

Research Article

Scalable AI-based Big Data Framework of Real-time Intrusion Detection and Threat Analytics based on Behavior

K.Prasanthi¹, P.Venkata Krishna², V.Saritha³

^{1,2,3}Dept. of Computer Science and Engineering, Sri Padmavati Mahila Visvavidyalayam, India, 2Dept. of Computer Science, Sri Padmavati Mahila Visvavidyalayam, India.

*Corresponding author: prasanthikadirimangalam@gmail.com

Received: 17/01/2025,

Revised: 23/04/2025,

Accepted:11/05/2025

Published:31/05/2025

Abstract: - Exponential growth in the amount of data and cyber threats is creating a challenge to the conventional security systems, which are not in a position to process and analyze large networks of heterogeneous data in real time. This study proposes Scalable Architecture for Big Data-based Attack Detection Applications, a modular and comprehensive framework that aims at facilitating intelligent detecting of cyber-attacks in distributed settings which uses artificial intelligence and deep learning to detect cyber-attacks. Proposed framework enables to process very large volumes of data that are very heterogeneous and they are coming out of all kinds of different sources network traffic, or cloud platforms, or IoT devices, or enterprise logs uses This architecture incorporates the distributed processing technology such as Apache spark and HDFS to allow a scalable intake, transformation, and real-time analysis of data. It has a deep learning-driven detection that is incorporated as part of the processing pipeline that employs neural networks to detect anomalous behaviors and emergent threats. In contrast to traditional rule-based systems, the Scalable Big data based architecture allows an adaptive learning and pattern recognition involving both historical data and streaming data. The framework is tested on benchmark datasets such as CICIDS2017 and UNSW-NB15 and proves to have high accuracies, poor false positives rates and perform efficiently. It offers the capability of proactive monitoring of cybersecurity in a complex environment with a powerful, scalable, and future-proof solution courtesy of its big data infrastructure and smart, AI-driven analytics

Keywords: - Anomaly Detection, Big Data Architecture, Cybersecurity, Cyber Threat Analytics, Deep Learning, Scalable Framework.

1. Introduction

The Frequency, scale, and complexity of cyber threats in the current hyper-connected digital environment have never been this high. Risks like the use of data breaches, ransom ware, insider threats, and distributed denial-of-service (DDoS) attacks are another volume of risk that has an even longer list of victims that include organizations across very critical areas of the economy such as finance and healthcare, government and cloud service providers. As recent studies show, the cost of cybercrime worldwide is expected to be around 10.5 trillion every year by 2025 [1]. Traditional rule-based and signature-based intrusion detection systems (IDS) have been ineffective in detecting new or zero-day attacks as cyber-attacks are increasingly becoming evolving and flexible [2].

At the same time, outbreak of data created by users, equipment and applications, and infrastructure makes cybersecurity harder to operate. Such large quantity of high velocity and heterogeneous data widely known as big data,

needs to be processed in scale and processed with smart analytics [3], [4]. Processing of such data includes issues pertaining to real-time ingestion, noise filtering, effective storage and timely analysis to derive actionable intelligence.

Recent advances in deep learning (DL) and artificial intelligence (AI) have enabled up revolutionary possibilities in cyber security. CNNs, LSTM, and GRUs are all forms of deep neural networks, and they have proved to be effective in spatial and temporal attack patterns capture [6]. The models are good at pointing out minute variations in normal behavior and discovering deeper relationships among the data. But they are computationally heavy and require well scaled frameworks, distributed systems, to be deployed. The current paper proposes Scalable Architecture for Big data based Detection Applications, unified framework of both big data based on threat detection based on deep learning, as well as a scalable and distributed iteration on data infrastructure. Apache Spark is a big data technology, and proposed approach uses this technology to ensure high-throughput data processing and failure resistant data storage



[5]. The architecture has the capacity to perform real-time and batch processing, and as a result, it can identify emerging threats in different spheres, such as IoT networks, cloud platforms, and enterprise systems.

In contrast to the traditional security systems, the proposed scalable architecture will enable heterogeneous data sources and open integration of AI models. It's structured architecture contains modular way to ingest all data, pre-processing and then transform those features and then put it in this deep learning system for classification and then the data can be visualized. The framework enables the analysis of threats in the past and at the moment of occurrence, which enables the cyber security professionals to better respond to incidents.

In order to measure its performance, Scalable Architecture for Big Data-based Attack Detection Applications is experimented with on baseline datasets such as CICIDS2017 and UNSW-NB15 [10], [11]. The experimental findings indicate that proposed framework is more accurate in detection, false positives and scalable than the conventional methods.

The significant contributions of this study is described as follows

1. Proposed a Scalable architecture for big data based attack detection for real-time and batch-mode cyber-attack detection by using AI-based models.
2. The deep learning techniques are integrated to discover requirements of various temporal and spatial based attacks in big data of heterogeneous nature.
3. Employed the Apache Spark and HDFS to parse, transform and analyse its data in large-scale environments in a performance-sensitive way in a distributed environment.
4. Benchmark data sets are evaluated to demonstrate the enhanced accuracy for detection and reducing the false positive rates.

This paper is organized to systematically address the development and evaluation of a scalable, deep learning-driven framework for cyberattack detection using big data technologies. It begins with an Introduction outlining the pressing challenges of modern cybersecurity, including the shortcomings of traditional intrusion detection systems and the opportunities presented by AI and big data platforms such as Apache Spark. Following this, the Related Work section surveys existing approaches in intrusion detection, emphasizing the limitations of current solutions in handling large-scale, real-time, and heterogeneous data, and highlighting the need for integrated architectures that combine big data processing and AI-based analytics. The Proposed Methodology section introduces a modular framework designed for scalable cyberattack detection, comprising layers for data ingestion, preprocessing, AI-based detection, threat scoring, and alert visualization. Each component is elaborated with technical detail, including how GRU, CNN, and hybrid models are employed for temporal and spatial pattern recognition. The architecture is shown to support both batch and real-time processing using distributed computing infrastructure. Next, the Experimental Results and Discussion section presents comprehensive performance evaluations using benchmark

datasets such as CICIDS2017 and UNSW-NB15, comparing detection accuracy, inference latency, and scalability against baseline models. Finally, the Conclusion and Future Work highlights the contributions of the framework in terms of detection precision, scalability, and real-time responsiveness, and outlines potential extensions such as online learning and deployment in edge or federated environments.

2. Related Work

Cyber threats are becoming more sophisticated with the second explosion of digital data that has encouraged researchers to explore new security architectures beyond the classical rule-based systems. In spite of being deployed extensively, the usefulness of signature-based intrusion detection systems (IDS) is restricted in adversely identifying new threats and dynamically changing attack patterns [12]. Such systems are easily ineffective against zero-day exploits, and behaviorally evasive malware especially in high volume settings.

To overcome such limitations, big data structure has been included to facilitate the processing and management of huge security data. The Internet environment will require cybersecurity tools that support parallel processing, in-memory computation, scalable data management, Apache Hadoop, and Apache Spark among others, that will be essential to cybersecurity affairs in the future [13], [14]. In particular, Apache Spark has gained popularity as a tool of choice to develop real-time data analysis pipelines because of these low latency and the ability to process data in batches and streams concurrently [15]. Developments in artificial intelligence, specifically deep learning, have paralleled the development of data infrastructure and provided a potent instrument of identifying slight abnormalities and attack patterns within large amounts of data.

Convolutional Neural Networks (CNN) effectively learns about spatial features of the packet and flow level network traffic data to classify intrusion patterns which cannot be learned via conventional feature engineering [16]. Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) networks belong to the family of architectures that are able to work well with time relationships and sequential dependencies in event logs and streams of user activity [17].

Spatial-temporal features have also been exploited with the hybrid deep learning models combining CNN and RNN to be proposed. A combination of these has proved to be better on identifying coordinated attacks and threat patterns which are time-evolving [18]. Although the models are accurate, such deep learning models also involve lots of backend infrastructure to support training and inference at scale.

A recent study of combining deep learning and big data platforms that utilize distributed frameworks to enable high-throughput detection as well as large-scale deployment has been developed [19]. These methods are capable not only of detection, but of adaptive learning of ever-changing threat landscapes, which make them the most appropriate tools in enterprise-level security monitoring.

One of the benchmark datasets utilized for assessment of cybersecurity frameworks is CICIDS2017 and UNSW-NB15. These datasets also present a range of categories of attacks, traffic patterns, and types of user activities that can be used in the training and tests of AI-based IDS modeling [20],[21].

A number of prior researches have tried to bridge big data analytics with cybersecurity, however, most of them were not end-to-end solutions that successfully integrated scalable architecture-based deep learning capabilities. In study [22], a big data-oriented architectural design of cyberattack detection by using such technologies as Apache Hadoop, Spark, etc. The adopted architecture covered the scalable data ingestion and processing workflow and lacked deep learning-based modules of detection or real-time visualization.

But most of the available literature either addresses the development of models or engineering of data without a central framework that integrates the two. To overcome the gap, Scalable Big Data-based Attack Detection

Applications architecture offers an end-to-end, module-based architecture that is based on distributed big data processing as well as capable of making use of the AI-based detection engines to provide reliable intrusion detection by being accurate and scalable in the detection process.

3. Proposed Methodology

The proposed Scalable Architecture for Big Data-based Attack Detection Applications framework described in Figure 1 is intended to address the needs of the scalable and efficient detection of cyber threats in high volumes specifically using deep learning models embedded into a distributed pipeline formed by the process of big data. The architecture is modular and layered, where each layer aims at responding to particular functional needs in cybersecurity analytics. The framework can be used in real time detection as well as in batch-mode detection, which makes the framework highly customizable to the various enterprise setups.

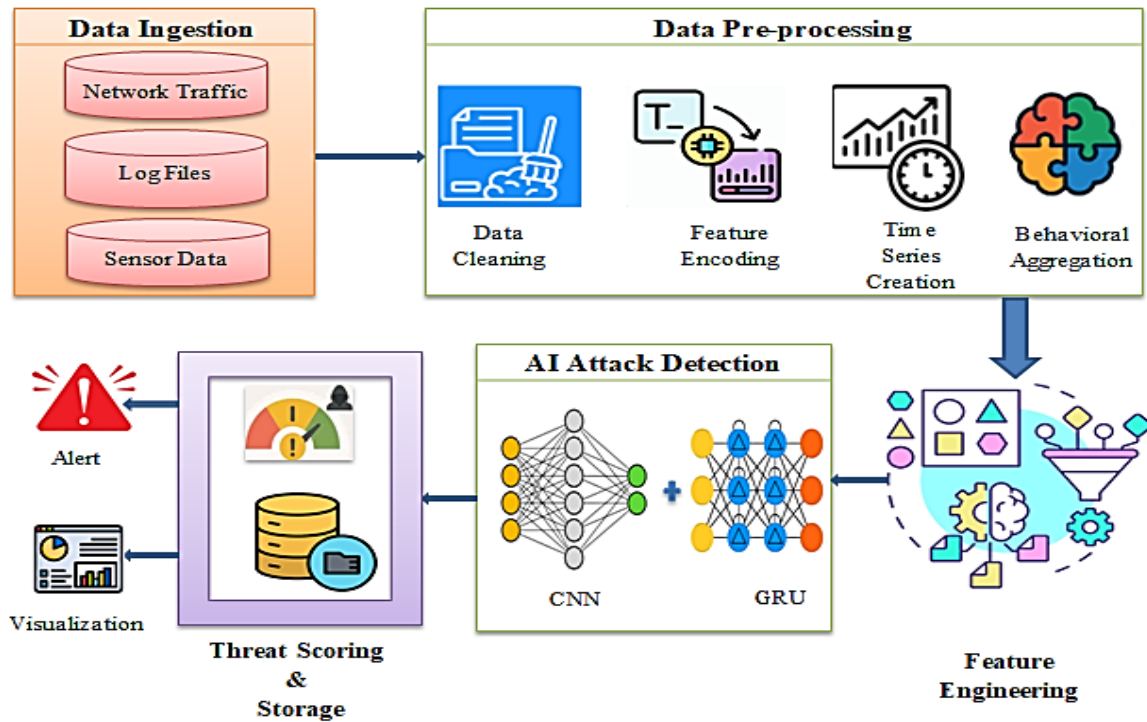


Fig 1: Proposed Framework for Attack Detection

A. Data Ingestion Layer

The architecture primary layer is involved with the process of collecting data provided by several heterogeneous sources. In contemporary infrastructures, information is generated by very diverse systems, such as Network traffic analyzers include Wireshark, NetFlow or PacketBeat, Cloud services such as aws, Microsoft Azure and Google cloud. Windows/Linux endpoint systems, server logs, firewall logs and virus scanning protocols. IoT devices and edge nodes contains the data related to streaming telemetry data and event data.

The tools used in this layer are Apache Kafka or Flume to aggregate the logs in real time and it takes data in various formats e.g. JSON, CSV, XML and PCAP. The acquired

data is passed on to the pre-processing layer and it undergoes clean up and normalization.

B. Pre-processing Layer

Cybersecurity raw information is usually noisy, incomplete, and inconsistent. This layer deals with:

Data cleaning: Elimination of duplicate or null values, correction of mismatched fields and noise including crashed pings or benign background traffics.

Feature encoding: One-hot encoding with Categorical features (i.e. protocol type, flag status) are coded so that they can be compatible with neural networks.

$$OneHot(C) = [\delta(C, c_1), \delta(C, c_2) \dots \delta(C, c_k)] \quad (1)$$

$$\text{where } \delta(a, b) = \begin{cases} 1 & \text{if } a = b \\ 0 & \text{if } a \neq b \end{cases}$$

Time-series creation: Time-series creation Events are aggregated and converted to sequences of instances in terms of session ID, user behavior or time windows and are used to learn sequential models (such as GRUs).

Behavioral aggregation: behavioral aggregation groups behavioral components related to the same host or user (e.g. number of login attempts, bytes transferred), aggregated to form behavioral vectors. Apache Spark [4] is applied in such cases because of distributed processing that allows parallelly feature engineering at scale.

C. AI Based Attack Detection Layer

In this layer, there is deep learning models [6] that can learn complex patterns and be able to predict attacks accurately. Architecture supported is as follows:

Gated Recurrent Unit (GRU): GRUs are a perfect fit to interpretations of time-series based behaviours like sequence of log-ins, network traffic or the activity of clients.

Convolutional Neural Networks (CNNs): CNNs are trained on the extraction of spatial features and they are deployed on log sequences or flow matrices as images or grid of numerical values.

Hybrid CNN-GRU model: In other cases a CNN might be followed by a GRU in order to learn temporal context [8]. This hybrid solution is associated with adding robustness in the detection of complicated multi-stage attacks.

The models trained are either loaded into Spark or used with TensorFlowOnSpark where the inference is performed on distributed data. The scores that such models provide are the probabilities of an instance being malicious.

D. Threat Scoring Layer & Storage

After a prediction is established, this layer assigns a threat level to every event such that it can be prioritized as far as response actions are concerned. The scoring individual is a cumulative measure of a variety of aspects:

Model probability output: The softmax output from the deep learning model.

Behavioral similarity scored score: Cosine similarity of the observed behavior vector and that known threat profile as.

Historical frequency: the frequency of occurrence of the above events in the past and the severity connected with those events.

All the data about the threat and the scores are stored in a scalable NoSQL database where it is possible to work fast

with the database and connect it to other downstream systems.

The formula of scoring will be:

$$TS = \alpha \cdot P_{model} + \beta \cdot S_{behavior} + \gamma \cdot H_{freq} \quad (2)$$

where $\alpha + \beta + \gamma = 1$ are tunable weights

E. Alert layer and Visualization

In the alert layer, custom alerting mechanisms can be sent through different mediums including the usage of emails, SMS or SIEM platforms. They can be presented in any common standardized protocol, such as JSON, Syslog, or Common Event Format (CEF) to be interoperable. Each alert is accompanied by the category of threat, its confidence rating, time stamp, and critical metadata so that they can be reviewed very briefly.

Apart from that, there is an investigation interface with the possibility to perform search in details including emotional reactions on raw log data and on coded features as well as include user behavior vectors and model predictions. This end-to-end visualization and alerting platform will guarantee that analysts are provided with approaches to intuitive and data-driven discoveries and that can lower the false positives drastically and speed up remediation processes.

The layer transforms detailed threat scoring and detection results to visual realizations able to take action on and provide information to perform effective incident response and forensic investigations. First, the threat scores created at the AI-based detection layer are prioritized in terms of severity so that the analyst can prioritize the response and such scores are used to develop ranked alerts such that triage and quick response is achieved.

The proposed framework connected to interactive dashboards constructed on such platform as Kibana. These dashboards provide real time displays such as timelines of attack progresses, heat mapping of areas of interest of malicious activity across IP or user, trend analysis of anomalies over various periods in time. Top-N types of attacks, most affected systems and distributions of activity by user can be displayed by analysts.

4. Experimental Results and Discussion

This section compares the proposed framework based on accuracy of classification results, percentage of accuracy per class, processing runtime and cost of resources. Baseline model comparators also made a part in the investigation to clarify the light on the positives of the hybrid deep learning as well as the big data-based models made by scalable big data based attack detection.

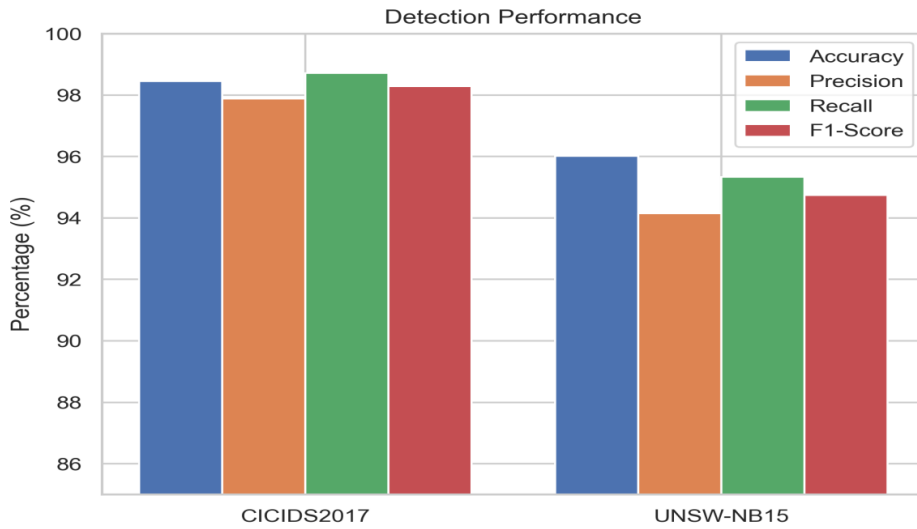


Fig 2: Performance Comparison of Different Data Sets

The performance indicators including accuracy, precision, recall and F1-score demonstrated in Figure 2 and portrayed that proposed framework had high detection efficacy. The framework achieved balanced and robust

results with marked good generalization to myriad patterns of attacks across datasets.

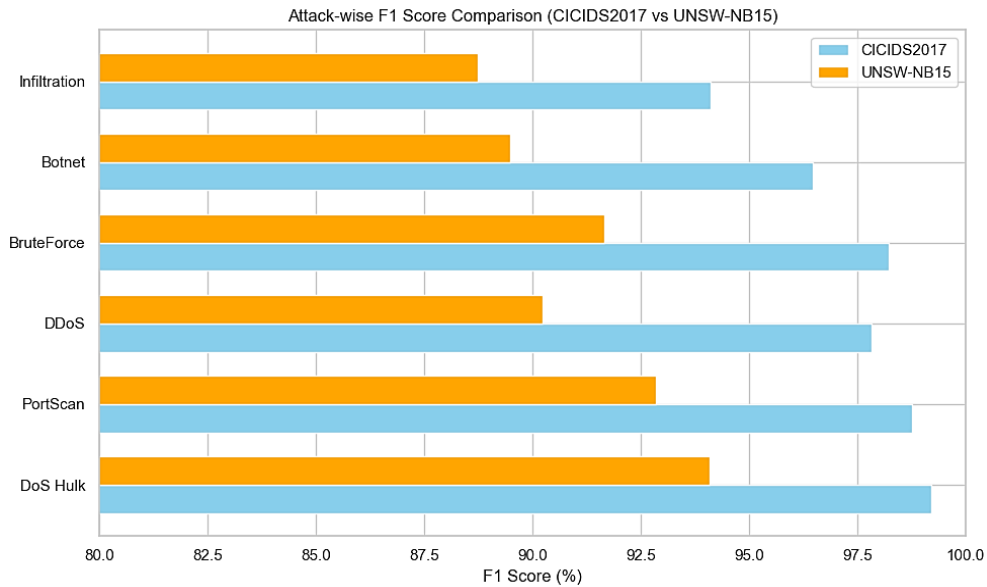


Fig 3: Attack Wise Comparison over various datasets

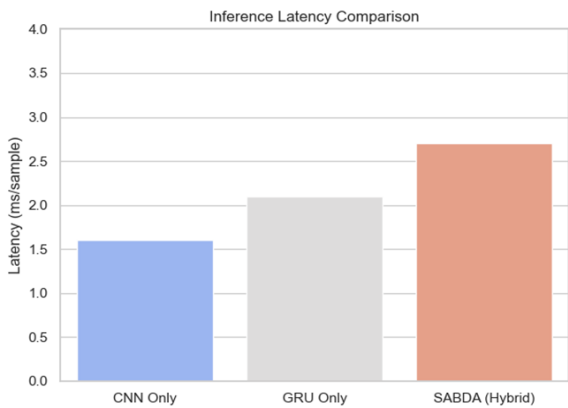


Fig 4: Comparison of Inference Latency

An evaluation using rank of alert analysis was conducted in Figure 5 in order to determine how useful the threat scoring and prioritization component. The system had

the ability to detect and categorize serious threats in the top-N alerts. This confirms that proposed can bring out the most vehement threats at their initial stage thus allows the fastest response of analysts and lowest mean-time-to-detect (MTTD) in security operations.

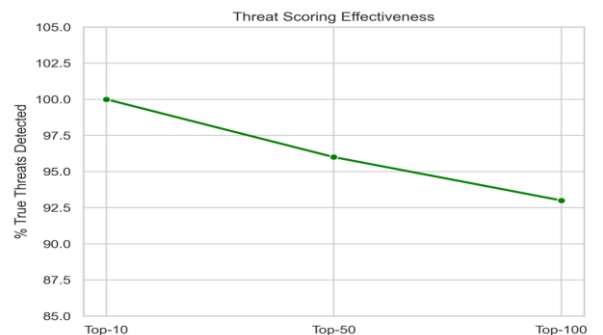


Fig 5: Scoring of Threats

Figure 6 represents the possibility of scaling the framework was tested by calculating streaming detections with different batches. Time taken in the processing increased proportionately to the size of the batch, but the accuracy in detecting the option was the same. This depicts the capacity of Scalable big data based architecture to perform well at high levels of data load, a major strength of big data-based cybersecurity systems.

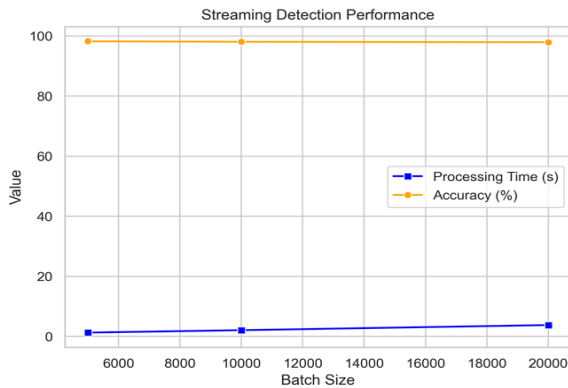


Fig 6: Performance Comparison of Streaming Detection

The proposed Scalable Big Data Based Attack Detection framework presents a high level of detection, responsiveness in real-time, intelligent threat priorities and scalability. Such features qualify it as a promising candidate to use in contemporary enterprises and cloud-based security infrastructures.

5. Conclusion and Future Work

This paper introduced a big data driven framework to help tackle the problem of intrusion detection in new cyber security systems. Proposed framework provides a solid model of combining deep learning mechanisms with a scalable architecture to detect a broad range of network attacks within real-time processing framework. The experimental analysis on a variety of data proved the stability and flexibility of the framework and its operational efficiency. The important contributions of this paper are the layered architecture which specializes in big data stream ingestion, hybrid deep learning based detection and threat scoring for effective alert order. The generality of the system in multitask settings also supports the value of this system in large network security surveillance. Future research will be directed to expanding the architecture to accommodate options of online learning and constant model adaptation. Moreover, it might be possible to improve privacy and minimize latency, deploying federated or edge computing environments.

Author Contributions: K. Prasanthi, P. Venkata Krishna, and V. Saritha all contributed substantially to the conception, design, development, and validation of this research work. K. Prasanthi led the implementation of the scalable big data-based attack detection framework, including data ingestion, preprocessing, and the development of deep learning models. P. Venkata Krishna provided expertise in distributed computing, guiding the integration of the architecture with Apache Spark and HDFS, and supervised the experimental design. V. Saritha contributed to the evaluation process by conducting experiments on benchmark datasets, analyzing results, and assisting in performance comparison and visualization. All

authors collaboratively drafted, reviewed, and approved the final manuscript, and each played an active role in shaping the research findings and conclusions.

Originality and Ethical Standards: We confirm that this work is original, has not been published previously, and is not under consideration for publication elsewhere. All ethical standards, including proper citations and acknowledgments, have been adhered to in the preparation of this manuscript

Data availability: Data available upon request.

Conflict of Interest: There is no conflict of Interest.

Ethical Statement: This research was conducted in accordance with ethical guidelines. Necessary approvals were obtained from the relevant ethical committee, and informed consent was secured from all participants. Confidentiality and anonymity were maintained. The authors declare no conflicts of interest and adhered to all applicable ethical standards.

Funding: The research received no external funding.

Similarity checked: Yes.

References

- [1] S. Morgan, "Cybercrime to cost the world \$10.5 trillion annually by 2025," *Cybersecurity Ventures*, 2020.
- [2] M. Roesch, "Snort—lightweight intrusion detection for networks," in Proc. 13th Systems Administration Conf. (LISA), pp. 229–238, 1999.
- [3] H. Hu, Y. Wen, T. S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," *IEEE Access*, vol. 2, pp. 652–687, 2014.
- [4] T. T. Thi, K. Zhang, and M. Li, "XFedHunter: An explainable federated learning framework for advanced persistent threat detection in SDN," *IEEE Transactions on Network and Service Management*, vol. 20, no. 3, pp. 2318–2332, 2023.
- [5] Latif, M. Ali, and R. Khan, "Securing federated learning with intrusion detection systems: A deep learning perspective," *Journal of Cybersecurity Research*, vol. 12, no. 1, pp. 45–58, 2025.
- [6] S. Vinayakumar, K. P. Soman, and P. Poornachandran, "Applying convolutional neural network for network intrusion detection," in Proc. ICACCI, pp. 1222–1228, 2017.
- [7] K. Kim, Y. Kim, and H. Kim, "LSTM-based system-call language modeling and robust ensemble method for designing host-based intrusion detection systems," *IEEE Access*, vol. 7, pp. 162894–162907, 2019.
- [8] X. Yuan, C. Li, and X. Li, "DeepDefense: Identifying DDoS attack via deep learning," in IEEE SMARTCOMP, pp. 1–8, 2017.
- [9] R. Vinayakumar, K. P. Soman, P. Poornachandran, and S. Sathya, "Evaluating deep learning approaches to intrusion detection," in IEEE ICICCS, pp. 141–146, 2018.
- [10] Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and

intrusion traffic characterization,” in Proc. ICISSP, pp. 108–116, 2018.

[11] H. Hindy, D. Brosset, E. Bayne, et al., “A taxonomy of network threats and the effect of current datasets on intrusion detection systems,” *IEEE Access*, vol. 8, pp. 104650–104675, 2020.

[12] K. Prasanthi, K. Sandhya Rani, and P. Venkata Krishna, “BACADA: Big Data Architecture for Cyber Security Attack Detection Applications,” *African Journal of Biological Sciences*, vol. 6, 2024

[13] C. Yin, Y. Zhu, S. Liu, and J. Fei, “An enhanced capsule network for intrusion detection,” *IEEE Access*, vol. 7, pp. 49699–49710, 2019.

[14] H. Xiao, R. Li, Y. Li, and H. Wang, “Efficient detection of DDoS attacks with ensemble learning,” *IEEE Access*, vol. 9, pp. 47636–47644, 2021.

[15] S. Shone, V. N. Ngoc, and Q. Phan, “A deep learning approach to network intrusion detection,” *IEEE Trans. Emerging Topics in Computational Intelligence*, vol. 2, no. 1, pp. 41–50, 2018.

[16] Y. Zhang, L. Wu, F. Guo, and C. Wang, “A deep learning-based framework for network intrusion detection,” *IEEE Access*, vol. 6, pp. 29085–29092, 2018.

[17] H. Gao, B. Liu, J. Li, and L. Zhang, “An attention-based LSTM-CNN hybrid model for anomaly detection in network traffic,” *IEEE Access*, vol. 9, pp. 106650–106660, 2021.

[18] F. Roy, K. Verma, and P. K. Gupta, “A comprehensive survey on deep learning-based methods for cybersecurity,” *ACM Computing Surveys*, vol. 55, no. 4, pp. 1–36, 2023.

[19] Y. Chen, M. Zhao, and W. Pan, “EdgeAI-NIDS: A lightweight edge computing-based deep learning approach for intrusion detection,” *IEEE Internet of Things Journal*, vol. 10, no. 1, pp. 710–721, 2023.

[20] J. L. Hernandez-Ramos, M. Fernandez, and A. Skarmeta, “Federated learning for anomaly detection in industrial control systems: A cybersecurity perspective,” *Computers & Security*, vol. 125, p. 102958, 2023.

[21] W. Lin, C. Yang, L. Zhang, and C. Xu, “A temporal convolutional network-based approach for intelligent cyber threat detection,” *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 2386–2399, 2021.

[22] M. Schmitt, “AI-enabled malware and intrusion detection for smart infrastructures,” *Journal of Digital Security*, vol. 5, no. 2, pp. 83–96, 2024.