Deep Learning-Driven Enhancement of IoT-Integrated Big Data Analytics in Edge-Cloud Architectures

Priyanka Patani
Apollo Institute of Engineering and Technology
Dr. Sanjay Gour
Gandhinagar University.

1. Introduction

The Internet of Things (IoT) has revolutionized how devices interact and communicate. IoT enables smarter devices to connect with each other and the Internet [1], allowing for remote organization, management, and control of these entities through unique digital identifiers [2]. This has led to an expansion of creative, inventive, and intelligent applications, including automated healthcare, autonomous vehicles, intelligent farming, crowd monitoring, and crowdsourcing [3-5]. Additionally, edge computing has arisen as a means to bring processing closer to the network edge and data sources [6]. It enables IoT devices to collect environmental data and facilitates crucial data analysis for implementing smart systems [7], as shown in Figure 1. Edge computing significantly reduces network traffic and latency by moving computation from centralized systems to the network periphery [8].

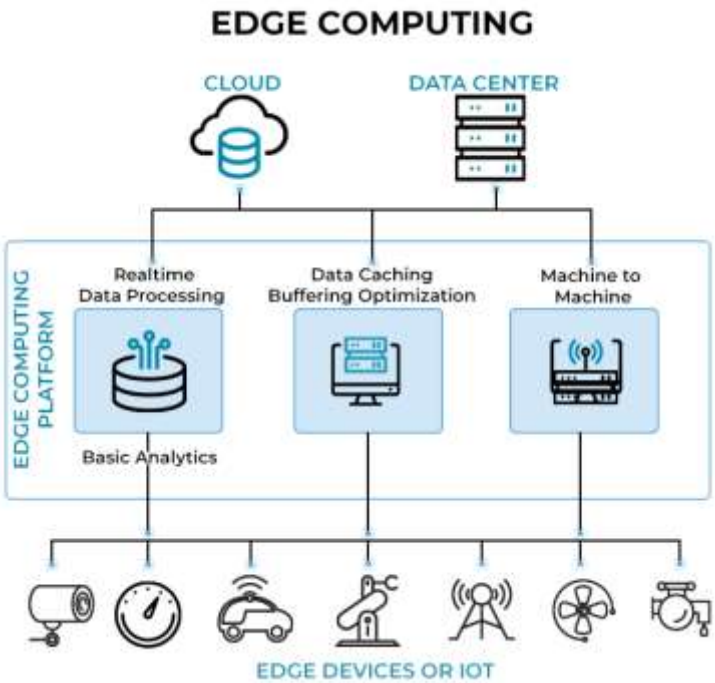

Figure 1. Edge Computing Architecture for IoT devices [9].

Despite these advancements, IoT-enabled big data analytics face several challenges, particularly in edge-cloud computing architectures. These include poor reaction times, high bandwidth costs, excessive energy consumption, and inadequate security [10]. Privacy concerns also arise when data is uploaded to the cloud,

making edge processing a potentially more secure option. Additionally, current methods often fail to keep pace with technological advancements, creating a demand for improved computing techniques [11].

Traditional approaches to address these challenges include cloud computing and edge computing. Cloud computing offers scalability, availability, and storage capabilities, making it ideal for data analytics and Artificial Intelligence (AI) training. However, it suffers from increased latency and requires constant maintenance. Edge computing, on the other hand, provides faster processing and improved privacy protection but lacks the scalability and redundancy of cloud systems [12]. The choice between these approaches often depends on specific use cases and requirements. Hence, to enhance IoT-integrated big data analytics in edge-cloud architectures, this study proposes deep learning-driven methods for anomaly detection.

Deep Learning has become increasingly relevant for big data analytics because of its ability to analyse both labelled and unlabeled data [13]. It can build hierarchical representations from lower-level inputs, enabling effective analysis of various data types, including text, images, and videos [14]. The proposed approach aims to handle huge real-time data, addressing the challenges posed by numerous real-world data sources. By applying deep learning techniques to a cleaned IoT Botnet Dataset, the study seeks to uncover patterns and representations that can improve network attack and anomaly detection. The study focuses on developing a framework to optimize IoT-enabled big data analytics using deep learning, with particular emphasis on security mechanisms for secure and real-time data transmission and processing in distributed edge-cloud architectures. Additionally, the approach aims to minimize latency and energy consumption in hybrid cloud and edge computing configurations.

Hence, this study seeks to overcome the drawbacks of traditional methods and improve the overall performance of IoT-integrated big data analytics in edge-cloud architectures, particularly in the realm of anomaly detection by leveraging deep learning's capabilities.

2. Literature review

Xin et al., (2024) [15] examined recognizing anomalies in IoT device data by employing Variational Autoencoders (VAE) and Convolutional Neural Networks (CNN) to improve security threat identification in IoT applications. The study aimed to enhance the detection and classification abilities of these models by adjusting hardware layouts, software settings, and hyperparameters. The CNN model attained a robust classification performance, with an efficiency of 95.8%, accurately differentiating among all different kinds of IoT device traffic. The VAE model demonstrated effectiveness in identifying anomalies through reconstructing loss and KL divergence, successfully recording unusual trends in the data. The integrated CNN and VAE techniques offered a thorough fix to IoT cybersecurity issues.

Akash et al., (2022) [16] focused on addressing the security challenges in IoT by developing a model for botnet detection using Machine Learning (ML) techniques. The study specifically examined anomalies, or botnets, in clusters of IoT devices trying to associate with a network, utilizing transport layer data generated by these devices. The proposed model combined Independent Component Analysis (ICA) with a Random Forest (RF) classifier for detecting botnets, offering a novel and efficient solution. The experimental results

demonstrated the performance across 3 diverse datasets, attaining up to 99.9% accuracy with the fastest estimation duration of 0.12s, highlighting the model's effectiveness in detecting botnets in IoT environments.

Ahmad et al., (2021) [17] addressed the growing security concerns in IoT networks by proposing an effective anomaly recognition method using mutual data together with a Deep Neural Network (DNN). The study proposed a Network-Based Intrusion Detection System (NIDS) for IoT environments to continuously observe network traffic and reduce False Alarm Rates (FAR) in detecting anomalies. A comparative analysis of various deep learning models was conducted using the IoT-Botnet 2020 dataset. The results showed a 0.57–2.6% improvement in model accuracy and a reduction in FAR by 0.23–7.98%, highlighting the DNN model's superior performance.

Hussain et al., (2020) [18] focused on the security vulnerabilities of IoT devices, which can be exploited to form large botnets. The study analysed that existing botnet detection techniques tend to perform poorly when applied to datasets other than the one they were trained on, due to the variety of attack approaches. To address this issue, the study presented a universal feature set for training ML models that would enhance botnet detection across multiple datasets. The proposed feature set demonstrated superior results when applied to three different botnet attack datasets, effectively improving detection performance regardless of the dataset used.

Hasan et al., (2019) [19] analysed the growing issue of various types of attacks in IoT infrastructures. Thus, the study aimed to address various types of threats and other anomalies that can lead to system failures. The study compared the performance of various ML models in accurately predicting these attacks. The results showed that all the Artificial Neural Network, Decision Tree and RF models achieved an accuracy of 99.40%. Additionally, other performance metrics indicated that RF outperformed the others, making it the most effective model in IoT systems.

3. Research Gaps

- Lack of exploration in integrating advanced feature selection techniques to enhance the feature engineering process for IoT botnet detection [15].
- Insufficient investigation of a hybrid deep learning classification model for IoT botnet detection that improves robustness across diverse datasets and attack patterns [18].
- Limited comprehensive performance evaluation metrics beyond accuracy in the analysis of deep learning models for IoT botnet detection [17].

4. Research Objectives

- To develop a robust IoT botnet detection system by leveraging a hybrid deep learning classification model in edge-cloud architecture
- To enhance the feature engineering process for botnet detection using Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE), ensuring only the most significant attributes are retained for analysis.

- To evaluate the performance of the proposed hybrid model through comprehensive performance analysis metrics, validating its effectiveness in detecting IoT botnet attacks.

5. Research Questions

RQ1: How effective is the hybrid model in detecting IoT botnet attacks compared to individual algorithms?

RQ2: Can the combination of PCA and RFE improve feature selection in IoT botnet detection?

RQ3: What performance metrics can best evaluate the efficiency of the proposed model in detecting IoT botnet attacks?

6. Research Hypothesis

H1: The hybrid model will significantly enhance the performance of IoT botnet detection compared to individual ML algorithms.

H2: The use of PCA and RFE will result in a more refined feature set, leading to better classification performance in identifying IoT botnet attacks.

H3: The proposed model will outperform traditional botnet detection systems when evaluated on various performance evaluation parameters.

7. Methodology

This section outlines the dataset and techniques used in developing a robust IoT botnet detection system using a hybrid deep learning model.

7.1 Dataset Description

The study utilized the publicly available "IoT Botnet Dataset," which contains extensive internet transactional data with numerous attributes for each transaction. Key features include unique identifiers, timestamps, flow state flags, protocol details, source and destination IP addresses and ports, packet and byte counts, transaction states, and duration metrics. Additionally, the dataset provides statistics on incoming connections by IP address, average rates per protocol and IP, categorical traffic information, and a class label (Attack) indicating normal (0) or malicious (1) traffic. This comprehensive dataset is primarily used to train and evaluate deep learning models for detecting intrusions in IoT botnets, supporting the development of effective intrusion detection systems [20][21].

7.2 Technique used

The techniques used in this study for feature extraction, feature selection, and classification are given below. These methods are selected for their ability to handle large-scale IoT botnet datasets, capture temporal patterns, and provide robust classification performance.

(i) PCA

It is a dimensionality reduction method that reduces complex datasets to a smaller collection of independent variables. These variables or components retain most of the variance from the original data, capturing its

essential patterns. PCA is employed for feature extraction because it reduces the complexity of the IoT botnet dataset by focusing on the most important features, which helps improve the model's efficiency and accuracy without losing critical information [22].

(ii) RFE

It is used for feature selection to further refine the feature set obtained from PCA. This backward elimination process eliminates the least significant features recursively, ensuring that the model only focuses on the most relevant attributes. RFE improves the model's efficiency and performance by reducing overfitting and ensuring that only impactful features contribute to botnet detection [23].

(iii) Long Short-Term Memory (LSTM)

It is designed to capture long-term dependencies in sequential data by overcoming the vanishing gradient problem. It is used because IoT traffic data has temporal dependencies, and botnet behaviours often evolve over time. LSTM helps the model recognize these long-term patterns, making it well-suited for accurately detecting anomalies and attacks in the dynamic nature of IoT networks [24].

(iv) Gated Recurrent Unit (GRU)

GRU is utilized due to its capability to capture temporal dependencies in the sequential nature of IoT traffic data. Botnets often execute attacks over time, and GRU's architecture allows the model to effectively track patterns across time steps, enhancing the detection of dynamic and evolving botnet activities [25].

(v) CNN

This deep learning technique enhances the classification process by extracting spatial patterns and local features from the data. CNN is particularly effective at capturing intricate details and anomalies within the dataset, making it well-suited for detecting complex botnet behaviours [26]. By complementing the strengths of LSTM and GRU, CNN adds an additional layer of feature extraction, ensuring more accurate and robust IoT botnet detection through its ability to recognize fine-grained patterns in network traffic data.

Each of these techniques is chosen because they collectively address the key challenges in IoT botnet detection: processing large datasets, capturing temporal and spatial patterns and providing robust, efficient, and accurate classification outcomes.

7.3 Proposed approach

The methodology begins with the collection of an IoT botnet dataset, which is subjected to a thorough data preprocessing phase. This involves data cleaning to remove any corrupt or irrelevant entries and normalization to scale the features, ensuring that no individual feature dominates the learning process due to its magnitude. Furthermore, missing values are carefully handled, either by imputation or removal, to ensure a complete and consistent dataset for further analysis. After preprocessing, feature extraction is performed using PCA, it ensures that the most informative features are extracted, enhancing the model's ability to learn from the data without being overwhelmed by irrelevant or redundant features. Once feature extraction is complete, the

dataset undergoes further feature selection using RFE. It refines the feature set, ensuring that only the highly relevant features are retained for classification.

After feature extraction and selection, the dataset is divided into training and testing sets. The classification model used is a hybrid approach combining three powerful Deep Learning methods: LSTM**,** GRU, and CNN. LSTM is used to handle the classification task, effectively capturing long-term dependencies and patterns within the IoT traffic data. The GRU component is incorporated to capture the temporal dependencies in the IoT traffic data, which is important because botnet behaviour often evolves. The CNN component is utilized to extract spatial patterns and local features from the network traffic data, helping to identify intricate patterns or anomalies within the data that are indicative of botnet activities. Finally, the performance of the employed model is evaluated using appropriate performance analysis metrics to validate its ability to detect IoT botnet attacks effectively. The combination of these methods ensures a powerful, accurate, and robust botnet detection system that leverages both sequential and spatial data features for enhanced detection accuracy.

8. Limitations

Despite the strengths of the proposed hybrid model combining LSTM**,** GRU, and CNN for IoT botnet detection, there are several limitations. The complexity of the model, due to the integration of multiple algorithms, may result in higher computational costs and longer training times, particularly with large IoT datasets. Additionally, while GRU captures temporal dependencies and CNN extracts spatial patterns, the model may struggle with handling new, unseen botnet behaviours if the training data does not cover a wide variety of attack types. Moreover, the reliance on feature engineering methods like PCA and RFE, while effective, could miss out on important features if not carefully tuned. Finally, scalability and real-time applicability might be constrained, as deploying such a resource-intensive model on edge devices with limited computational power could be challenging.

9. Conclusion and Future Scope

This study demonstrates the potential of deep learning techniques for optimizing an IoT-enabled big data analytics architecture within an edge-cloud computing environment, focusing on anomaly detection. IoT systems generate huge amounts of data, presenting challenges related to real-time processing, network congestion, and security vulnerabilities. Traditional approaches often fall short due to latency, bandwidth limitations, and insufficient security measures. By leveraging deep learning models this framework captures temporal dependencies, identifies spatial patterns, and enhances classification accuracy, enabling real-time detection of anomalies and network attacks. The integration of edge computing further reduces network traffic and processing delays by moving computation closer to the data source, thus improving efficiency and privacy. This deep learning-driven approach not only enhances anomaly detection but also optimizes resource allocation, reduces latency, and minimizes energy consumption, providing a scalable and secure solution for IoT-enabled big data analytics across various applications, such as smart cities and industrial automation.

Future research could focus on expanding this framework to handle larger and more complex IoT networks, integrating more advanced neural network architectures like transformers to further improve detection

capabilities. Moreover, exploring the application of federated learning could allow for more distributed and privacy-preserving anomaly detection across edge devices. Lastly, optimizing the system for deployment on low-power IoT devices would help make this solution more scalable and energy-efficient for real-world implementations in various industries.

References

[1] Evtodieva, T. E., D. V. Chernova, N. V. Ivanova, and J. Wirth. "The internet of things: possibilities of application in intelligent supply chain management." *Digital transformation of the economy: Challenges, trends and new opportunities* (2019): 395-403.

[2] Piccialli, Francesco, and Jai E. Jung. "Understanding customer experience diffusion on social networking services by big data analytics." *Mobile Networks and Applications* 22 (2017): 605-612.

[3] Baker, Stephanie B., Wei Xiang, and Ian Atkinson. "Internet of things for smart healthcare: Technologies, challenges, and opportunities." *Ieee Access* 5 (2017): 26521-26544.

[4] Babar, Muhammad, Fazlullah Khan, Waseem Iqbal, Abid Yahya, Fahim Arif, Zhiyuan Tan, and Joseph M. Chuma. "A secured data management scheme for smart societies in industrial internet of things environment." *IEEE Access* 6 (2018): 43088-43099.

[5] Lashkari, Bahareh, Javad Rezazadeh, Reza Farahbakhsh, and Kumbesan Sandrasegaran. "Crowdsourcing and sensing for indoor localization in IoT: A review." *IEEE Sensors Journal* 19, no. 7 (2018): 2408-2434.

[6] Ghosh, Ananda Mohon, and Katarina Grolinger. "Edge-cloud computing for Internet of Things data analytics: Embedding intelligence in the edge with deep learning." *IEEE Transactions on Industrial Informatics* 17, no. 3 (2020): 2191-2200.

[7] L'heureux, Alexandra, Katarina Grolinger, Hany F. Elyamany, and Miriam AM Capretz. "Machine learning with big data: Challenges and approaches." *Ieee Access* 5 (2017): 7776-7797.

[8] Roman, Rodrigo, Javier Lopez, and Masahiro Mambo. "Mobile edge computing, fog et al.: A survey and analysis of security threats and challenges." *Future Generation Computer Systems* 78 (2018): 680-698.

[9] https://www.spiceworks.com/tech/cloud/articles/edge-vs-cloud-computing/

[10] Mitra, Anuran, Soumita Biswas, Tinku Adhikari, Arindam Ghosh, Soumalya De, and Raja Karmakar. "Emergence of edge computing: An advancement over cloud and fog." In 2020 11th International Conference on Computing, Communication and Networking Technologies (ICT), pp. 1-7. IEEE, 2020

[11] Satyanarayanan, Mahadev, Guenter Klas, Marco Silva, and Simone Mangiante. "The seminal role of edge-native applications." In *2019 IEEE International Conference on Edge Computing (EDGE)*, pp. 33-40. IEEE, 2019.

[12] Shi, Weisong, Jie Cao, Quan Zhang, Youhuizi Li, and Lanyu Xu. "Edge computing: Vision and challenges." *IEEE internet of things journal* 3, no. 5 (2016): 637-646.

[13] Ravi, Daniele, Charence Wong, Benny Lo, and Guang-Zhong Yang. "A deep learning approach to on-node sensor data analytics for mobile or wearable devices." *IEEE journal of biomedical and health informatics* 21, no. 1 (2016): 56-64.

[14] Tannahill, Barnabas K., Chris E. Maute, Yunus Yetis, Maryam N. Ezell, Aldo Jaimes, Roberto Rosas, Azima Motaghi, Halid Kaplan, and Mo Jamshidi. "Modeling of system of systems via data analytics—Case for "Big Data" in SoS." In *2013 8th International Conference on System of Systems Engineering*, pp. 177-183. IEEE, 2013.

[15] Xin, Qi, Zeqiu Xu, Linfeng Guo, Fanyi Zhao, and Binbin Wu. "IoT traffic classification and anomaly detection method based on deep autoencoders." (2024).

[16] Akash, Nazmus Sakib, Shakir Rouf, Sigma Jahan, Amlan Chowdhury, and Jia Uddin. "Botnet detection in IoT devices using random forest classifier with independent component analysis." *Journal of Information and Communication Technology* 21, no. 2 (2022): 201-232.

[17] Ahmad, Zeeshan, Adnan Shahid Khan, Kashif Nisar, Iram Haider, Rosilah Hassan, Muhammad Reazul Haque, Seleviawati Tarmizi, and Joel JPC Rodrigues. "Anomaly detection using deep neural network for IoT architecture." *Applied Sciences* 11, no. 15 (2021): 7050.

[18] Hussain, Faisal, Syed Ghazanfar Abbas, Ubaid U. Fayyaz, Ghalib A. Shah, Abdullah Toqeer, and Ahmad Ali. "Towards a universal features set for IoT botnet attacks detection." In *2020 IEEE 23rd international multitopic conference (INMIC)*, pp. 1-6. IEEE, 2020.

[19] Hasan, Mahmudul, Md Milon Islam, Md Ishrak Islam Zarif, and M. M. A. Hashem. "Attack and anomaly detection in IoT sensors in IoT sites using machine learning approaches." *Internet of Things* 7 (2019): 100059.

[20] Ullah, Imtiaz, and Qusay H. Mahmoud. "A technique for generating a botnet dataset for anomalous activity detection in IoT networks." In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 134-140. IEEE, 2020.

[21] https://sites.google.com/view/iotbotnetdatset/home

[22] Haq, Mohd Anul, and Mohd Abdul Rahim Khan. "DNNBoT: Deep neural network-based botnet detection and classification." *Computers, Materials & Continua* 71, no. 1 (2022).

[23] Pektaş, Abdurrahman, and T. Acarman. "Effective feature selection for botnet detection based on network flow analysis." In *International Conference Automatics and Informatics*, pp. 1-4. 2017.

[24] Kim, Jiyeon, Hyerin Won, Minsun Shim, Seungah Hong, and Eunjung Choi. "Feature analysis of iot botnet attacks based on RNN and LSTM." Int J Eng Trends Technol 68, no. 4 (2020): 43-47.

[25] Ur Rehman, Saif, Mubashir Khaliq, Syed Ibrahim Imtiaz, Aamir Rasool, Muhammad Shafiq, Abdul Rehman Javed, Zunera Jalil, and Ali Kashif Bashir. "DIDDOS: An approach for detection and identification of Distributed Denial of Service (DDoS) cyberattacks using Gated Recurrent Units (GRU)." *Future Generation Computer Systems* 118 (2021): 453-466.

[26] Bajao, Naomi A., and Jae-an Sarucam. "Threats Detection in the Internet of Things Using Convolutional neural networks, long short-term memory, and gated recurrent units." *Mesopotamian journal of cybersecurity* 2023 (2023): 22-29.