CYBER THREAT DETECTION IN EVENT PROFILES VIA ARTIFICIAL NEURAL NETWORKS

[1] K.Venkata Ramaiah, Associate Professor, Department of CSE, Chalapathi Institute of Technology, Guntur.
[2] Lekkala Madhavi, B.Tech, Department of CSE, Chalapathi Institute of Technology, Guntur.
[3] Bathina Naga Nivesh, B.Tech, Department of CSE, Chalapathi Institute of Technology, Guntur.
[4] Deevi Hari Krishna Deekshith, B.Tech, Department of CSE, Chalapathi Institute of Technology, Guntur.
[5] Golla Gangadhara Rao, B.Tech, Department of CSE, Chalapathi Institute of Technology, Guntur.

Abstract: Introducing an automated and effective cyber-threat detection technique remains a significant challenge in the field of cyber security. This paper aims to address this challenge by presenting an innovative approach that utilizes artificial neural networks and AI technology. Our proposed technique revolutionizes cyber-threat detection by converting a multitude of collected security events into individual event profiles and employing a deep learning-based detection method. To facilitate this research, we have developed an AI-SIEM system that combines event profiling for data pre-processing with various artificial neural network methods, including FCNN, CNN, and LSTM.The primary objective of our system is to distinguish between true positive and false positive alerts, enabling security analysts to respond promptly to cyber threats. To evaluate the effectiveness of our approach, we conducted extensive experiments using two benchmark datasets (NSLKDD and CICIDS2017) as well as two real-world datasets. Through these experiments, we were able to demonstrate the superior performance and accuracy of our AI technique for cyber-threat detection. By introducing this cutting-edge methodology, we aim to contribute to the advancement of cyber security and provide valuable insights for researchers and practitioners in the field. Our innovative approach not only enhances the overall efficiency of cyber-threat detection but also empowers security analysts to combat emerging threats more effectively.

1. INTRODUCTION

The advent of artificial intelligence (AI) techniques has significantly enhanced learning-based approaches for detecting cyber attacks, yielding promising results in numerous studies. However, safeguarding IT systems against evolving threats and malicious behaviors in networks remains a formidable challenge. The constantly changing landscape of cyber attacks necessitates effective defense mechanisms and robust security considerations to identify reliable solutions [1]–[4]. Traditionally, two primary systems have been employed for cyber-threat detection and network intrusion detection. The intrusion prevention system (IPS) is deployed within enterprise networks, utilizing signature-based methods to examine network protocols and flows. It generates security events, or intrusion alerts, and forwards them to another system such as Security Information and Event Management (SIEM). SIEM plays a pivotal role in collecting and managing IPS alerts and is widely recognized as a dependable solution among various security operations tools for analyzing accumulated security events and logs [5]. Despite the efforts of security analysts to investigate suspicious alerts using policies, thresholds, and event correlation analysis, the recognition and detection of intrusions, particularly intelligent network attacks, remain challenging due to high false alerts and the vast amount of security data [6], [7]. Consequently, recent advancements in intrusion detection research have placed a heightened emphasis on leveraging machine learning and artificial intelligence techniques to detect attacks. The progress in AI fields enables timely and automated investigation of network intrusions by security analysts. These learning-based approaches involve training models on historical threat data to learn the attack patterns and subsequently utilize these models to detect intrusions caused by unknown cyber threats [8], [9]. A learning-based method focused on efficiently determining the occurrence of attacks within large datasets can greatly assist analysts in promptly analyzing numerous events. Information security solutions, as mentioned in [10], generally fall into two categories: analyst-driven and machine learning-driven solutions, each with its own merits and applications. While analyst-driven solutions rely on predefined rules set by security experts, machine learning-driven solutions offer the potential to detect rare or anomalous patterns, thereby improving the identification of new cyber threats [10]. However, we have identified several key limitations in existing learning-based approaches used for detecting cyber attacks in systems and networks.

Firstly, learning-based detection methods heavily rely on labeled data for training and evaluating models. Acquiring a sufficient amount of accurately labeled data is a significant challenge, as it is not readily available on a scale required for effective model training. Many commercial Security Information and Event Management

(SIEM) solutions do not maintain labeled data that can be utilized for supervised learning models [10].

Secondly, most of the learning features used in research studies are not generalizable to real-world scenarios as they are not commonly found in network security systems [3]. This lack of generalizability limits the practical applicability of these features. Recent efforts in intrusion detection research have explored automation approaches using deep learning technologies and evaluated their performance on well-known datasets such as NSLKDD [11], CICIDS2017 [12], and Kyoto-Honeypot [13]. However, these benchmark datasets, while accurate, do not adequately represent real-world scenarios due to their limited feature sets. To address this limitation, it is crucial to evaluate learning models using datasets collected from real-world environments.

Thirdly, utilizing anomaly-based methods for network intrusion detection can help identify unknown cyber threats but often leads to a high false alert rate [6]. The generation of numerous false positive alerts incurs significant costs and requires substantial effort from personnel to investigate them thoroughly.

Fourthly, some hackers deliberately alter their behavior patterns gradually to evade detection [10], [14]. Even with appropriate learning-based models in place, the constant evolution of attacker behavior renders these models ineffective over time. Furthermore, the majority of security systems primarily focus on analyzing short-term network security events. To effectively defend against ever-evolving attacks, we propose that analyzing the long-term history of security events associated with event generation can be an effective means of detecting malicious behavior in cyber attacks.

Our research is driven by a set of challenges that serve as the primary motivation for this work. To effectively tackle these challenges, we introduce an innovative AI-SIEM (Artificial Intelligence-Security Information and Event Management) system that leverages deep learning techniques to discern genuine alerts from false positives. By employing our proposed system, security analysts can swiftly respond to cyber threats, even when dealing with an extensive volume of security events. The core of our AI-SIEM system lies in its ability to extract meaningful event patterns by aggregating and correlating security events through a concurrency feature. This process empowers analysts to efficiently handle vast amounts of data by providing concise input for various deep neural networks. Additionally, our system enables analysts to compare current data with long-term historical data, ensuring prompt and efficient data processing. The key contributions of our work can be summarized as follows: Firstly, our proposed system aims to convert a large influx of security events into individual event profiles, enabling the processing of vast-scale data. We have developed a versatile security event analysis methodology that learns normal and threat patterns from extensive datasets while considering their frequency of occurrence. To accomplish this, we propose a method to characterize the datasets using basepoints in the data preprocessing stage. This approach significantly reduces the dimensionality space, overcoming a major challenge commonly associated with traditional data mining techniques in log analysis. Our innovative event profiling method, which incorporates artificial intelligence techniques, distinguishes itself from traditional sequence-based pattern approaches. By providing feature-rich input data, our technique enables the utilization of various deep learning methods. Consequently, our approach surpasses conventional machine learning methods by significantly improving the classification accuracy of true alerts. This breakthrough leads to a remarkable reduction in the number of alerts that analysts need to handle, streamlining their workflow. To validate the applicability of our system, we conducted evaluations using real IPS (Intrusion Prevention System) security events obtained from an actual Security Operations Center (SOC). Performance metrics such as accuracy, true positive rate (TPR), false positive rate (FPR), and F-measure were employed to assess the effectiveness of our system. Furthermore, we compared the performance of our method with five conventional machine learning algorithms (SVM, k-NN, RF, NB, and DT) and evaluated its performance using two widely recognized benchmark datasets in the network intrusion detection field (NSLKDD and CICIDS2017). In our study, we leveraged the TF-IDF (Term Frequency-Inverse Document Frequency) mechanism to decompose a large collection of events into individual event occurrence profiles. These profiles were generated by computing similarity values among TF-IDF event sets and designated basepoints. The resulting event profiles were then fed into the input layers of FCNN (Fully Connected Neural Network), CNN (Convolutional Neural Network), and LSTM (Long Short-Term Memory) models integrated into our AI-SIEM system. We aimed to showcase the practical applicability of our system in defending IT systems against cyber threats by conducting

evaluations using two benchmark datasets and two real datasets collected from operational IPS. While we acknowledge the limitations of the NSLKDD and CICIDS2017 datasets, we utilized them as widely accepted benchmarks for comparing machine learning methodologies. Additionally, we performed a performance comparison with existing methods using the real datasets and the two additional benchmark datasets. Our ultimate goal was to ensure that our system achieves satisfactory performance not only on benchmark datasets but also on real-world data, as this is of utmost importance in practical applications.

2. LITERATURE SURVEY

[1] The research conducted by A. SHABTAI, E. MENAHEM, AND Y. ELOVICI introduces a groundbreaking method called F-Sign, designed to automatically extract unique signatures from malware files. This method targets high-speed network traffic filtering devices that rely on deep-packet inspection. The analysis of malicious executables employs two approaches: disassembly with IDA-Pro and the application of a dedicated state machine to identify the executable's set of functions.

The process of signature extraction in F-Sign involves comparing the executable's functions with those in a common function repository. By eliminating functions present in the repository from the list of potential signature candidates, F-Sign significantly reduces the risk of false-positive detection errors. To further enhance the accuracy of detection, F-Sign introduces intelligent candidate selection using an entropy score to generate signatures. The effectiveness of F-Sign was thoroughly evaluated under various conditions. The results demonstrate that this proposed method is capable of automatically generating signatures that are both highly specific and sensitive to the presence of malware. The research not only highlights the potential of F-Sign in combating malware but also underscores its ability to minimize false-positive rates, making it a valuable tool in the field of cyber security.

[2] The research paper authored by D. Kong, J. Gong, S. Zhu, P. Liu, and H. Xi introduces a pioneering approach named SAS (Semantics Aware Statistical) for the automatic generation of effective worm signatures in the presence of adversarial environments. While string extraction and matching techniques have been commonly employed for signature generation, dealing with the challenges posed by an adversarial setting remains a significant problem.

In such environments, attackers possess the capability to manipulate byte distributions within attack payloads, allowing them to inject well-crafted noisy packets that contaminate the suspicious flow pool. To counteract these attacks, the SAS algorithm is proposed. When processing packets within the suspicious flow pool, SAS utilizes data flow analysis techniques to remove non-critical bytes. Subsequently, a hidden Markov model (HMM) is applied to the refined data, enabling the generation of signatures based on state-transition graphs. Notably, this work represents the first endeavor to combine semantic analysis with statistical analysis to automatically generate worm signatures. Through extensive experiments, it was demonstrated that the proposed SAS technique exhibits accurate worm detection capabilities, utilizing concise signatures. Furthermore, the results indicate that SAS demonstrates enhanced resilience against changes in byte distribution and noise injection attacks when compared to existing approaches like Polygraph and Hamsa. This research contributes significantly to the field of worm detection by addressing the challenges posed by adversarial environments and offering a novel and robust approach for signature generation.

[3] The rapid expansion of cloud services, the proliferation of users, dynamic changes in network infrastructure connecting mobile operating system devices, and continuous advancements in network technology have presented unprecedented challenges in the realm of cyber security.

These challenges were unforeseen, necessitating the development of novel approaches to counter emerging threats. Network security mechanisms, sensors, and protection schemes must adapt and evolve to cater to the needs and address the complex problems faced by modern-day users. The landscape of cyber security is constantly evolving, requiring proactive measures to stay ahead of the ever-evolving threat landscape and safeguard sensitive information in this dynamic environment.

[4] Efficiently detecting deviations from normal behavior in computer networks poses a significant challenge in traffic monitoring. In this research paper, authored by M. H. A. C. ADANIYA, M. F. LIMA, J. J. P. C. RODRIGUES, T. ABRAO, AND M. L.

PROENCA, two models are presented for network anomaly detection utilizing flow data, specifically bits and packets per second. The first model leverages the Firefly Algorithm, while the second model utilizes the Genetic Algorithm.

Both models were evaluated extensively to assess their effectiveness in detecting network anomalies, and the obtained results were meticulously compared. The research team experienced positive outcomes using data collected at the backbone of a university network. This implies that the proposed models demonstrate promising capabilities in identifying anomalies within network traffic, contributing to enhanced network security. By employing advanced algorithms and analyzing flow data, these models aim to enhance the overall efficiency of network anomaly detection, enabling proactive responses to potential threats in computer networks. The findings of this study provide valuable insights into the field of traffic monitoring and anomaly detection, paving the way for further advancements in network security.

[5] Network anomaly detection plays a critical role in network management, ensuring quality of service (QoS), security, and overall network integrity. The ever-evolving landscape of anomalies and attacks poses an ongoing challenge, requiring effective strategies to safeguard networks. Most existing network anomaly detection systems employ supervised approaches, relying on either signature-based detection methods or supervised learning techniques. However, these approaches have significant limitations. Signature-based methods fail to detect and characterize unknown anomalies, leaving networks vulnerable for extended periods. On the other hand, supervised learning requires training data with labeled traffic, which is often difficult and costly to obtain.

These limitations present a significant bottleneck in addressing network anomaly detection. To overcome these challenges, we propose an unsupervised approach that can detect and characterize network anomalies without depending on signatures, statistical training, or labeled traffic. This represents a major advancement towards achieving network autonomy. Our unsupervised detection method utilizes robust data-clustering techniques, combining sub-space clustering with evidence accumulation or inter-clustering results association, to blindly identify anomalies within traffic flows. Additionally, we perform correlation analysis on the results of unsupervised detection to enhance robustness. Characterization of detected anomalies is accomplished by constructing efficient filtering rules that describe the nature of the identified anomalies. To evaluate the performance of our unsupervised approach, we conducted experiments using real network traffic, providing a realistic assessment of its effectiveness in detecting and characterizing network anomalies. By introducing an unsupervised approach that does not rely on signatures or labeled traffic, our research takes a significant stride towards achieving more autonomous network anomaly detection. This novel methodology demonstrates the potential to improve network security by autonomously detecting and characterizing anomalies, reducing the reliance on manual intervention and expensive training data.

3. EXISTING SYSTEM

Despite the usefulness of learning-based approaches in detecting cyber attacks in systems and networks, several limitations have been observed in existing methods. Firstly, these approaches heavily rely on labelled data for model training and evaluation. However, acquiring a sufficient amount of labelled data at scale is a challenging task, especially since commercial Security Information and Event Management (SIEM) solutions often lack the necessary labelled data for supervised learning models. Secondly, many of the learning features utilized in research studies are not applicable to real-world scenarios as they are not commonly found in standard network security systems. This lack of generalized features makes it difficult to implement these approaches in practical cases. While recent efforts in intrusion detection research have explored automation using deep learning technologies and evaluated performance using benchmark datasets like NSLKDD, CICIDS2017, and Kyoto-Honeypot, these datasets, although accurate, may not accurately represent real-world scenarios due to their limited features. To overcome this limitation, it is crucial to evaluate learning models with datasets collected from real-world environments. Thirdly, employing anomaly-based methods for network intrusion detection can help identify unknown cyber threats, but it often results in a high rate of false positive alerts. Dealing with numerous false alerts is not only costly but also requires significant effort from personnel to investigate each alert. Fourthly, malicious hackers can intentionally alter their behaviour patterns gradually to evade detection. This constant adaptation makes detection models

ineffective, even if appropriate learning-based models are employed. Additionally, most security systems focus on analyzing short-term network security events, which may not be sufficient to defend against continuously evolving attacks. To effectively detect malicious behaviour in cyber attacks, analyzing the long-term security event history associated with the generation of events can be a viable approach. Addressing these limitations is crucial for developing more robust and effective learning-based approaches in cyber security. By overcoming the challenges related to labelled data, incorporating generalized features, reducing false positive rates, and considering long-term event analysis, we can enhance the accuracy and practicality of learning-based detection systems in real-world cyber security scenarios.

DISADVANTAGES OF EXISTING SYSTEM

- Accuracy is Low.

4. PROPOSED SYSTEM

This research paper introduces a novel approach to threat detection utilizing the AI-SIEM (Artificial Intelligence-Security Information and Event Management) technique. By harnessing the power of deep learning algorithms, specifically FCNN (Fully Connected Neural Network), CNN (Convolution Neural Network), and LSTM (Long Short-Term Memory), the proposed technique leverages event profiling, including attack signatures, to identify and mitigate potential threats. To evaluate the performance of the proposed approach, the author compares it with traditional algorithms such as SVM (Support Vector Machine), Decision Tree, Random Forest, KNN (K-Nearest Neighbours), and Naïve Bayes. In this implementation, the focus lies on implementing CNN and LSTM algorithms, which have proven to be effective in various domains. By utilizing the AI-SIEM technique and employing deep learning algorithms, this research aims to enhance the accuracy and efficiency of threat detection. Through comparative analysis with conventional algorithms and the utilization of specific CNN and LSTM models, the proposed approach seeks to provide insights into the efficacy of this advanced methodology for detecting and mitigating security threats.

ADVANTAGES

- Accuracy is Very high.

SYSTEM ARCHITECTURE

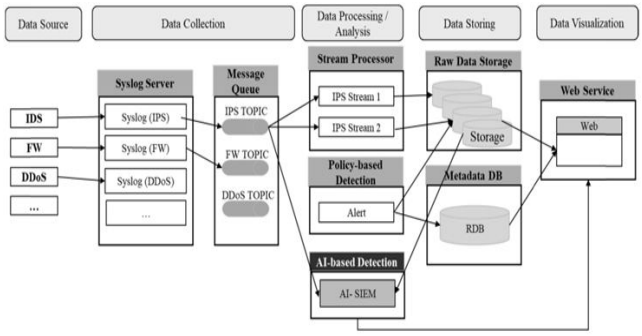

Fig 1: System Architecture

5. DATASET:

We can collect the dataset from the kaggle.com site and placed into our project folder used for detecting the cyber attacks.

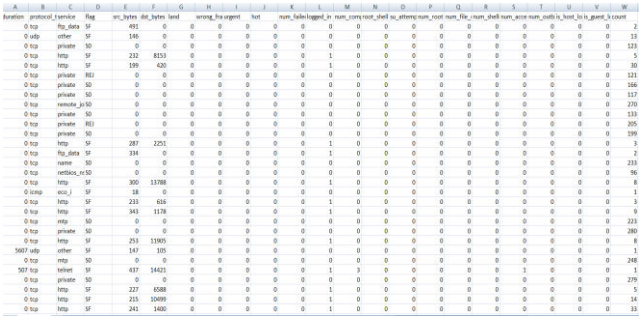

Table 5.1: Cyber Attack data

6. UML DIAGRAMS

1. CLASS DIAGRAM

The cornerstone of event-driven data exploration is the class outline. Both broad practical verification of the application's precision and fine-grained demonstration of the model translation into software code rely on its availability. Class graphs are another data visualisation option.

The core components, application involvement, and class changes are all represented by comparable classes in the class diagram. Classes with three-participant boxes are referred to be "incorporated into the framework," and each class has three different locations:

• The techniques or actions that the class may use or reject are depicted at the bottom.

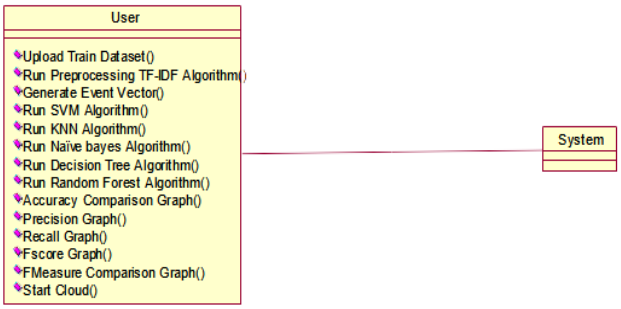

Fig 6.1 shows the class diagram of the project

2. USECASE DIAGRAM:

A use case diagram in the Unified Modeling Language (UML) is a type of behavioral diagram defined by and created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The main purpose of a use case diagram is to show what system functions are performed for which actor. Roles of the actors in the system can be depicted.

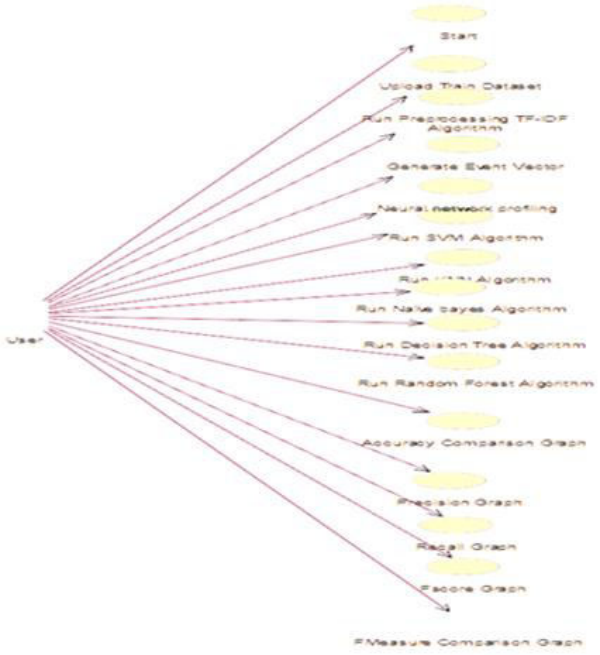

Fig 6.2 Shows the Use case Diagram

3. SEQUENCE DIAGRAM:

A sequence diagram in Unified Modeling Language (UML) is a kind of interaction diagram that shows how processes operate with one another and in what order. It is a construct of a Message Sequence Chart. Sequence diagrams are sometimes called event diagrams, event scenarios, and timing diagrams.

Fig 7.3 Shows the Sequence Diagram

7. RESULTS

7.1 Output Screens

Fig 7.1 Upload the Dataset

In above screen click on 'Upload Train Dataset' button and upload dataset

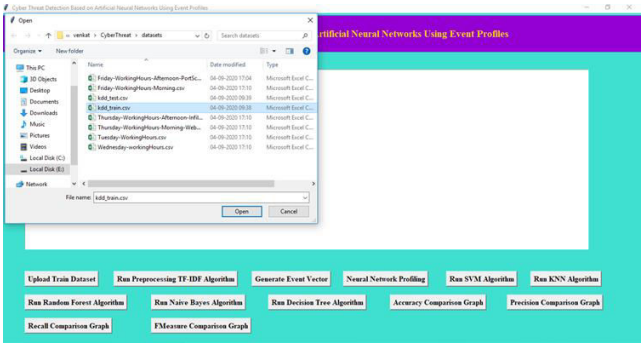

Fig 7.2 Uploading the Dataset File

In above screen uploading 'kdd_train.csv' dataset and after upload will get below screen.

Fig 7.3 Preprocess the dataset

In above screen we can see dataset contains 9999 records and now click on 'Run Preprocessing TF-IDF Algorithm' button to convert raw dataset into TF-IDF values

Fig 7.4 Run the Generate Event Vector Algorithm

In above screen TF-IDF processing completed and now click on 'Generate Event Vector' button to create vector from TF-IDF with different events

Fig 7.5 Run the LSTM and CNN Algorithm

In above screen we can see algorithms accuracy, precision, recall and FMeasure values. Now click on 'Run SVM Algorithm' button to run existing SVM algorithm

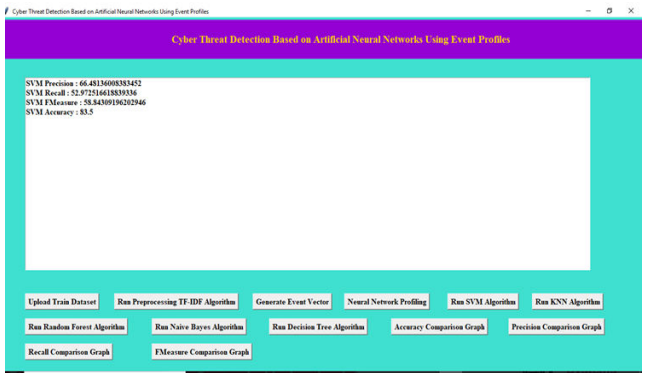

Fig 7.6 Run the SVM Algorithm

In the above screen shows the new remote user registration form.

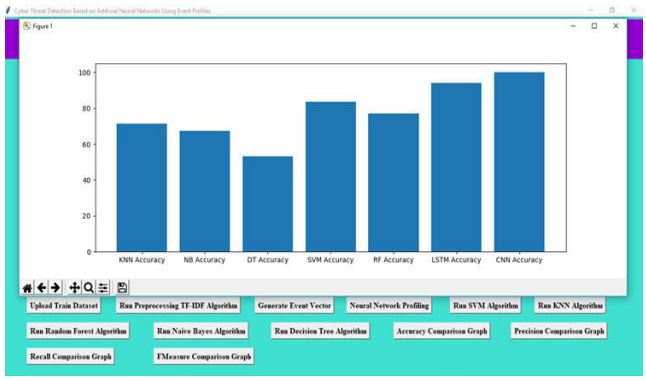

Fig 7.7 Accuracy Comparison Graph

In above graph x-axis represents algorithm name and y-axis represents accuracy of those algorithms and from above graph we can conclude that LSTM and CNN perform well.

8. CONCLUSION

In this research paper, we propose an innovative AI-SIEM system that leverages event profiles and artificial neural networks to enhance cyber-threat detection capabilities. Our approach stands out by efficiently condensing large-scale data into event profiles and utilizing deep learning-based detection methods. The AI-SIEM system empowers security analysts to effectively handle significant security alerts by comparing long-term security data. It also aids in reducing false positive alerts, enabling swift responses to cyber threats scattered across numerous security events. To evaluate the system's performance, we conducted a comparative analysis using two benchmark datasets (NSLKDD and CICIDS2017) as well as real-world datasets. Firstly, through the comparison experiments with her methods using well-known benchmark datasets, we

demonstrated that our mechanisms can serve as effective learning-based models for network intrusion detection. Secondly, by evaluating our technology with two real datasets, we showcased promising results, surpassing conventional machine learning methods in terms of accurate classifications. Moving forward, our future research aims to address the evolving challenge of cyber attacks by focusing on enhancing early threat predictions through a multiple deep learning approach that captures long-term patterns in historical data. Additionally, we plan to collaborate with SOC analysts to improve the precision of labeled datasets for supervised learning, making concerted efforts to record labels of raw security events over an extended period of time.

9. REFERENCES

[1] S. Naseer, Y. Saleem, S. Khalid, M. K. Bashir, J. Han, M. M. Iqbal, and K. Han, 'Enhancing Network Anomaly Detection through Deep Neural Networks,' IEEE Access, vol. 6, pp. 48231–48246, 2018.

[2] B.-C. Zhang, G.-Y.Hu, Z.-J.Zhou, Y.-M.Zhang, P.-L.Qiao, and L.-L. Chang, 'Network Intrusion Detection using Directed Acyclic Graph and Belief Rule Base,' Electronic Telecommunications Research Institute Journal, vol. 39, no. 4, pp. 592–604, Aug. 2017.

[3] W. Wang, Y. Sheng, and J. Wang, 'HAST-IDS: Learning Hierarchical Spatial-Temporal Features for Improved Intrusion Detection,' IEEE Access, vol. 6, pp. 1792–1806, 2018.

[4] M. K. Hussein, N. Bin Zainal, and A. N. Jaber, 'Data Security Analysis for DDoS Defense in Cloud-Based Networks,' in Proceedings of the IEEE Student Conference on Research and Development (SCOReD), Kuala Lumpur, Malaysia, Dec. 2015, pp. 305–310.

[5] S. S. Sekharan and K. Kandasamy, 'Profiling SIEM Tools and Correlation Engines for Security Analytics,' in Proceedings of the International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), Mar. 2017, pp. 717–721.

[6] N. Hubballi and V. Suryanarayanan, 'Minimizing False Alarms in Signature-Based Intrusion Detection Systems: A Survey,' Computer Communications, vol. 49, pp. 1–17, Aug. 2014.

[7] A. Naser, M. A. Majid, M. F. Zolkipli, and S. Anwar, 'Trusting Cloud Computing for Personal Files,' in Proceedings of the International Conference on Information and Communication Technology Convergence (ICTC), Busan, South Korea, Oct. 2014, pp. 488–489.

[8] Y. Shen, E. Mariconti, P. A. Vervier, and G. Stringhini, 'Tiresias: Deep Learning-based Prediction of Security Events,' in Proceedings of the ACM Conference on Computer and Communications Security (CCS), Toronto, ON, Canada, Oct. 2018, pp. 592–605.

[9] K. Soska and N. Christin, 'Automatically Detecting Vulnerable Websites Before They Turn Malicious,' in Proceedings of the USENIX Security Symposium, San Diego, CA, USA, 2014, pp. 625–640.

[10] K. Veeramachaneni, I. Arnaldo, V. Korrapati, C. Bassias, and K. Li, 'AI2: Training a Big Data Machine to Defend,' in Proceedings of the IEEE BigDataSecurity HPSC IDS, New York, NY, USA, Apr. 2016, pp. 49–54.

[11] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, 'A Detailed Analysis of the KDD Cup 99 Dataset,' in Proceedings of the 2nd IEEE Symposium on Computers and Communications for Security and Defense vol. 5, no. 1, pp. 291–302, 2013.