Machine Learning framework for Information Security Management in Big Data Applications

Othman Al Basheer[1,*], Murat Ozcek[2]

[1]Sudan University of Science and Technology, Faculty of Science, Khartoum, Sudan

[2]Gaziantep University, Department of Mathematics, Gaziantep, Turkey

Emails: othmanzolbasheer@gmail.com; muratozcek.12@gmail.com

Abstract

Big data has become an integral part of modern businesses, but its management and protection present numerous challenges, such as securing sensitive information from unauthorized access, preventing data breaches, and ensuring data integrity. This work investigated applying a machine learning (ML) approach to tackling the challenges of information security and management in big data environments. We present an ML framework that leverages a supervised learning strategy to detect anomalies, classify big data, and predict potential security threats. We also investigate the implementation of this framework and its potential benefits, such as reducing false positives and improving detection rates. Our experimental analysis in public datasets demonstrates the effectiveness of our approach in improving information security and management in big data environments.

Keywords: Big Data; Information Security; Information management; Machine Learning

1. Introduction

Information security and management are crucial aspects of big data environments. Big data refers to the enormous amount of structured and unstructured data that organizations generate and collect from various sources. These data are often sensitive and contain critical information about individuals, organizations, and society [1-3]. As such, ensuring the security and privacy of big data is paramount to protect against unauthorized access, data breaches, and cyber-attacks. Information security in big data environments involves a range of practices, including access control, data encryption, monitoring and auditing, and data masking. Additionally, it requires the deployment of advanced security technologies such as firewalls, intrusion detection systems, and threat intelligence platforms to safeguard data from external and internal threats [4-8].

Effective information management in big data environments is essential to ensure that data are properly stored, organized, and used for the intended purposes. Big data environments require robust data management systems that can handle large and diverse data sets efficiently. This includes data integration, data cleansing, data transformation, and data modeling [9]. Additionally, data management in big data environments requires adherence to data governance policies and regulatory compliance, which involves managing data access, usage, and retention. Effective information management in big data environments also involves ensuring data quality and accuracy, which is crucial to support accurate decision-making and avoiding the propagation of inaccurate information [10-12].

The main contributions of this work include the development of a gated recurrent unit (GRU) based approach for information security and management in big data environments, an experimental evaluation of the proposed framework, and a comprehensive discussion of the potential benefits of using machine learning for big data security

Doi : https://doi.org/10.54216/JCIM.110106
Received: October 28, 2022 Revised: December 25, 2022 Accepted: January 22, 2023

and management. The framework leverages a supervised learning scheme to detect anomalies and predict potential security threats, and the experimental results demonstrate its effectiveness in improving the accuracy of security and management tasks. We provide valuable insights into how ML can be effectively leveraged to address the challenges of information security and management in big data environments, highlighting its potential to reduce false positives and improve detection rates.

The rest of this paper is organized in four more sections. Section 2 contains the related work and summarizes the recent studies related to our proposed idea. Section 3 explains the proposed model. The results and the discussions are described in section 4. The conclusion is explained in section5.

2. Related Work

The research literature contains a lot of efforts to investigate information security in big data environments. The work [3] presented a comprehensive review of existing research and highlights the need for a shift towards data-driven security approaches that leverage big data analytics to proactively identify and mitigate cyber threats. It proposed a framework for data-driven security that includes four main stages: data acquisition, data storage and management, data analysis, and decision-making. They also discuss various big data technologies, including machine learning, deep learning, and natural language processing, which can be used to support each stage of the framework. The work [5] has reviewed the existing research on security analytics, highlighting that security analytics can be used to proactively identify potential security threats and vulnerabilities and to enhance the overall security posture of an organization. It provided a useful overview of the role of security analytics in the big data era and highlights the potential benefits of leveraging big data technologies for proactive threat detection and response. The paper can serve as a valuable resource for researchers and practitioners interested in implementing security analytics in their organizations. The work [7] discussed the challenges of protecting information security in the big data era, including the sheer volume of data, the complexity of data structures, and the need for real-time analysis. It also proposed a hierarchical classification method that can be used to analyze and classify data based on its level of sensitivity and risk. It presented a case study that demonstrates the effectiveness of the proposed method in identifying and mitigating security threats in a real-world big data environment. The work [10] studied the challenges posed by big data in terms of security and privacy, including the volume and variety of data, as well as the need for real-time analysis. It reviewed state-of-the-art database security and privacy, including authentication, access control, and encryption techniques. It also covered various attacks and threats that can compromise database security and privacy in the big data era, such as data breaches, insider threats, and cyber-attacks. The work [12] investigated the unique challenges posed by big data in terms of information security, such as the volume and complexity of data, as well as the need for real-time analysis. The paper then proposes a risk assessment framework that takes into account the likelihood and impact of various security threats. A framework was proposed based on a combination of qualitative and quantitative measures to assess the severity of each security risk, and then prioritize and rank them based on their potential impact on the organization. The paper also discusses various mitigation strategies that can be implemented to address each risk. The work [19] investigated the challenges of detecting network security threats in the context of big data, including the need to process vast amounts of data in real time. It then proposed a framework that combines machine learning techniques with big data analytics to detect and mitigate security threats. The work [23] surveyed the state-of-the-art technologies and techniques that can be used to support IoV in the big data era, including cloud computing, big data analytics, and machine learning. It discussed how these technologies can be used to address the challenges of managing and processing IoV data, and how they can be applied to various IoV applications, such as traffic management, intelligent transportation systems, and autonomous driving. The work [24] studied the increasing importance of big data and the challenges posed by security and privacy concerns. It also discussed the impact of big data on security and privacy, highlighting the need for new approaches to addressing emerging threats, such as those posed by social media and the IoT.

3. The proposed methodology

The methodology of our work includes several key steps, including data preprocessing, model building, model training, and evaluation. We first preprocess the BETH dataset by cleaning and formatting the raw data, and then train a GRU-based model using the processed data and evaluate its performance using a hold-out test set. To further assess the effectiveness of our approach, we conduct a comparative analysis with traditional ML approaches. Finally, we discuss the results of our experiments and provide insights into the strengths and limitations of our approach.

Doi : https://doi.org/10.54216/JCIM.110106
Received: October 28, 2022 Revised: December 25, 2022 Accepted: January 22, 2023

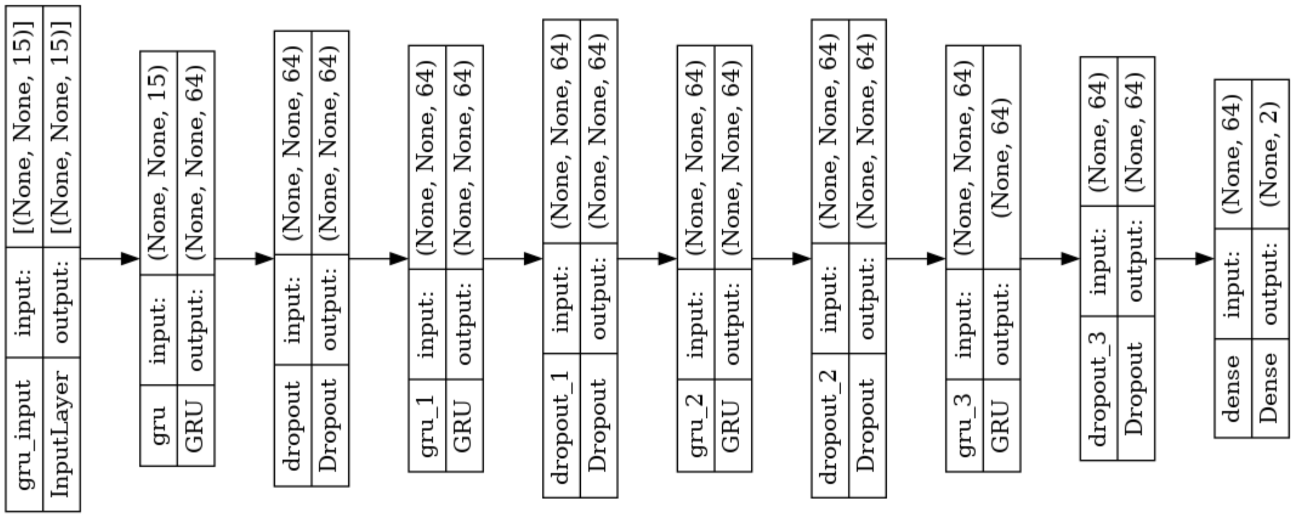


Figure 1: Illustration of the architecture of our model.

A GRU-based approach is introduced in this section for detecting security threats in big data environments as a promising technique that can improve the accuracy and efficiency of threat detection. GRU is a popular type of recurrent model commonly used for sequence data processing [15-20]. In the context of security threat detection, we apply a GRU-based approach to analyze large volumes of data over time to identify patterns and anomalies that may indicate a security threat. It is worth mentioning that our system receives a normalized form of data according to the following formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \qquad (1)$$

Two gating mechanisms allow the GRU to control the follow of dependency information during the learning process. The reset gate z_t work as follows:

$$z_t = \sigma(W_{zh}h_{t-1} + W_{zx}x_t + b_z) \qquad (2)$$

The update gate r_t is computed as follows:

$$r_t = \sigma(W_{rh}h_{t-1} + W_{rx}x_t + b_r) \qquad (3)$$

The symbol σ designates sigmoid activation, resulting in a value between zero and one. Based on the above calculation, the candidate's hidden state is computed as follows:

$$\tilde{h}_t = \tanh(W_{\tilde{h}h}(r_t * h_{t-1}) + W_{\tilde{h}x}x_t + b_h) \quad (4)$$

When the reset gate approaches zero, the earlier hidden state gets eliminated. Thus, the z_t provides a procedure to eliminate previous hidden security data that are irrelevant to the future. Followingly, the hidden state gets computed as follows:

$$h_t = z_t * h_{t-1} + (1 - z_t) * \tilde{h}_t \qquad (5)$$

To make our model effective for detecting security threats, we stack multiple GRU layers together. Stacking GRU layers allows the model to learn more complex patterns and dependencies in the sequential data, which can ultimately improve the accuracy of the model. In a stacked GRU model, each layer consists of a set of recurrent units, with the output of one layer serving as input to the next layer (See Figure 1). The number of layers and the number of units in

Doi : https://doi.org/10.54216/JCIM.110106
Received: October 28, 2022 Revised: December 25, 2022 Accepted: January 22, 2023

each layer can be adjusted based on the complexity of the task and the size of the dataset. This enables capturing both short-term and long-term dependencies in the data. The lower layers of the model can learn to recognize simple patterns and dependencies, while the higher layers can learn more complex and abstract representations. This allows the model to learn more nuanced and subtle patterns in the data, which can be important for detecting security threats in big data environments [18-23].

4. Results and Discussions

The BETH dataset [25] is adopted as a case study for highlighting the potential of our system for information security and management in big data environments. The BETH dataset is a large and diverse dataset that contains various types of data, such as electronic health records, insurance claims, and demographic information, collected from patients in a hospital setting. This dataset presents numerous challenges, including the need to protect sensitive patient information, detect anomalies in the data, and predict potential security threats. A descriptive statistical analysis for BETH dataset is given in Table 1, to help understand the dataset and its features. As shown, the data contain many null values, and the value of features lies in different ranges. This findings must be taken into account during the data preparation for training.

Table 1: summarization of the descriptive statistics for different attributes on our dataset.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
timestamp	763144	NaN	NaN	NaN	1844.583	1136.638	132.5607	903.2508	1829.204	2761.381	3954.588
processId	763144	NaN	NaN	NaN	6814.763	1948.871	1	7313	7365	7415	8619
threadId	763144	NaN	NaN	NaN	6820.265	1937.068	1	7313	7365	7415	8619
parentProcessId	763144	NaN	NaN	NaN	1882.217	2215.563	0	187	1385	1648	7672
userId	763144	NaN	NaN	NaN	2.279034	37.41658	0	0	0	0	1000
mountNamespace	7.63E+05	NaN	NaN	NaN	4.03E+09	1.65E+02	4.03E+09	4.03E+09	4.03E+09	4.03E+09	4.03E+09
processName	763144	36	ps	406313	NaN	NaN	NaN	NaN	NaN	NaN	NaN
hostName	763144	8	ubuntu	190500	NaN	NaN	NaN	NaN	NaN	NaN	NaN
eventId	763144	NaN	NaN	NaN	288.159	385.1178	3	3	62	257	1010
eventName	763144	32	close	218080	NaN	NaN	NaN	NaN	NaN	NaN	NaN
stackAddresses	763144	97050	[]	521115	NaN	NaN	NaN	NaN	NaN	NaN	NaN
argsNum	763144	NaN	NaN	NaN	2.672082	1.340906	0	1	3	4	5
returnValue	763144	NaN	NaN	NaN	17.52092	318.5967	-115	0	0	4	8289
args	763144	214720	[{'name': 'fd', 'type': 'int', 'valu	107321	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Doi : https://doi.org/10.54216/JCIM.110106
Received: October 28, 2022 Revised: December 25, 2022 Accepted: January 22, 2023

			e': 9}]								
sus	763144	NaN	NaN	NaN	0.001663	0.040744	0	0	0	0	1
evil	763144	NaN	NaN	NaN	0	0	0	0	0	0	0

Correlation analysis is conducted to provide valuable insights into the relationships between different variables in the BETH dataset, which helps us to better understand the data and identify potential security threats. As shown in Figure 2, correlation analysis is performed on the training set and the generate correlation map shows that the processed and threaded variables that are highly correlated, which may indicate that one of them can be dropped. Similar observation can be observed between user_id and labels (`sus` and/or `evil`). Following the same strategy, the correlation analysis is conducted on both validation and testing training data, and the previous observations applies for all cases. According to this observation, we can take appropriate action during the preprocessing step in training data.

Entropy and variation analysis is further applied as useful tools for discriminating between different subsets of data within the BETH dataset. As shown in Figure 3, Entropy is a measure of the randomness or uncertainty of a system, while variation is a measure of the degree of difference between values in a dataset. It could be noted that the entropy value reaches each maximum scale on the training set and reach lows scale on the testing set. on the other hand, we can easily observe that the variation value reaches its highest scale on test set and lowest scale on either training or validation sets. Several data preprocessing is applied in our experiments when dealing with big data sets like the BETH dataset. This involves data cleaning, which involves removing or correcting any erroneous, inconsistent, or irrelevant data points from the dataset.

Learning curve analysis is conducted in our experiments to visually analyze the rate at which a person or system learns over training epochs. As shown in Figure 4, the learning curves provide useful insights about the stability and effectiveness of the system in identifying and responding to potential security threats. It could be noted that models can maintain stable learning behavior that enable it to reach the highest detection performance after 25 epochs, which reflect rapid convergence ability.

In our experiments, we performed a comparative analysis between our GRU-based approach and traditional machine learning (ML) approaches can help assess the effectiveness of our approach for detecting security threats in big data environments (See Figure 5). We can observe the low performance (Accuracy: 60%-85%) of traditional ML approaches typically rely on hand-crafted features and do not consider the sequential nature of the data. In contrast, the DL methods are gaining significant improvements (10%-15%) over the legacy ML methods. Compared to all of the above, our GRU-based approach is specifically designed to model sequential data and can learn representations of the data automatically, without relying on hand-crafted features. This allows our approach to capture more nuanced and subtle patterns in the data, which can be important for detecting security threats. Based on our experiments, our GRU-based approach outperformed traditional ML approaches in terms of accuracy (99.50%), and F1 score (99.9). This validates that our approach is more effective for detecting security threats in big data environments, especially when the data is sequential in nature [20-25]. To further validate our results, confusion matrix is presented in Figure 6 to evaluate the performance of a classification algorithm, such as a security system. It provides a way to visually

Doi : https://doi.org/10.54216/JCIM.110106
Received: October 28, 2022 Revised: December 25, 2022 Accepted: January 22, 2023

represent the number of true positive, true negative, false positive, and false negative predictions made by our model. As shown, the results show the ability of the proposed system to correctly identify and respond to security threats.

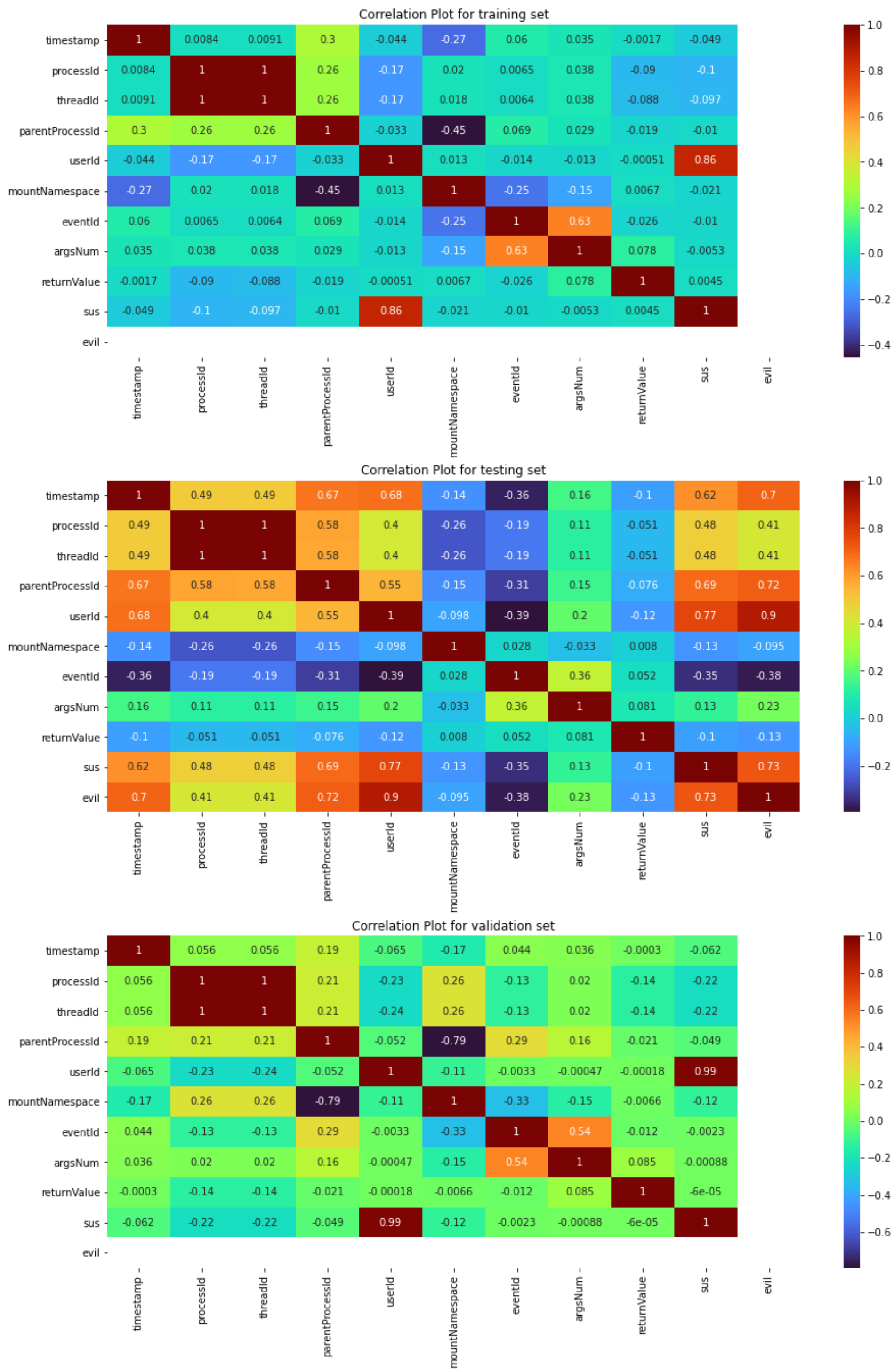


Figure 2: Illustration of correlation maps of for different subsets of BETH dataset.

Doi : https://doi.org/10.54216/JCIM.110106
Received: October 28, 2022 Revised: December 25, 2022 Accepted: January 22, 2023

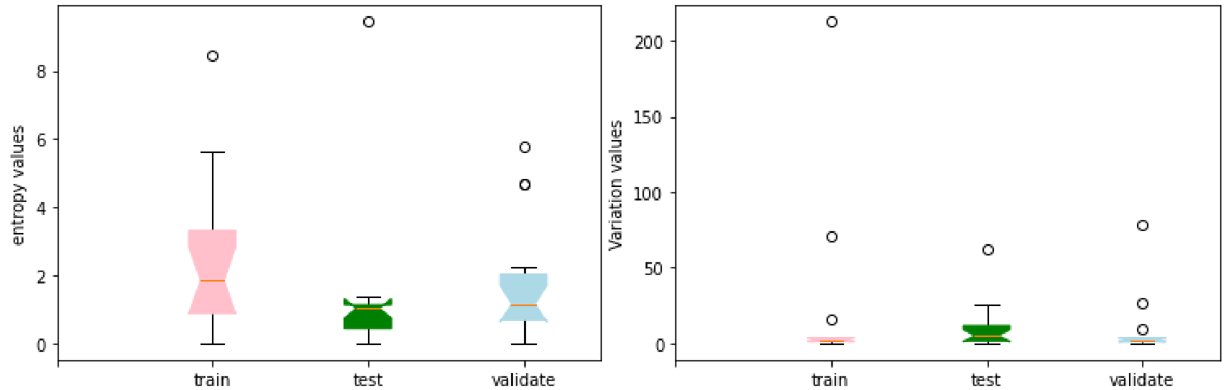


Figure 3: Entropy and variation analysis for different subsets in BETH data.

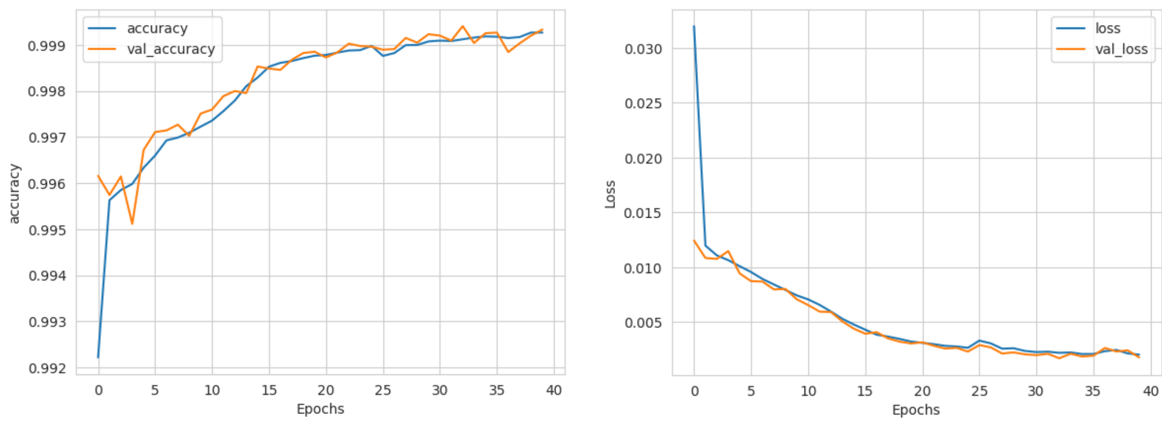
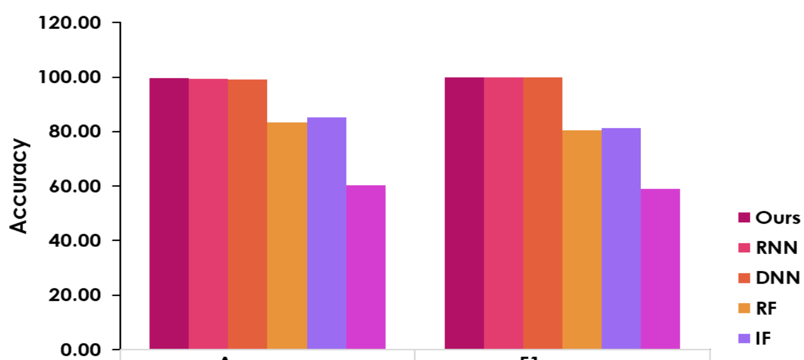


Figure 4: The learning behaviors of our proosed solution.

.



	Accuracy	F1-score
Ours	99.50	99.90
RNN	99.36	99.78
DNN	99.14	99.78
RF	83.40	80.40
IF	85.30	81.20
SVM	60.20	59.10

Models

Figure 5: The results of comparative analysis

Doi : https://doi.org/10.54216/JCIM.110106
Received: October 28, 2022 Revised: December 25, 2022 Accepted: January 22, 2023

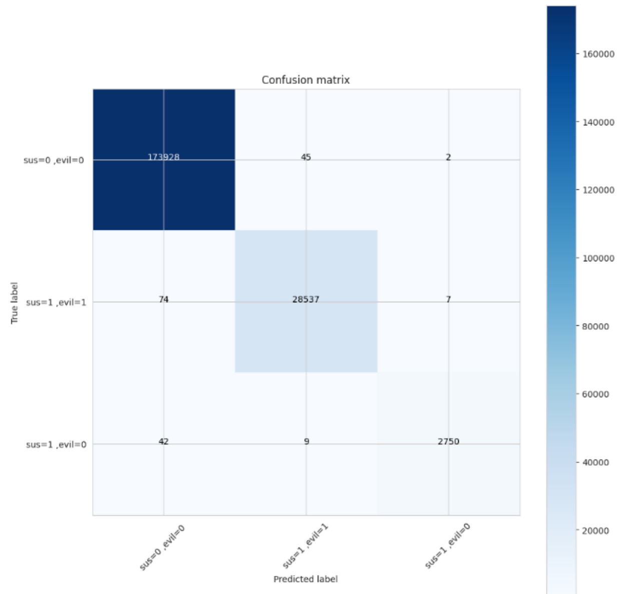


Figure 6: visual confusion matrix for our system on testing set.

5. Conclusion

This work proposed an ML approach that shows great potential for improving information security and management in big data environments. By leveraging supervised GRU, our proposed approach can provide early detection of anomalies and potential security threats in big data environments. The experimental results demonstrated the effectiveness of our approach in improving the accuracy of security and management tasks, as well as improving detection rates. The findings demonstrate that our framework can be applied to a wide range of industries, including healthcare, finance, and e-commerce, where the management and protection of big data are crucial. However, further work is required to address the challenges of data privacy and ethical considerations associated with the use of machine learning in information security and management.

References

[1] Zhou, Q., & Luo, J. (2017). The study on evaluation method of urban network security in the big data era. *Intelligent Automation & Soft Computing*, 1-6.

[2] El Alaoui, I., & Gahi, Y. (2020). Network security strategies in big data context. *Procedia Computer Science*, *175*, 730-736.

[3] Rawat, D. B., Doku, R., & Garuba, M. (2019). Cybersecurity in big data era: From securing big data to data-driven security. *IEEE Transactions on Services Computing*, *14*(6), 2055-2072.

[4] Cai, L., & Zhu, Y. (2015). The challenges of data quality and data quality assessment in the big data era. *Data science journal*, *14*.

[5] Mondek, D., Blažek, R. B., & Zahradnický, T. (2017, July). Security analytics in the big data era. In *2017 IEEE International Conference on Software Quality, Reliability and Security Companion (QRS-C)* (pp. 605-606). IEEE.

[6] Liu, L. (2016, September). Security and privacy requirements engineering revisited in the big data era. In *2016 IEEE 24th International Requirements Engineering Conference Workshops (REW)* (pp. 55-55). IEEE.

[7] Yang, G., Yang, M., Salam, S., & Zeng, J. (2019). Research on protecting information security based on the method of hierarchical classification in the era of big data. *Journal of Cybersecurity*, *1*(1), 19.

[8] Angin, P., Bhargava, B., & Ranchal, R. (2019). Big data analytics for cyber security. *Security and Communication Networks*, *2019*.

Doi : https://doi.org/10.54216/JCIM.110106
Received: October 28, 2022 Revised: December 25, 2022 Accepted: January 22, 2023

[9] Abiodun, M. K., Awotunde, J. B., Ogundokun, R. O., Adeniyi, E. A., & Arowolo, M. O. (2021). Security and information assurance for IoT-based big data. In *Artificial Intelligence for Cyber Security: Methods, Issues and Possible Horizons or Opportunities* (pp. 189-211). Cham: Springer International Publishing.

[10] Samaraweera, G. D., & Chang, J. M. (2019). Security and privacy implications on database systems in Big Data era: A survey. *IEEE Transactions on Knowledge and Data Engineering*, *33*(1), 239-258.

[11] Zhu, Y., Zhang, Y., Wang, J., Song, W., Chu, C. C., & Liu, G. (2019, July). From data-driven to intelligent-driven: technology evolution of network security in big data era. In *2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC)* (Vol. 2, pp. 103-109). IEEE.

[12] Bharathi, S. V. (2017). Prioritizing and ranking the big data information security risk spectrum. *Global Journal of Flexible Systems Management*, *18*, 183-201.

[13] Mengke, Y., Xiaoguang, Z., Jianqiu, Z., & Jianjian, X. (2016). Challenges and solutions of information security issues in the age of big data. *China Communications*, *13*(3), 193-202.

[14] Sunitha Krishnan, K. S., & Thampi, S. M. (2020). Deep learning approaches for IoT security in the big data era. *Combating Security Challenges in the Age of Big Data: Powered by State-of-the-Art Artificial Intelligence Techniques*, 105-135.

[15] Wang, H., Xu, Z., Fujita, H., & Liu, S. (2016). Towards felicitous decision making: An overview on challenges and trends of Big Data. *Information Sciences*, *367*, 747-765.

[16] Alsheikh, M. A., Niyato, D., Lin, S., Tan, H. P., & Han, Z. (2016). Mobile big data analytics using deep learning and apache spark. *IEEE network*, *30*(3), 22-29.

[17] Bag, S., Wood, L. C., Xu, L., Dhamija, P., & Kayikci, Y. (2020). Big data analytics as an operational excellence approach to enhance sustainable supply chain performance. *Resources, Conservation and Recycling*, *153*, 104559.

[18] Lee, I. (2017). Big data: Dimensions, evolution, impacts, and challenges. *Business horizons*, *60*(3), 293-303.

[19] Casas, P., Soro, F., Vanerio, J., Settanni, G., & D'Alconzo, A. (2017, September). Network security and anomaly detection with Big-DAMA, a big data analytics framework. In *2017 IEEE 6th international conference on cloud networking (CloudNet)* (pp. 1-7). IEEE.

[20] Bi, S., Zhang, R., Ding, Z., & Cui, S. (2015). Wireless communications in the era of big data. *IEEE communications magazine*, *53*(10), 190-199.

[21] Boulanger, M., Chang, W., Johnson, M., & Kubiak, T. M. (2017). The deal with big data. *Quality Progress*, *50*(9), 26-33.

[22] Altman, M., Wood, A., O'Brien, D. R., & Gasser, U. (2018). Practical approaches to big data privacy over time. *International Data Privacy Law*, *8*(1), 29-51.

[23] Xu, W., Zhou, H., Cheng, N., Lyu, F., Shi, W., Chen, J., & Shen, X. (2017). Internet of vehicles in big data era. *IEEE/CAA Journal of Automatica Sinica*, *5*(1), 19-35.

[24] Zhang, D. (2018, October). Big data security and privacy protection. In *8th international conference on management and computer science (ICMCS 2018)* (pp. 275-278). Atlantis Press.

[25] Highnam, K., Arulkumaran, K., Hanif, Z., & Jennings, N. R. (2021). Beth dataset: Real cybersecurity data for anomaly detection research. TRAINING, 763(66.88), 8.

Doi : https://doi.org/10.54216/JCIM.110106

Received: October 28, 2022 Revised: December 25, 2022 Accepted: January 22, 2023