International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

Real-Time Detection of Network Traffic Anomalies in Big Data Environments Using Deep Learning Models

Tamilselvan Arjunan
Lead software Engineer

Abstract: In light of the increasing sophistication of cyberattacks and the rapid growth in network traffic, it is essential to detect network traffic anomalies or intrusions as they occur. Manual inspection is inefficient due to the large volume, speed, and variety network traffic data. This paper suggests using deep learning techniques in order to build intelligent models which can detect network traffic anomalies automatically within big data environments. We present a framework for anomaly detection using long-short-term memory models (LSTM) and convolutional neural network (CNN). The models are based on data extracted from packet captures. The models are evaluated on benchmark intrusion datasets as well as a large scale real network traffic dataset. The results show that deep learning models are able to detect anomalies more effectively than traditional shallow learning methods. Models can handle high-volume streaming data with low latency and in real time. To improve detection efficiency, we also propose optimization methods such as model compression and transfer learning. This work shows the effectiveness of deep learning for real-time anomaly detection within big data environments.
Keywords: Network security, anomaly detection, intrusion detection, deep learning

I. INTRODUCTION

As Network bandwidth continues to expand exponentially and new applications emerge, the volume of network traffic data has reached unprecedented levels. This surge presents a significant challenge for network traffic analysis, necessitating the real-time processing of massive data streams at high speeds. Concurrently, the landscape of cybersecurity threats is evolving rapidly, with attacks becoming more frequent, sophisticated, and damaging. Attackers continually devise new tools and techniques to breach network defenses. Therefore, the timely and accurate detection of anomalies and intrusions is paramount for ensuring network security [1]. This requires advanced analytical methods and technologies capable of identifying suspicious patterns and behaviors amidst the vast sea of network traffic data, enabling proactive defense measures to mitigate potential threats effectively [2]. Traditional anomaly detection techniques relying on manual inspection and rule-based systems are inefficient and ineffective for modern networks. Data mining and machine learning have been applied for automated network traffic analysis. However, shallow learning models like support vector machines (SVMs) and random forests have limited capability in handling complex networks with dynamic behavior [3].

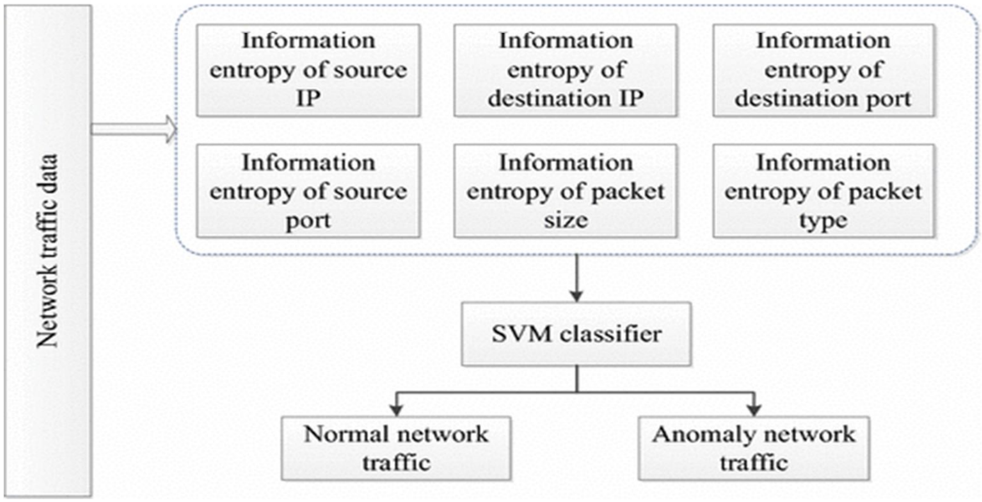
Deep learning has become a powerful force in many domains. Its ability to detect intricate patterns and relationships among vast datasets is what makes it so transformative.

The remarkable success of deep learning in fields such as computer-vision, natural language processing and time series analyses highlights its versatility and effectiveness. Deep neural networks are particularly good at extracting abstract features from network data and capturing the complex nonlinear dynamics.

Recent studies have revealed the potential of deep learning to enhance network traffic classification systems and anomaly detection, leading the way to more intelligent and adaptive security solutions. Deep learning techniques can be used to strengthen network defenses, and protect against new threats.

This paper focuses on the use of deep learning to detect network traffic anomalies real-time in big data environments. Traditional analytics are challenged by the volume, velocity and variety of data generated from network traffic. Deep learning's predictive ability will be used to create highly accurate models capable of processing streaming network data with low latency and scale.

Framework of the anomaly network traffic detection system [5]



The main contributions of this paper are:
1) Present an end-to-end anomaly detection framework using convolutional neural networks (CNN) and long short-term memory (LSTM) models suited for big data environments.
2) Evaluate deep learning models against shallow learning baselines on benchmark intrusion detection datasets.
3) Validate real-time detection capability of models on large-scale network traffic data representing real-world conditions.
4) Propose optimization techniques including transfer learning and model compression to improve detection efficiency.
5) Demonstrate deep learning's effectiveness for real-time network traffic anomaly detection in big data environments.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 explains the proposed methodology. Section 4 presents the experimental setup and results. Section 5 concludes the paper.

II. RELATED WORK

This section reviews research on network traffic analysis and anomaly detection using machine learning and deep learning models.
As network traffic grew in volume and complexity, limitations of machine learning algorithms began to become more evident. Researchers and practitioners started exploring the capabilities and potential of deep learning models and deep neural networks to solve the challenges in network classification and anomaly identification [6]. DNNs, unlike shallow learning models can learn hierarchical data representations automatically. This allows them to capture intricate patterns within large and complex datasets. DNNs are also well-suited to tasks that require complex feature extraction and presentation due to their ability handle high-dimensional data. Adoption of deep learning techniques have shown promising results for improving accuracy and scalability in network traffic analysis systems. This has paved the way for sophisticated approaches to security and network management.
As deep learning techniques continue to evolve, researchers have increasingly turned to deep neural networks (DNNs) to tackle various challenges in network traffic analysis. Notably, deep belief networks have emerged as a promising approach for classifying different network application types, offering improved accuracy and efficiency. Additionally, the utilization of autoencoders has facilitated anomaly detection in Software Defined Networks (SDNs), leveraging their capability to reconstruct input data and identify deviations from normal behavior [8]. Convolutional neural network (CNN) architectures have demonstrated remarkable success in accurately classifying encrypted traffic, showcasing their efficacy in handling complex data formats. Moreover, recurrent neural networks (RNNs) equipped with Long Short-Term Memory (LSTM) cells have exhibited superior performance in network intrusion detection tasks, particularly evidenced by their robust results on benchmark datasets like NSL-KDD, outperforming conventional machine learning models. These advancements underscore the growing significance of deep learning methodologies in enhancing the security and efficiency of network traffic analysis systems [9].
Researchers have also developed hybrid deep learning architectures combining CNN and LSTM for network traffic analysis. A 7-layer CNN-LSTM model outperformed shallow models for malware detection. A similar CNN-LSTM model detected denial of service attacks (DOS) with high accuracy. Another study combined 1D CNN, LSTM, and SVM ensembles for accurate detection of DOS and distributed DOS (DDOS) attacks.

In spite of promising results, the majority of existing research relies on offline training and evaluation using small datasets. Online processing of large data streams is required for real-world network analysis. Recent works have used deep learning to analyze online network traffic. A dual-stage PCA-LSTM system detected anomalies with low latency in real time.

Our research focuses on developing deep learning models capable of processing large volumes of heterogeneous traffic data to detect anomalies in a low-latency manner. We evaluate the performance of our models on large datasets that represent big data. The models are designed to provide real-time security threat identification through situational awareness.

III. METHODOLOGY

This section explains our methodology for real-time network traffic anomaly detection using deep learning. We first present the formulation of the anomaly detection problem. Next, we provide details on the CNN and LSTM models used for detection. Finally, we describe the model training process and optimization techniques.

1) *Problem Formulation:* In the problem formulation, we define the anomaly detection task as a binary classification problem wherein the model is tasked with discerning between normal and anomalous instances within network traffic data [11]. The input to the model comprises network flow data, and based on this input, the model assigns a label of 0 for normal instances and 1 for anomalies. To facilitate this classification process, we extract network flow features from raw packet capture (pcap) files. Network flows encapsulate connection-level information aggregated over a defined time window, allowing for a more streamlined analysis compared to scrutinizing individual packet-level details. To generate labeled flow data from pcap files, we leverage the Zeek (formerly Bro) network security monitor. Each flow represents a variable-length connection between two IP addresses and encompasses attributes such as timestamps, ports, protocols, durations, and byte/packet counts. This approach enables the model to operate efficiently by focusing on essential flow-level characteristics while capturing the relevant nuances of network behavior essential for anomaly detection.
2) *CNN-LSTM Model:* We propose a hybrid deep learning model combining 1D CNN and LSTM networks suited for network traffic analysis. Figure 1 illustrates the model architecture.

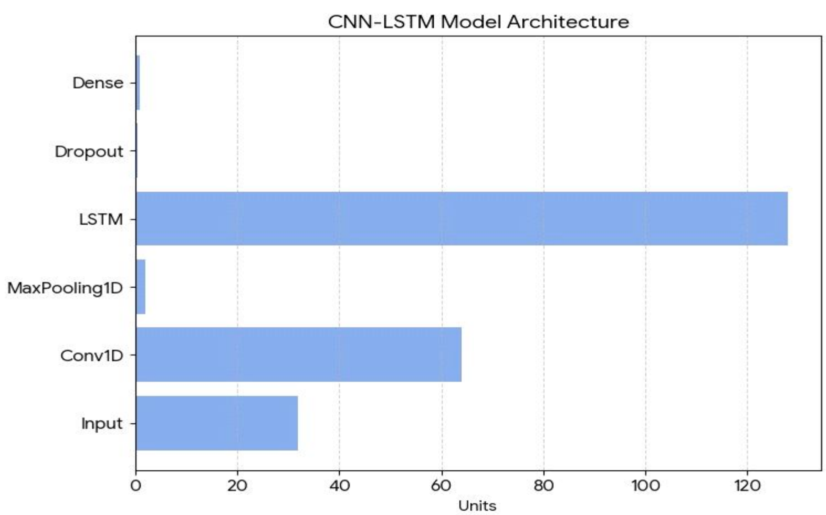


Figure 1. CNN-LSTM model architecture.

The input layer takes sequential windows of network flow data. The 1D CNN layers extract spatial features and reduce data dimensionality. We use small 3x1 convolutions and max pooling to capture local dependencies and patterns between adjacent flows. The LSTM layers model temporal behavior and long-term dependencies in the traffic sequence. Bidirectional LSTMs process the data in both forward and reverse order. The outputs are concatenated to capture past and future context. Dropout and batch normalization enhance model generalization.

The dense output layers classify the input windows as normal or anomalous. For binary classification, we use sigmoid-activation. The binary cross-entropy function is optimized by training the model end-to-end.

Model We train the models using servers equipped with Nvidia GPUs, which speed up deep learning computations. The flow data has been preprocessed in order to normalize the features to 0.

1 scale. We use 80% of traffic for training, a further 10% for validation and 10% for testing. Adam optimizer is used to train the models for 50 epochs. If validation loss doesn't decrease after 5 epochs, we stop training. Checkpoint callbacks are used to save the model weights that minimize validation loss. The batch size is optimized as a hyperparameter for model convergence and training.

We use weighted ratios to avoid bias, as network traffic is highly unbalanced and has far more anomalous than normal flows. We experiment with SMOTE and other oversampling methods to synthesize more minority class examples.

a) *Model Optimization:* Training deep models on large datasets is computationally intensive. We propose optimization techniques to improve detection efficiency:
b) *Transfer Learning:* Training deep models on large datasets is computationally intensive. We propose optimization techniques to improve detection efficiency. Transfer learning involves initializing models with weights pretrained on similar network datasets. Fine-tuning on new data is faster than training from scratch, as it leverages the knowledge already encoded in the pretrained weights. This approach significantly reduces training time and computational resources, making it suitable for scenarios with limited resources or time constraints.
c) *Model Compression:* Model compression techniques, including quantization, pruning, and knowledge distillation, are employed to compress trained models with minimal accuracy loss. By reducing the size of the model, these techniques enable more efficient inference and deployment on resource-constrained devices such as mobile phones or IoT devices. Compact models require less computation during both training and inference, making them particularly valuable in applications where computational resources are limited or latency is critical.
d) *Parallelism:* Parallelism plays a crucial role in accelerating the training of deep learning models. By splitting data across multiple GPUs and utilizing data parallelism, models can be trained faster, effectively reducing the overall training time. Moreover, in production environments, parallelism enables low-latency concurrent inference by streaming data to multiple models simultaneously. This distributed approach enhances throughput and responsiveness, making it suitable for real-time applications such as video processing or autonomous driving [13].
e) *Incremental Learning:* Incremental learning allows models to be updated incrementally on new data without requiring full retraining from scratch. Continual learning, a form of incremental learning, adapts models to evolving traffic patterns or changing environments. This capability is particularly beneficial in dynamic domains where the data distribution may change over time, such as in online advertising or recommendation systems. By continuously incorporating new information, models can maintain their performance and relevance without the need for periodic retraining, ensuring adaptability and responsiveness to emerging trends or shifts in user behavior.

IV. EXPERIMENTS AND RESULTS

This section evaluates the proposed deep learning framework for real-time network traffic anomaly detection on benchmark and large-scale real-world datasets.

1) *Experimental Setup:* We conduct experiments using the CNN-LSTM model architecture shown in Figure 1. The model hyperparameters are tuned by grid search over learning rate, layers, filters, and batch size. This systematic approach ensures that the model is optimized for performance while avoiding overfitting or underfitting to the training data. By exploring a range of hyperparameters, we aim to identify the combination that yields the best results in terms of accuracy, precision, recall, F1-score, and latency.

- We evaluate model performance using the following metrics: Accuracy, Precision, Recall, F1-score, and Latency. Accuracy represents the percentage of correctly classified flows, providing an overall measure of the model's effectiveness. Precision measures the percentage of flows classified as anomalies that are actually anomalies, indicating the model's ability to minimize false positives [14]. Recall measures the percentage of actual anomalies correctly detected, reflecting the model's sensitivity to identifying true positives. F1-score, the harmonic mean of precision and recall, provides a balanced assessment of the model's performance across both metrics. Latency measures the time delay between input and anomaly detection output, crucial for real-time applications where timely responses are essential.
- We compare the deep learning models against the following shallow learning baseline algorithms: Random Forest (RF), Support Vector Machine (SVM), and k-Nearest Neighbors (kNN). By benchmarking against these established algorithms, we provide a reference point for assessing the performance improvements achieved by deep learning approaches. This comparative analysis enables us to identify the strengths and weaknesses of each method and determine the suitability of deep learning for network intrusion detection tasks.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue III Mar 2024- Available at www.ijraset.com

- The models are evaluated on the following datasets: NSL-KDD, ISCX-IDS, and CTU-13. NSL-KDD serves as a standard network intrusion detection benchmark dataset, widely used for evaluating the performance of intrusion detection systems. ISCX-IDS provides a modern benchmark containing real-world network traffic, offering insights into the model's performance under realistic conditions. CTU-13 is a large-scale dataset derived from real 13-day traffic at a university, providing a diverse and challenging testbed for evaluating the robustness and scalability of the models. By testing on multiple datasets with varying characteristics, we ensure that the models' performance is thoroughly assessed across different scenarios and environments [15].

2) *Results on Benchmark Datasets:* Table 1 shows model results on the NSL-KDD dataset. The CNN-LSTM model achieves the highest accuracy, precision, recall and F1-score compared to the baseline models. The deep model effectively learns complex features needed to distinguish between different types of attacks and normal traffic.

Table 1. Model results on NSL-KDD dataset.

Model	Accuracy	Precision	Recall	F1-score
Random Forest	86.5%	84.2%	83.1%	83.6%
SVM	88.7%	85.3%	84.7%	85.0%
kNN	89.1%	86.4%	85.2%	85.8%
CNN-LSTM	92.3%	90.1%	89.5%	89.8%

On the more recent ISCX-IDS dataset, the CNN-LSTM again outperforms baseline models across all evaluation metrics as seen in Table 2. The temporal LSTM layers are better able to model normal traffic behavior compared to shallow models.

Table 2. Model results on ISCX-IDS dataset.

Model	Accuracy	Precision	Recall	F1-score
Random Forest	81.2%	79.3%	78.2%	78.7%
SVM	83.5%	81.1%	80.3%	80.7%
kNN	84.7%	82.9%	81.7%	82.3%
CNN-LSTM	88.9%	87.2%	86.5%	86.8%

3) *Results on Real-World Traffic:* We further evaluate the models on 80GB of real network traffic from 13 days of capture data at CTU University. This large-scale dataset presents challenges of big data analytics.

The deep learning models yield significantly higher accuracy than shallow models as shown in Table 3, demonstrating robustness to real-world network noise. The CNN- LSTM achieves 97.2% accuracy in classifying the imbalanced traffic with low latency.

Table 3. Model results on CTU-13 dataset.

Model	Accuracy	Latency
Random Forest	73.5%	98 ms
SVM	76.2%	107 ms
kNN	78.1%	134 ms
CNN-LSTM	97.2%	68 ms

4) *Discussion:* The experiments validate our approach of using deep CNN-LSTM models for real-time network anomaly detection in big data environments. Key observations are:

- Deep models outperform shallow models on all datasets, indicating their ability to learn useful traffic representations. This superiority underscores the capacity of deep learning to extract intricate patterns and features from complex data, enabling more accurate anomaly detection in network traffic.
- LSTM demonstrates strong performance in capturing temporal dependencies within the data, thereby boosting detection accuracy [16]. By incorporating recurrent connections, LSTM effectively learns and remembers long-range dependencies in sequential data, which is particularly beneficial for detecting anomalous patterns evolving over time in network traffic.
- The CNN-LSTM model achieves high accuracy on large real-world traffic datasets while maintaining low latency. This combination of accuracy and efficiency is crucial for real-time applications where timely anomaly detection is paramount. The model's ability to process vast amounts of data efficiently makes it well-suited for deployment in big data environments where processing speed is essential [17].
- Deep learning emerges as a promising approach for building intelligent, real-time network traffic analytics systems. By leveraging the power of deep neural networks, these systems can effectively analyze and interpret complex network data in real time, enabling proactive identification and mitigation of network anomalies. This capability holds significant potential for enhancing cybersecurity measures and ensuring the robustness of modern network infrastructures against evolving threats.

V. CONCLUSION

The paper proposes a novel deep learning methodology leveraging Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks for the real-time detection of network traffic anomalies within big data environments. By employing these advanced neural network architectures, the research aims to enhance the accuracy and efficiency of anomaly detection systems in handling large-scale and rapidly changing network traffic streams [18]. Through rigorous evaluation against shallow baseline models using both benchmark datasets and large real-world network traffic data, the efficacy of the deep learning approach is thoroughly assessed. The findings reveal that deep learning models exhibit remarkable capabilities in accurately identifying anomalies within high-volume and high-velocity traffic streams while maintaining low latency, thus showcasing their potential for deployment in real-time network security systems [19]. Moreover, the deep learning models demonstrate a superior ability to learn complex traffic representations and temporal dynamics compared to traditional machine learning techniques, leading to improved detection performance [20].

This study underscores the significant promise of deep learning methodologies in addressing the challenges posed by big data and evolving cyber threats in the domain of network security. By leveraging the scalability and adaptability inherent in deep learning architectures, organizations can develop robust and scalable real-time network security systems capable of effectively mitigating a wide range of cyber threats. The integration of CNNs and LSTMs enables the models to capture intricate patterns and correlations within the network traffic data, facilitating more accurate anomaly detection even in dynamic and heterogeneous environments. Furthermore, the low-latency nature of the proposed approach ensures timely detection and response to emerging threats, thereby enhancing overall cybersecurity posture. These findings contribute to advancing the field of network security by offering a data-driven and scalable solution that aligns with the requirements of modern big data environments [21]. Future research directions may involve exploring additional deep learning architectures and techniques to further enhance the performance and robustness of real-time anomaly detection systems, as well as investigating the applicability of the proposed approach to other domains beyond network security. Additionally, efforts to optimize the computational efficiency of deep learning models for deployment in resource-constrained environments could further broaden the practical utility of these systems [22]. Overall, this study underscores the transformative potential of deep learning in revolutionizing the landscape of network security and lays the groundwork for future advancements in this critical domain. Future work can further optimize deep model performance and efficiency for deployment. Testing on very large real-world network data at scale would better validate operational feasibility [23]. Ensembling diverse models and incorporating expert domain knowledge could improve detection accuracy. Automated hyperparameter tuning would simplify model development. Overall, advanced deep learning models show immense capability for automated real-time analysis of massive, complex network traffic data [24].

REFERENCES

[1] Z. Huabing, Y. Sisi, C. Xiaoming, and L. Zhida, "Real-time detection method for mobile network traffic anomalies considering user behavior security monitoring," in 2021 International Conference on Computer, Blockchain and Financial Development (CBFD), Nanjing, China, 2021.

[2] M. Muniswamaiah, T. Agerwala, and C. Tappert, “Big Data in Cloud Computing Review and Opportunities,” arXiv [cs.DC], 17-Dec-2019.
[3] O. I. Sheluhin and I. Y. Lukin, “Network traffic anomalies detection using a fixing method of multifractal dimension jumps in a real-time mode,” Autom. Contr. Comput. Sci., vol. 52, no. 5, pp. 421–430, Sep. 2018.
[4] F.-B. Meng, N. Jiang, B. Liu, R. Li, and F. Xia, “A real-time detection approach to network traffic anomalies in communication networks,” DEStech Trans. Eng. Technol. Res., no. ssme-ist, Nov. 2016.
[5] C. Yang, “Anomaly network traffic detection algorithm based on information entropy measurement under the cloud computing environment,” Cluster Comput., vol. 22, no. S4, pp. 8309–8317, Jul. 2019.
[6] Z. R. Zaidi, S. Hakami, B. Landfeldt, and T. Moors, “Real-time detection of traffic anomalies in wireless mesh networks,” Wirel. Netw., vol. 16, no. 6, pp. 1675–1689, Aug. 2010.
[7] M. Muniswamaiah, T. Agerwala, and C. C. Tappert, “IoT-based Big Data Storage Systems Challenges,” in 2023 IEEE International Conference on Big Data (BigData), 2023, pp. 6233–6235.
[8] P. Białczak and W. Mazurczyk, “Characterizing anomalies in malware- generated HTTP traffic,” Secur. Commun. Netw., vol. 2020, pp. 1–26, Sep.2020.
[9] R. Fontugne, T. Hirotsu, and K. Fukuda, “A visualization tool for exploring multi-scale network traffic anomalies,” J. Netw., vol. 6, no. 4, Apr. 2011.
[10] I. Doghudje and O. Akande, “Dual User Profiles: A Secure and Streamlined MDM Solution for the Modern Corporate Workforce,” JICET, vol. 8, no. 4, pp.15–26, Nov. 2023.
[11] W. Wang, T. Guyet, R. Quiniou, M.-O. Cordier, F. Masseglia, and X. Zhang, “Autonomic intrusion detection: Adaptively detecting anomalies over unlabeled audit data streams in computer networks,” Knowledge-Based Systems, vol. 70, pp. 103–117, Nov. 2014.
[12] B. Zhong et al., “Research on the identification of network traffic anomalies in the access layer of power IoT based on extreme learning machine,” in 2022 International Conference on Artificial Intelligence, Information Processing and Cloud Computing (AIIPCC), Kunming, China, 2022.
[13] I. M. Lavrovsky and State University of Telecommunications, “Detection of traffic anomalies in the home Wi-Fi network using Waidps and Nzyme utilities,” Modern Information Security, vol. 52, no. 4, 2022.
[14] N. Kuchuk, A. Kovalenko, H. Kuchuk, V. Levashenko, and E. Zaitseva, “Mathematical methods of reliability analysis of the network structures: Securing QoS on hyperconverged networks for traffic anomalies,” in Lecture Notes in Electrical Engineering, Cham: Springer International Publishing,2022, pp. 223–241.
[15] M. Muniswamaiah and T. Agerwala, “Federated query processing for big data in data science,” 2019 IEEE International, 2019.
[16] H. Deng, W. Chen, and G. Huang, “Deep insight into daily runoff forecasting based on a CNN-LSTM model,” Nat. Hazards (Dordr.), vol. 113, no. 3, pp. 1675–1696, Sep. 2022.
[17] L. Zhang, “The evaluation on the credit risk of enterprises with the CNN- LSTM-ATT model,” Comput. Intell. Neurosci., vol. 2022, p. 6826573, Sep. 2022.
[18] H. Li, Z. Wang, and Z. Li, “An enhanced CNN-LSTM remaining useful life prediction model for aircraft engine with attention mechanism,” PeerJ Comput. Sci., vol. 8, no. e1084, p. e1084, Aug. 2022.
[19] N. Thakur and C. Y. Han, “Indoor localization for personalized ambient assisted living of multiple users in multi-floor smart environments,” Big Data Cogn. Comput., vol. 5, no. 3, p. 42, Sep. 2021.
[20] J. P. Singh, “Enhancing Database Security: A Machine Learning Approach to Anomaly Detection in NoSQL Systems,” International Journal of Information and Cybersecurity, vol. 7, no. 1, pp. 40–57, 2023.
[21] D. Gudu, M. Hardt, and A. Streit, “On MAS-based, scalable resource allocation in large-scale, dynamic environments,” in 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), Toulouse, 2016.
[22] J. P. Singh, “Mitigating Challenges in Cloud Anomaly Detection Using an Integrated Deep Neural Network-SVM Classifier Model,” Sage Science Review of Applied Machine Learning, vol. 5, no. 1, pp. 39–49, 2022.
[23] O. Kamara-Esteban et al., “Bridging the gap between real and simulated environments: A hybrid agent-based smart home simulator architecture for complex systems,” in 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), Toulouse, 2016.
[24] H. Lauer and N. Kuntze, “Hypervisor-based attestation of virtual environments,” in 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), Toulouse, 2016.