International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 13 Issue 4, April, 2024, Impact Factor: 8.285

Log-based Anomaly Detection: Unveiling Patterns in ELK Stack Data for Enhanced Financial Security

Saurabh Shete[1], Mukta Takalikar[2], Dipesh Sonawane[3],
Sara Shaikh[1], Nidhi Yadav[1], Diksha Shivarkar[1]

[1]Student,Computer Engineering, Pune Institute of Computer Technology, Pune, Maharashtra, India.
[2]Assistant Professor,Computer Engineering, Pune Institute of Computer Technology, Pune, Maharashtra, India.
[3]Student,Information Technology, Sinhgad College of Engineering, Pune, Maharashtra, India.

ABSTRACT

In this research, we delve into the realm of anomaly detection within log data obtained from the ELK (Elasticsearch, Logstash, Kibana) stack. With a specific focus on enhancing security and performance monitoring, our study involves a rigorous investigation into data pre-processing techniques and feature extraction methods. The implementation leverages the Isolation Forest algorithm to discern anomalies within transactional patterns. The findings underscore the efficacy of our approach, providing valuable insights into the nuanced landscape of anomaly detection using ELK logs. This research contributes to the evolving field of log-based anomaly detection, offering a practical and efficient methodology for monitoring and securing diverse systems.

Keywords: ELK Stack, Anomaly Detection, Isolation Forest, Security Monitoring, Performance Monitoring, Machine Learning,Log Analysis, Cybersecurity, Data Preprocessing, Feature Extraction

INTRODUCTION

In the ever-evolving landscape of information technology, the proactive identification and mitigation of irregularities in system behavior have become imperative. Anomaly detection, a fundamental aspect of cybersecurity and system monitoring, plays a pivotal role in identifying unexpected patterns that may indicate security threats, performance issues, or potential vulnerabilities. As organizations increasingly rely on sophisticated software and infrastructure to manage and process vast amounts of data, the need for robust anomaly detection methodologies becomes paramount. At the core of this study lies the ELK stack, a powerful and widely adopted suite of tools comprising Elasticsearch, Logstash, and Kibana. Elasticsearch serves as a distributed search and analytics engine, Logstash facilitates log parsing and processing, while Kibana provides a user-friendly interface for data visualization. The ELK stack has emerged as a go-to solution for organizations seeking efficient log management, real-time analytics, and streamlined data visualization.

The motivation behind this research stems from the critical importance of anomaly detection within the ELK stack, particularly in addressing the evolving challenges posed by sophisticated cyber threats and the growing complexity of system architectures. As organizations increasingly rely on the ELK stack for log management and analysis, the need to fortify these systems against anomalies becomes more pronounced.

In pursuit of this goal, various anomaly detection approaches will be explored within the ELK context. These approaches include statistical methods such as Z-score and interquartile range, machine learning algorithms like Isolation Forest and One-Class SVM, and time-series analysis to identify patterns and deviations from expected behavior.The central problem addressed by this research lies in optimizing anomaly detection methodologies within ELK logs, with a specific focus on enhancing security and performance monitoring. By delving into the intricacies of log data generated by applications and systems, we aim to develop a nuanced understanding of how anomalies manifest within the ELK stack. Through this exploration, we seek to contribute practical insights and methodologies that can be employed by organizations relying on the ELK stack for log management and real-time analytics.

In the subsequent sections, we will delve into the specifics of ELK-based anomaly detection, exploring data preprocessing, feature extraction, and the implementation of the Isolation Forest algorithm to discern and classify

anomalies within transactional patterns. The findings of this research aim to not only enhance the anomaly detection capabilities of ELK-based systems but also contribute to the broader field of log-based anomaly detection methodologies.

LITERATURE SURVEY

In the realm of cybersecurity and log analysis, the integration of anomaly detection techniques within the ELK (Elasticsearch, Logstash, Kibana) stack has garnered considerable attention from researchers. A survey of pertinent literature reveals significant contributions in diverse applications.

Real-time detection systems leveraging the ELK stack for identifying malicious activities within organizational environments have been proposed. The frameworks include a threat intelligence interface, enabling the prioritization of detected threats on the network. The practicality is demonstrated through manual intervention to remove files associated with identified threats [1]. A comparative analysis of the ELK stack's performance in security log analysis against commercial systems has been conducted. The study highlighted the cost-effectiveness and efficiency of ELK, showcasing its comparable or superior performance in specific security log searches. This finding positions ELK as an attractive solution for small to medium-sized enterprises seeking robust log analysis capabilities[2]. Efforts to enhance network security through effective log analysis using ELK have identified critical gaps in existing commercial safeguards. Continuous log analysis is emphasized to monitor network efficiency, highlighting the significance of log analysis in identifying vulnerabilities with potential repercussions for network security [3].

A comprehensive framework for network anomaly detection, combining big-data-based analysis with machine learning technologies, has been proposed. The introduced multi-source log analyzing system, tested in a real-world network environment, demonstrated notable advantages in terms of timeliness, accuracy, and scalability. The study positions the framework as a practical solution for network anomaly detection [4]. An anomaly detection system within the Hadoop environment, employing log data management and analysis architecture, has been devised. The study introduced basic anomaly detection methods using HiveQL, further enhancing them with weight-based methods. Experimental results showcased improved accuracy in anomaly detection, particularly with weighted methods, opening avenues for future applications in large-scale Hadoop environments [5].

[6] An operational logs analysis solution at the ALMA observatory using the ELK stack has been presented. Configured as a decoupled system, the solution incurs minimal operational impact and allows storage of over six months of activity logs online. The study provides insights into specific failures and long-term performance analysis, showcasing the practicality of ELK in large-scale operational settings[7].

Moreover, a comprehensive analysis of the growing problem of financial fraud in the contemporary technological and economic milieu has been undertaken through survey papers [2], [5], [6], and [7].

Survey paper [2] provides an insightful analysis of the growing problem of financial fraud.[8] It emphasizes the shortcomings in current fraud prevention strategies and advocates for robust fraud detection schemes. The progression of anomaly detection techniques is explored, from supervised to semi-supervised and unsupervised methods. Customized approaches for different types of financial fraud are highlighted, with a focus on auto-generating models like GANS and different AE networks, demonstrating high efficiency in feature extraction and oversampling. Credit card fraud detection tools, including deep learning architectures with CNNs and LSTMs, are discussed.

The analysis of financial fraud based on a manager knowledge graph is introduced in paper [6][9], presenting a Knowledge Graph (KG) framework for financial fraud detection. The KG strategy utilizes relationships among managers and relevant institutions to discover tacit knowledge, showcasing the potential of SVM models with topological features from the KG for superior fraud detection. This approach signifies a shift towards holistically-based and data-rich methodologies, considering complex relationships within the financial ecosystem.[10] In paper [7], research on Fraud Detection in Credit Card Transactions using Anomaly Detection is outlined. It delves into the peculiarities associated with credit card fraud, emphasizing the advantages and dangers of using credit cards for online purchases. The study discusses predictive algorithms and data mining for fraud prevention, highlighting the competence of the Isolation Forest Algorithm for precise and effective fraud detection. The report recognizes the complexities of the issue but emphasizes how machine learning enhances credit card users' financial safety in making secure transactions.

This amalgamation of literature surveys underscores the comprehensive understanding of both ELK-based anomaly detection and the broader landscape of financial fraud detection methodologies. The studies collectively emphasize the efficacy, efficiency, and versatility of ELK-based solutions while providing insights into evolving strategies for combating financial fraud [11].

METHODOLOGY

A. Data Collection

The first step in our methodology involves the collection of log data from various sources within the organization's IT infrastructure. This encompasses system logs, application logs, network logs, and any other relevant sources of data. The collected data should cover a diverse range of activities and events to ensure the effective detection of anomalies within the system.

B. Feature Extraction

With the preprocessed data in hand, the next step is to identify and extract relevant features that can be used for anomaly detection. Features such as timestamps, event types, source and destination IP addresses, user IDs, and other pertinent information are extracted from the log data. This process is guided by domain knowledge and insights from the literature survey conducted as part of our research.

C. Anomaly Detection Algorithms

In this phase, we implement various anomaly detection algorithms within the ELK stack environment. Leveraging the capabilities of Elasticsearch, Logstash, and Kibana, we explore a range of techniques including statistical methods such as Z-score and interquartile range, machine learning algorithms like Isolation Forest and One-Class SVM, and time-series analysis approaches. These algorithms are applied to the extracted features to identify abnormal patterns indicative of anomalies within the transactional data.

D. Model Training and Evaluation

The developed anomaly detection models are trained and evaluated using appropriate datasets. The data is split into training and testing sets for model development and evaluation purposes. Parameters of the algorithms are adjusted to optimize performance, and model performance is assessed using metrics such as precision, recall, F1-score, and ROC-AUC. Cross-validation and sensitivity analysis are conducted to ensure the robustness and generalization capabilities of the models.

E. Integration with ELK Stack

An important aspect of our methodology is the integration of the developed anomaly detection models into the ELK stack infrastructure. This involves configuring Logstash to ingest log data and apply anomaly detection algorithms for real-time analysis. Detected anomalies are visualized using Kibana dashboards, providing actionable insights for security and performance monitoring within the organization.

F. Validation and Deployment

Following the integration phase, the effectiveness of the anomaly detection system is validated through pilot testing and validation against known anomalies. Feedback from stakeholders and domain experts is solicited to refine the system and address any identified issues or limitations. Once validated, the anomaly detection system is deployed into production, ensuring seamless integration with existing IT infrastructure and operational workflows.

MODEL AND RESULT

A. DeepLog

This framework, DeepLog, introduces a novel approach to online log anomaly detection and diagnosis leveraging deep neural networks. DeepLog effectively learns and encodes complete log messages, including timestamps, log keys, and parameter values, enabling anomaly detection at the per-log-entry level. This stands in contrast to previous methods, which often operate at the per-session level. Moreover, DeepLog is capable of task separation within log files, constructing workflow models for each task using a combination of deep learning (LSTM) and classical mining (density clustering) techniques, thus facilitating efficient anomaly diagnosis. Additionally, DeepLog supports online update/training of its LSTM models based on user feedback, allowing for adaptation to new execution patterns. Extensive evaluation on large system logs has demonstrated the superior effectiveness of DeepLog compared to existing methods. Future directions for research include exploring the efficiency of incorporating other types of recurrent neural networks (RNNs) into DeepLog and integrating log data from diverse applications and systems to enable more comprehensive system diagnosis. For instance, failures in a MySQL database may be linked to disk failures as reflected in separate system logs.

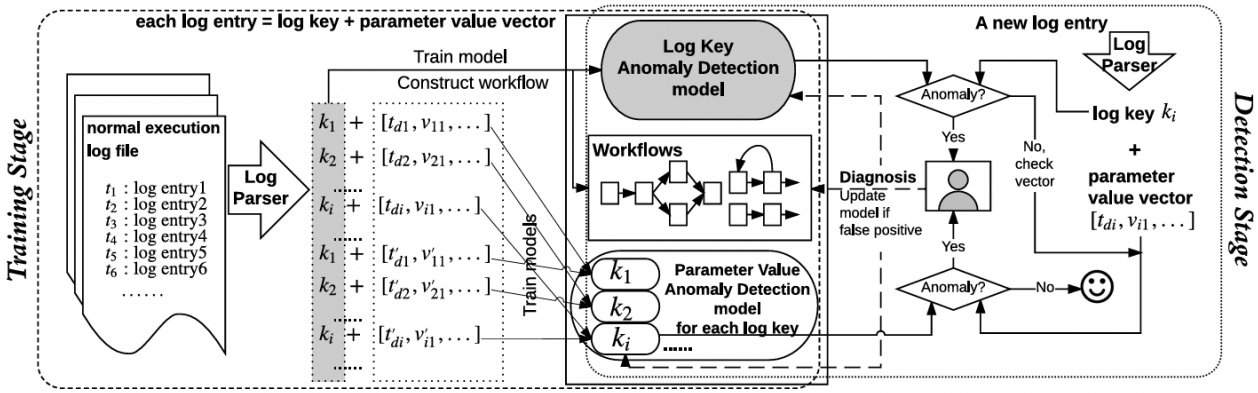

Fig 1. DeepLog Architecture

B. LogAnomaly

Many computer systems rely on logs to track their runtime status, and identifying anomalies in these logs is crucial for detecting system malfunctions promptly. However, manual anomaly detection in logs is time-consuming, error-prone, and impractical. Existing automated approaches often rely on indexes rather than the semantics of log templates, leading to frequent false alarms. In this study, we introduce LogAnomaly, a framework designed to treat unstructured log streams as natural language sequences. Utilizing a novel method called template2vec, LogAnomaly extracts semantic information embedded within log templates. This approach enables LogAnomaly to simultaneously detect both sequential and quantitative log anomalies, a capability not previously achieved. Additionally, LogAnomaly addresses the issue of false alarms caused by newly appearing log templates between periodic model retrainings. Evaluation on two publicly available production log datasets demonstrates that LogAnomaly outperforms existing methods for log-based anomaly detection.

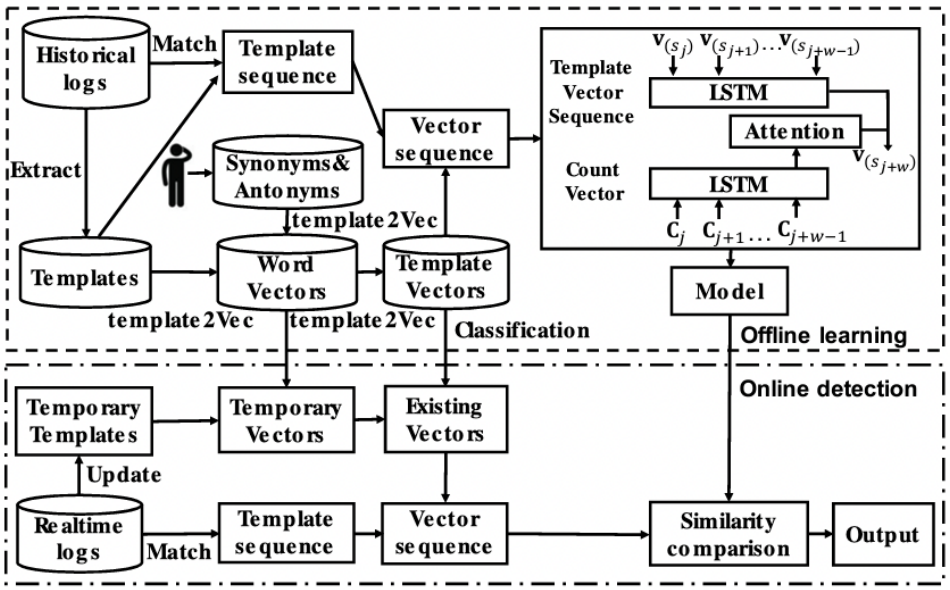

Fig 2. LogAnamoly Architecture

C. RobustLog

Logs play a vital role in troubleshooting large and complex software-intensive systems. Despite extensive research on log-based anomaly detection, existing methods often fall short in practical applications. These methods typically rely on historical log event data to construct detection models, assuming a closed-world scenario where log data remains stable over time and the set of distinct log events is known. However, empirical evidence reveals that real-world log data frequently contains previously unseen events or sequences due to the evolution of logging statements and processing noise. To address this challenge, we introduce LogRobust, a novel log-based anomaly detection approach. LogRobust leverages semantic information extracted from log events, representing them as semantic vectors. It employs an attention-based Bidirectional Long Short-Term Memory (Bi-LSTM) model to detect anomalies, enabling it to capture contextual information within log sequences and automatically learn the importance of different log events. By doing so, LogRobust effectively identifies and handles unstable log events and sequences. We evaluate LogRobust using logs from the Hadoop system and a Microsoft online service system. Results from our experiments demonstrate that LogRobust successfully

addresses the issue of log instability and achieves accurate and robust anomaly detection performance on real-world, dynamically changing log data.

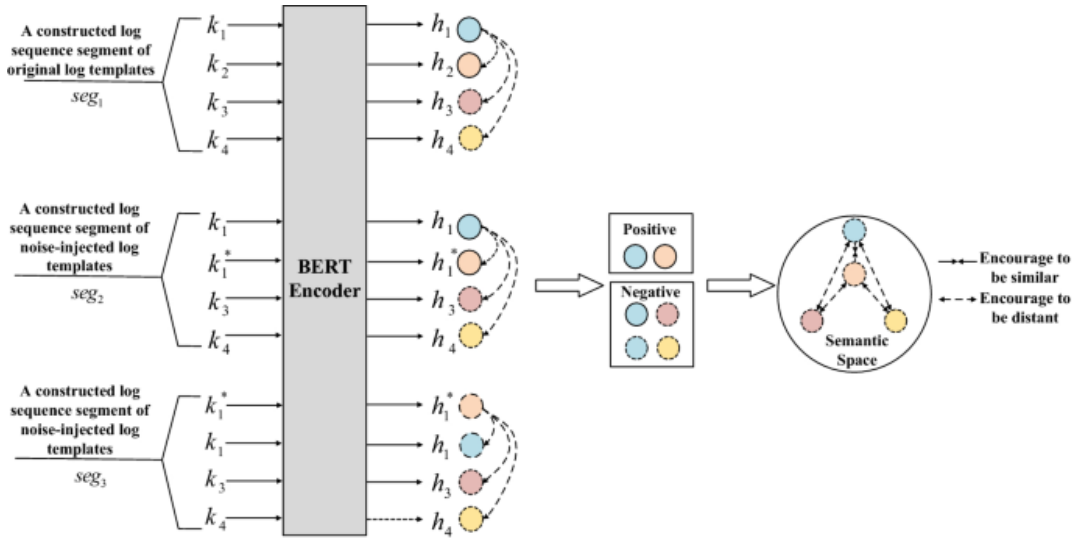

Fig 3.DeepLogArchitechture

Table1. Results

Model	Feature	Precision	Recall	F1
DeepLog(unsupervised)	seq	0.9583	0.9330	0.9454
LogAnamoly(unsupervised)	Seq+quan	0.9690	0.9825	0.9757
RobustLog	Semantic	0.9216	0.9586	0.9397

CONCLUSION AND FUTURE SCOPE

In conclusion, our study highlights the critical role of logs in large-scale software systems for troubleshooting and anomaly detection. We have presented two novel frameworks, DeepLog and LogAnomaly, each offering unique approaches to address the challenges associated with log-based anomaly detection.

DeepLog introduces a general-purpose framework utilizing deep neural networks for online log anomaly detection and diagnosis. By encoding entire log messages, including timestamp, log key, and parameter values, DeepLog enables anomaly detection at the per log entry level. Through the integration of deep learning and classic mining approaches, DeepLog constructs workflow models for different tasks within log files, facilitating effective anomaly diagnosis. Additionally, DeepLog's capability to incorporate user feedback supports online update/training to adapt to new execution patterns, as demonstrated in extensive evaluations on large system logs.

On the other hand, LogAnomaly proposes a framework to model unstructured log streams as natural language sequences, empowering the detection of both sequential and quantitative log anomalies simultaneously. By leveraging template2vec, LogAnomaly extracts semantic information from log templates, mitigating false alarms caused by newly appearing log templates between periodic model retrainings.Our evaluation on public production log datasets showcases LogAnomaly's superiority over existing log-based anomaly detection methods.

Looking ahead, future research directions include exploring the integration of other types of recurrent neural networks (RNNs) into DeepLog to further enhance its efficiency. Additionally, integrating log data from diverse applications and systems will enable more comprehensive system diagnosis, addressing real-world challenges such as identifying the root causes of failures across interconnected systems. These advancements will contribute to the ongoing evolution of log-based anomaly detection methods, improving their effectiveness and applicability in complex software environments.

REFERENCES

[1]. Mohiuddin Ahmed, Abdun Naser Mahmood, Md. Rafiqul Islam, A survey of anomaly detection techniques in financial domain, Future Generation Computer Systems, Volume 55, 2016, Pages 278-288, ISSN 0167-739X, https://doi.org/10.1016/j.future.2015.01.001.

[2]. Waleed Hilal, S. Andrew Gadsden, John Yawney, Financial Fraud: A Review of Anomaly Detection Techniques and Recent Advances, Expert Systems with Applications, Volume 193, 2022, 116429, ISSN 0957-4174,https://doi.org/10.1016/j.eswa.2021.116429.

[3]. Pankaj Richhariya and Prashant K Singh. Article: A Survey on Financial Fraud Detection Methodologies. International Journal of Computer Applications 45(22):15-22, May 2012.

[4]. Martin Jullum, Anders Løland, Ragnar Bang Huseby, Geir Ånonsen, Johannes Lorentzen. Detecting money laundering transactions with machine learning. Journal of Money Laundering Control, January 2020.https://doi.org/10.1108/jmlc-07-2019-0055

[5]. Archana Anandakrishnan, Senthil Kumar, Alexander Statnikov, Tanveer Faruquie, Di Xu. Anomaly Detection in Finance: Editors' Introduction. Proceedings of the KDD 2017: Workshop on Anomaly Detection in Finance, Proceedings of Machine Learning Research, 2018.https://proceedings.mlr.press/v71/anandakrishnan18a.html

[6]. Wen, Shigang& Li, Jianping & Zhu, Xiaoqian& Liu, Mingxi. (2022). Analysis of financial fraud based on manager knowledge graph. Procedia Computer Science. 199. 773-779.mhttps://doi.org/10.1016/j.procs.2022.01.096

[7]. Asheesh Kumar Dwivedil, Ashish Kumar Rai, Ashish Kashyap. (2021). Fraud Detection in Credit Card Transactions using Anomaly Detection. Vol. 12 No. 12 (2021)

[8]. AL. Sayeth Saabith, T. Vinothraj, MMM.Fareez, MM. Marzook. A survey of machine learning techniques for anomaly detection in cybersecurity. International Journal of Research in Engineering and Science (IJRES) ISSN: 2320-9364 https://www.ijres.org/papers/Volume-11/Issue-10/1110183193.pdf

[9]. Ashfaq T, Khalid R, Yahaya AS, Aslam S, Azar AT, Alsafari S, Hameed IA. A Machine Learning and Blockchain Based Efficient Fraud Detection Mechanism. Sensors. 2022; 22(19):7162. https://doi.org/10.3390/s22197162

[10]. Megdad, Mosa M. M. ; Abu-Naser, Samy S. & Abu-Nasser, Bassem S. (2022). Fraudulent Financial Transactions Detection Using Machine Learning. International Journal of Academic Information Systems Research (IJAISR) 6 (3):30-39.

[11]. Maya B Dhone, E.Nitya. BIG DATA ANALYTICS FOR FRAUD DETECTION IN FINANCIAL TRANSACTIONS, Journal of Data Acquisition and Processing, 2023, 38 (3): 290-307 . ISSN 1004-9037 https://sjcjycl.cn/article/view-2023/pdf/03_290.pdf

[12]. M. Aschi, S. Bonura, N. Masi, D. Messina and D. Profeta, "Cybersecurity and fraud detection in financial transactions", Big Data and Artificial Intelligence in Digital Finance, pp. 269-278, 2022.

[13]. Kotsiantis, Sotiris & Kanellopoulos, D. &Pintelas, P.. (2005). Handling imbalanced datasets: A review. GESTS International Transactions on Computer Science and Engineering. 30. 25-36.

[14]. [14] Hall, M. (1999) Correlation-based feature selection for machine learning. PhD Thesis, The University of Waikato. https://www.cs.waikato.ac.nz/~mhall/thesis.pdf

[15]. Hodge, V. J., & Austin, J. (2013). A Survey of Outlier Detection Methodologies. In S. Babones (Ed.), Fundamentals of Regression Modeling (SAGE Benchmarks in Social Research Methods). https://www-users.york.ac.uk/~vjh5/myPapers/Hodge+Austin_OutlierDetection_AIRE381.pdf

[16]. Félix Iglesias Vázquez, Alexander Hartl, Tanja Zseby, Arthur Zimek,Anomaly detection in streaming data: A comparison and evaluation study,Expert Systems with Applications, Volume 233, 2023, 120994,ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2023.120994.

[17]. H. He and E. A. Garcia, "Learning from Imbalanced Data," in IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp. 1263-1284, Sept. 2009, doi: https://doi.org/10.1109/TKDE.2008.239

[18]. Wang, H., Yang, R., Shi, J. (2023). Anomaly Detection in Financial Transactions Via Graph-Based Feature Aggregations. In: Wrembel, R., Gamper, J., Kotsis, G., Tjoa, A.M., Khalil, I. (eds) Big Data Analytics and Knowledge Discovery. DaWaK 2023. Lecture Notes in Computer Science, vol 14148. Springer, Cham. https://doi.org/10.1007/978-3-031-39831-5_6

[19]. Boukherouaa, E. B., AlAjmi, K., Deodoro, J., Farias, A., & Ravikumar, R. (2021). Powering the Digital Economy: Opportunities and Risks of Artificial Intelligence in Finance, Departmental Papers, 2021(024), A001. Retrieved Nov 15, 2023, from https://doi.org/10.5089/9781589063952.087.A001

[20]. Misra, S., Thakur, S., Ghosh, M., & Saha, S. K. (2019). An Autoencoder Based Model for Detecting Fraudulent Credit Card Transaction. Procedia Computer Science, 167, 254-262. https://doi.org/10.1016/j.procs.2020.03.219