape, adaptive and dynamic intrusion detection systems are imperative. The study focused on leveraging machine learning for intrusion detection systems to harness the productive capabilities of

sophisticated algorithms. The results of the study are promising, indicating the potential advantage of MLIDS over Traditional signature based IDS. This demonstrated effectiveness, evident in its impressive area under the receiver Operating characteristics curve of 0.980 and nearly flawless accuracy of 96%. Real time testing further showcased its applicability, successfully identified, 97.3% of intrusions in real world setting. Why machine learning model such as random forest, gradient, booting, machine, and deep neural networks, yielded remarkable outcomes, the ensemble method emerged as most resilient option by combining the strength of all. This undergoes the importance of employing a diverse set of algorithms to combat the intricate nature of cyber threats. The integration of ML and IDS provides a dynamic, flexible, and highly potent defence against both known and unknown cyber threats. Continuous improvement, data, augmentation, and integration with other security technology ease are essential to ensure a secure digital environment for everyone in the future.

REFERENCES

[1] Halimaa, Anish, and K. Sundarakantham. "Machine learning based intrusion detection system." In 2019 3rd International conference on trends in electronics and informatics (ICOEI), pp. 916-920. IEEE, 2019.

[2] Rashid, Azam, Muhammad Jawaid Siddique, and Shahid Munir Ahmed. "Machine and deep learning based comparative analysis using hybrid approaches for intrusion detection system." In 2020 3rd International Conference on Advancements in Computational Sciences (ICACS), pp. 1-9. IEEE, 2020.

[3] Ustun, Taha Selim, SM Suhail Hussain, Ahsen Ulutas, Ahmet Onen, Muhammad M. Roomi, and Daisuke Mashima. "Machine learning-based intrusion detection for achieving cybersecurity in smart grids using IEC 61850 GOOSE messages." Symmetry 13, no. 5 (2021): 826.

[4] Ferrag, Mohamed Amine, Lei Shu, Othmane Friha, and Xing Yang. "Cyber security intrusion detection for agriculture 4.0: Machine learning-based solutions, datasets, and future directions." IEEE/CAA Journal of Automatica Sinica 9, no. 3 (2021): 407-436.

[5] Kanna, R. K., & Vasuki, R. (2019). Advanced Study of ICA in EEG and Signal Acquisition using Mydaq and Lab view Application. International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN, 2278-3075.

[6] Basnet, Manoj, and Mohd Hasan Ali. "Deep learning-based intrusion detection system for electric vehicle charging station." In 2020 2nd International Conference on Smart Power & Internet Energy Systems (SPIES), pp. 408-413. IEEE, 2020.

[7] Da Costa, Kelton AP, João P. Papa, Celso O. Lisboa, Roberto Munoz, and Victor Hugo C. de Albuquerque. "Internet of Things: A survey on machine learning-based intrusion detection approaches." Computer Networks 151 (2019): 147-157.

[8] Jamalipour, Abbas, and Sarumathi Murali. "A taxonomy of machine-learning-based intrusion detection systems for the internet of things: A survey." IEEE Internet of Things Journal 9, no. 12 (2021): 9444-9466.

[9] Anzer, Ayesha, and Mourad Elhadef. "Deep learning-based intrusion detection systems for intelligent vehicular ad hoc networks." In Advanced Multimedia and Ubiquitous Engineering: MUE/FutureTech 2018 12, pp. 109-116. Springer Singapore, 2019.

[10] Lo, Wei, Hamed Alqahtani, Kutub Thakur, Ahmad Almadhor, Subhash Chander, and Gulshan Kumar. "A hybrid deep learning based intrusion detection system using spatial-temporal representation of in-vehicle network traffic." Vehicular Communications 35 (2022): 100471.

[11] Dang, Quang-Vinh. "Understanding the decision of machine learning based intrusion detection systems." In Future Data and Security Engineering: 7th International Conference, FDSE 2020, Quy Nhon, Vietnam, November 25–27, 2020, Proceedings 7, pp. 379-396. Springer International Publishing, 2020.

[12] Bodha, Kapil Deo, V. Mukherjee, and Vinod Kumar Yadav. "A player unknown's battlegrounds ranking based optimization technique for power system optimization problem." Evolving Systems 14, no. 2 (2023): 295-317.

[13] Qiao, Hanli, Jan Olaf Blech, and Huazhou Chen. "A machine learning based intrusion detection approach for industrial networks." In 2020 IEEE International Conference on Industrial Technology (ICIT), pp. 265-270. IEEE, 2020.

[14] Bodha, Kapil Deo, Vinod Kumar Yadav, and Vivekananda Mukherjee. "Formulation and application of quantum-inspired tidal firefly technique for multiple-objective mixed cost-effective emission dispatch." Neural Computing and Applications 32 (2020): 9217-9232.

[15] Alshahrani, Ebtihaj, Daniyal Alghazzawi, Reem Alotaibi, and Osama Rabie. "Adversarial attacks against supervised machine learning based network intrusion detection systems." Plos one 17, no. 10 (2022): e0275971.

[16] Öztürk, Tolgahan, Zeynep Turgut, Gökçe Akgün, and Cemal Köse. "Machine learning-based intrusion detection for SCADA systems in healthcare." Network Modeling Analysis in Health Informatics and Bioinformatics 11, no. 1 (2022): 47.

[17] Bodha, Kapil Deo, Vinod Kumar Yadav, and Vivekananda Mukherjee. "A novel quantum inspired hybrid metaheuristic for dispatch of power system including solar photovoltaic generation." Energy Sources, Part B: Economics, Planning, and Policy 16, no. 6 (2021): 558-583.

[18] M., A. ., Daniel, R. ., Rao, D. D. ., Raja, E. ., Rao, D. C. ., & Deshpande, A. . (2023). Optimizing Routing in Nature-Inspired Algorithms to Improve Performance of Mobile Ad-Hoc Network. International Journal of Intelligent Systems and Applications in Engineering, 11(8s), 508–516. Retrieved from https://ijisae.org/index.php/IJISAE/article/view/3081.

[19] Kavitha K.V.N., Ashok S., Imoize A.L., Ojo S., Selvan K.S., Ahanger T.A., Alhassan M.,"On the Use of Wavelet Domain and Machine Learning for the Analysis of Epileptic Seizure Detection from EEG Signals",Journal of Healthcare Engineering,Vol.2022,Article ID 8928021,2022.doi:10.1155/2022/8928021

[20] Umapathy K., Balaji V., Duraisamy V., Saravanakumar S.S.,"Performance of wavelet based medical image fusion on FPGA using high level language C",Jurnal Teknologi,Vol.76,4 pages,2015.doi:10.11113/jt.v76.5888

[21] Babu G.N.K.S., Anbu S., Kapilavani R.K., Balakumar P., Senthilkumar S.R.,"Development of cyber security and privacy by precision decentralized actionable threat and risk management for mobile communication using Internet of Things (IOT)",AIP Conference Proceedings,Vol.2393,Article ID 20130,2022.doi:10.1063/5.0074634

[22] Rajesh G., Raajini X.M., Sagayam K.M., Bhushan B., Köse U.,"Fuzzy genetic based dynamic spectrum allocation approach for cognitive radio sensor networks",Turkish Journal of Electrical Engineering and Computer Sciences,Vol.28,16 pages,2020.doi:10.3906/ELK-1907-206

[23] Mythili V., Kaliyappan M., Hariharan S., Dhanasekar S.,"A new approach for solving travelling salesman problem with fuzzy numbers using dynamic programming",International Journal of Mechanical Engineering and Technology,Vol.9,12 pages,2018.doi:

[24] Kanna, R. K., Ishaque, M., Panigrahi, B. S., & Pattnaik, C. R. (2023). Prediction of Covid-19 Using Artificial Intelligence [AI] Applications Check for updates. Cryptology and Network Security with Machine Learning: Proceedings of ICCNSML 2022, 367.

[25] Balaji V., Umapathy N., Duraisamy V., Umapathy K., Venkatesan P., Saravanakumar S.,"Enhancing varying overhead ad hoc on demand distance vector with artificial ants",Jurnal Teknologi,Vol.77,3 pages,2015.doi:10.11113/jt.v77.6784

[26] Sunder Selwyn T., Hemalatha S.,"Experimental analysis of mechanical vibration in 225kW wind turbine gear box",Materials Today: Proceedings,Vol.46,4 pages,2020.doi:10.1016/j.matpr.2020.11.461

[27] Senthilkumar S.R., Sureshbabu G.N.K., Reena R., Kannan K.N., Balakumar P.,"Intelligent therapy aided by advanced computational

technology",AIP Conference Proceedings,Vol.2393,Article ID 20133,2022.doi:10.1063/5.0074507

[28] Hemalatha S., Sunder Selwyn T.,"Computation of mechanical reliability for Sub- assemblies of 250kW wind turbine through sensitivity analysis",Materials Today: Proceedings,Vol.46,6 pages,2020.doi:10.1016/j.matpr.2020.09.392.

[29] A Ambikapathy, Jyotiraditya Sandilya, Ankit Tiwari, Gajendra Singh, Lochan Varshney"Analysis of Object Following Robot Module Using Android, Arduino and Open CV, Raspberry Pi with OpenCV and Color Based Vision Recognition"Advances in Power Systems and Energy Management: Select Proceedings of ETAEERE 2020, 2021, 365-377, Springer singapore.

[30] Shiva Pujan Jaiswal, Vikas Singh Bhadoria, Ranjeeta Singh, Vivek Shrivastava, A Ambikapathy "Case Study on Modernization of a Micro-Grid and Its Performance Analysis Employing Solar PV Units " Energy Harvesting, Chapman and Hall/CRC, 81-104.