Enhancing Cyber Detection with Machine Learning: A Stacking-Based Approach for Large and Imbalanced Datasets

P. Usha Manikyam[1], G.V.N. Kishore[2], Sheik. Faridha Akather[3]
[1]*Asst Professor, Dept. of CSE, Srinivasa Institute of Engineering & Technology*
[2]*Asst Professor, Dept. of AI&ML, SASI Institute of Technology and Engineering*
[3]*Asst Professor, Dept. of CSE, Srinivasa Institute of Engineering & Technology*

Abstract: The present world has become dependent on cyberspace based on every aspect of our daily life. The usage of cyberspace is rising with each passing day. The world is spending more time on the internet than before. As a result, the risk of cyber threats and cyberattacks is increased. The term cyber threat refers to the illegal activity performed using the internet. Cyber criminals are changing their techniques over time to pass through the wall of protection. Conventional techniques are not capable of detecting zero-day attacks and sophisticated attacks. So that this is to investigate the use of Machine Learning techniques to improve cybersecurity measures, with a particular emphasis on threat detection, prevention, and response. To begin, an examination of the principles of Machine Learning and the importance of this bill to cybersecurity is presented. When it comes to recognising and mitigating cyber threats, a number of different Machine Learning methodologies, including deep learning, signature-based detection, and anomaly detection, are evaluated in terms of how effective they are. Machine Learning based behaviour analysis within the IDS has considerable potential for detecting dynamic cyber threats. I didn't find abnormalities, good identities malicious conduct within the network. However, as the number of data points grows, dimension reduction becomes an increasingly difficult task when training Machine Learning models. At present, we are going to introduce an ML-based network intrusion detection model that uses Random Oversampling (RO) to address data imbalance and stacking feature embedding based on clustering results, as well as Principal Component Analysis(PCA) for dimension reduction, and is specifically designed for large and imbalanced datasets. This model preference is carefully evaluated using three cutting mark datasets: UNSW-NB15, CIC-IDS-2017, and CIC-IDS-2018. On the UNSW-NB15 dataset, the trial shows the RF and ET model your accuracy rates of 99.59% and 99.95%, respectively. Furthermore, using the CIC-IDS-2017 dataset, DT, RF, and ET obtained 99.99% accuracy, while DT and RF models obtained 99.94% accuracy on CIC-IDS-2018. This performance result continuously outperforms the state of our indicating significant process in the field of network impression detection. This achievement demonstrates the efficiency of the suggested methodology, which can be used particularly to accurately monitor and identify network traffic intrusions, thereby blocking possible threats.

Keywords: Machine Learning, Random Oversampling, Cyber Threats, Principal Component Analysis, Dynamic Cyber Threats, Recognising, Mitigating, Network, Dimension, Reduction, Recognising.

I. INTRODUCTION

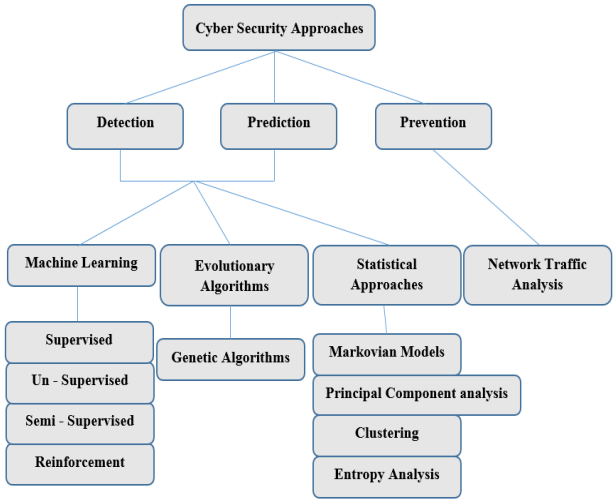
In this digital age, security has become an issue of the utmost importance. Based on the rapid evaluation of threats in terms of both spoken complexities when it tries to combine the never-ending environment of cyberattacks and traditional methods of threat detection and prevention are frequently in effect. In this purpose is to investigate the use of Machine Learning techniques to improve cybersecurity measures with a particular emphasis on prevention, detection, and response to threats. In the field of cybersecurity, it serves as an essential defence against the constantly shifting terrain of cyber threats. In this era, it is categorised by the pervasive presence of digital technology.

Despite the fact that the growth of network technologies, cloud computing, and internet of things (IoT) has reached new levels of convenience and

efficiency, and also made individuals and organisations vulnerable to a wide variety of threats. The foundation of modern Cybersecurity advancements is Machine learning, because it enables the analysis of massive datasets, the recognition of patterns, and the formation of predictions that are essential for the detection, prevention response to threats. The traditional approaches to cybersecurity, despite their importance, frequently struggle to keep up with the level of sophistication and the volume of assaults that are occurring in the modern era. In response to this unrelenting silence, the information of techniques that use Machine Learning as a powerful ally in the fight to perfect digital assets. A dynamic way to augment standard security measures is a significant advancement in the field of security research. Through the utilisation of algorithms that are able to acquire knowledge from Data professionals in the field of cybersecurity, which are able to acquire tools that are of great value in order to proactively protect network systems and sensitive information.

II. CYBERSECURITY AND MACHINE LEARNING APPROACH

Cybersecurity is not one size fits all. It is a layered approach involving prediction, prevention, and detection forward by Machine learning, evolutionary algorithms, statistical models, and traffic analysis. Generally divided into three primary categories:



Detection focuses on identifying malicious activities for security bridges that have already occurred. Detection mechanism helps organisations react from minimise damage. Prediction trends and data to anticipate security threats before they occur. This proactive approach uses advanced analytics and AI to recognise patterns indicative of Future attacks. Prevention enters the system in the place first. This includes enforcing access control, firewalls, encryption, and more.

Moualla et al.[6] Proposed a revolutionary network IDS model that places a crucial role in network security and combines existing cyberattacks on the network, utilising the UNSW-NB15 data as a baseline. It was a dynamically Machine Learning based network with several phases. The imbalance was handled by the SMOTE technique, after which he based on the Gini impurity criterion he tempered the 80 classification and finally tried Extreme Learning Machine(ELM), which was utilised to classify each of the attacks using bibliographies. Using the outputs of employing the ETM classifier as inputs to a fully connected layer, a logistic regression layer was employed to produce soft judgements for all classes and attained 98.43% accuracy.

III. MACHINE LEARNING - CYBERSECURITY NEXUS

It is very important to have a solid understanding of the fundamental principles that support Machine Learning algorithms in order to have a complete comprehension of the combination of Machine Learning and cybersecurity.

With the purpose of focusing on various applications of supervisor Learning and semi-supervisor Learning and semi supervisor Learning in the field of cybersecurity. By using the three types of Machine

Learning fields, the implementation of rigorous training and evolution processes is as follows:

- The selection of features, also known as input variables, has a significant impact on the performance of the model.
- This process of selection and engineering features involves determining which data properties are most effective for a certain endeavour.
- To guarantee the generalizability of the result and cross-validation technique, divide the dataset into training and testing subsets. This makes it possible to conduct more thorough model evaluations.
- Machine Learning systems can automate responses to detect threats, such as isolating compromised devices or recommending mitigating steps, thereby reducing response times and minimizing damage.
- By continuously monitoring the user and network behaviour, Machine Learning models can identify deviations from normal patterns, indicating potential insider threats or compromised accounts.
- Advanced Machine Learning models learn to differentiate between legitimate activities and real threats, reduce the number of false alarms, and also security terms to purpose on genuine risks.
- American credit potential is vulnerable by analysing trends in known attack vectors and historical data, helping organisations to actually address weaknesses.

IV. STACKING-BASED APPROACHES AND TECHNIQUES

1.Data-free Processing and Imbalance to Mitigation:
Random Oversampling (RO): Balances class distribution by replicating minority samples, improving detection of rare attacks.

Stacking Feature Embedding (SFE): integrates clustering results into the feature space to enrich data representation.

2. Dimensionality Reduction:
Principal component analysis (PCA): reduces features phase while preserving critical information, optimising model training.
Linear discriminant analysis (LDA): identifies discriminative features for malware detection, paired with scaling for stability.

3. Stacking Ensembles:
Homogeneous Stacking: combines multiple XGboost classifiers with hyperparameter optimisation, achieving superior F1-scores for underrepresented classes like "Backdoor" or "DoS".
Heterogeneous Stacking: uses meta classifiers to aggregate predictions from base models, improving robustness on datasets like UNSW-NB15 and UGR'16.

V. PERFORMANCE COMPARISON

Approach	Dataset	Key Metrics	Highlights
SFE + RO + PCA [1]	UNSW-NB15	99.95% accuracy (ET)	Outperforms state-of-the-art IDS models
FPA-XGBoost Stacking [3]	UNSW-NB15	Weighted F1-score: 0.95	Effective for highly imbalanced classes
Heterogeneous Stacking [5]	UNSW-NB15	Improved AUC-ROC	Robust to real-world network variability

The above summarizes advanced Machine Learning approaches for intrusion detection using the UNSW-NB 15 cybersecurity dataset.

- SFE + RO + PCA: This approach combines Statistical Feature Engineering (SFE), Random Oversampling (RO), and Principal Component Analysis (PCA), achieving an exceptionally high accuracy of 99.95% using an Extra Tree (ET) classifier. Its performance surpasses the current state-of-the-art intrusion detection system (IDS) models, making it highly effective for identifying threats.

- FPA-XGBoost Stacking: Here, a stacking ensemble method integrates the Flower Pollination Algorithm (FPA) with XGBoost, resulting in a weighted F1-score of 0.95. This method is particularly effective for datasets with highly imbalanced classes, which is common in cybersecurity, ensuring reliable detection of both common and rare attack types.

- Heterogeneous Stacking: This technique uses an ensemble of diverse Machine Learning models to improve the area under the Curve – Receiver Operating Characteristics (AUC–ROC) metric. Its main strength is robustness to real-world network variability, making it suitable for deployment in dynamic and unpredictable environments.

Collectively, this approach demonstrates how tired Machine Learning strategies can significantly enhance interaction detection performance, adaptability, and reliability in cybersecurity applications.

Finally, the system integrates with cybersecurity operations to provide real-time threat detection and response, automating alerts or mitigation actions. This modular approach, Data collection, feature processing, model training, and operational integration, forms the backbone of effective Machine Learning driven cybersecurity.

VI. FUTURE DIRECTIONS

- Extending stacking frameworks to real-time detection systems.
- Exploring advanced clustering techniques for SFE.
- Integrating adversarial training to counter evolving cyber threats

By combining pre-processing innovations with stacking Ensembles, these approaches significantly advance the accuracy and scalability of ML-based intrusion detection systems.

VII. CONCLUSION

This research demonstrates the transformative potential of Machine Learning a in cyber-security, presenting is stacking-based approach that achieves exponential performance on large and imbalanced datasets. The proposed methodology combines random oversampling, stacking feature embedding, and principal component analysis to address critical challenges in network intrusion detection. The results are remarkable, with accuracy rates reaching 95.95% on UNSW-NB15, 99.99% on CIS-IDS-2017, and 99.94% on CIC-IDS-2018 datasets, consistently outperforming state-of-the-art systems. The integration of heterogeneous and homogeneous tagging approaches, combined with advanced pre-processing techniques, efficiently handles data imbalance while maintaining high detection accuracy for both common and trade attack types. This work addresses fundamental cybersecurity challenges, including zero-day attack detection, sophisticated threat recognition, and false positive reduction. The methodologies' scalability and adaptability make them suitable for real-world deployment in an enterprise environment, offering partial solutions for automated threat detection and response. Future research directions include extending frameworks to real-time systems, exploring advanced clustering techniques, and integrating adversarial training to counter evolving threats. This advancement represents a significant step toward proactive cybersecurity defence, positioning Machine Learning as an essential component in protecting digital infrastructure against increasingly sophisticated cyber threats in our interconnected world.

REFERENCE

[1] An intelligent cyber threat detection: A swarm-optimized machine learning approach https://doi.org/10.1016/j.aej.2024.12.039

[2] Cyber Threat Detection Using Machine Learning Techniques: A Performance Evaluation Perspective https://ieeexplore.ieee.org/document/9292388 Published in: 2020 International Conference on Cyber Warfare and Security (ICCWS)DOI: 10.1109/ICCWS48432.2020.9292388

[3] Machine Learning for Cyber Security: Threat Detection, Prevention, and Response https://www.researchgate.net/publication/377990654_Machine_Learning_for_Cybersecurity_Threat_Detection_Prevention_and_Response

[4] Cybersecurity Threat Detection using Machine Learning and Network Analysis https://www.researchgate.net/publication/378892626_Cybersecurity_Threat_Detection_using_Machine_Learning_and_Network_Analysis

[5] Machine learning-based network intrusion detection for big and imbalanced data using oversampling, stacking feature embedding and feature extraction https://journalofbigdata.springeropen.com/articles/10.1186/s40537-024-00886-w

[6] Moualla S, Khorzom K, Jafar A. Improving the performance of machine learning-based network intrusion detection systems on the UNSW-NB15 dataset. Comput Intel Neurosci. 2021; 2021:1–13.