International Journal Software Engineering and Computer Science (IJSECS)
4 (3), 2024, 983-989
Published Online December 2024 in IJSECS (http://www.journal.lembagakita.org/index.php/ijsecs)
P-ISSN: 2776-4869, E-ISSN: 2776-3242. DOI: https://doi.org/10.35870/ijsecs.v4i3.3358.

RESEARCH ARTICLE Open Access

The Application of Artificial Intelligence for Anomaly Detection in Big Data Systems for Decision-Making

Cut Susan Octiva *
Universitas Amir Hamzah, Deli Serdang Regency, North Sumatra Province, Indonesia.
Corresponding Email: cutsusan875@gmail.com.

Dikky Suryadi
STMIK Al Muslim, Bekasi Regency, West Java City, Indonesia.
Email: dikky98@gmail.com.

Loso Judijanto
IPOSS Jakarta Indonesia, South Jakarta City, Special Capital Region of Jakarta, Indonesia.
Email: osojudijantobumn@gmail.com.

Mitranikasih Laia
Universitas Nias Raya, South Nias Regency, North Sumatra Province, Indonesia.
Email: mitracheersly@uniraya.ac.id.

Dedy Irwan
Universitas Harapan Medan, Medan City, North Sumatra Province, Indonesia.
Email: ddirwan@gmail.com.

Received: September 28, 2024; Accepted: October 30, 2024; Published: December 1, 2024.

Abstract: The development of big data technology has generated huge volumes of diverse data, creating challenges in detecting anomalies that could potentially affect decision-making. This research aims to examine the application of artificial intelligence (AI) in detecting anomalies in big data systems to support faster, more accurate and effective decision-making. The approach used includes the integration of machine learning algorithms, such as classification-based detection, clustering, and deep learning, in identifying abnormal patterns in large datasets. The research method involves real-time dataset-based simulations by measuring the performance of AI models using accuracy, precision, recall, and F1-score metrics. The results show that the application of AI can significantly improve the anomaly detection capability compared to conventional methods, with an average accuracy of 92%.

Keywords: Artificial Intelligence; Anomalies; Big Data.

1. Introduction

The digital era has ushered in significant advancements in data collection, processing, and analysis, giving rise to what is now known as big data. Big data is characterized by its immense volume, diverse variety, and rapid velocity, making it an essential component in various sectors such as industry, finance, healthcare, and security. However, managing data on this massive scale presents unique challenges, particularly in the area of anomaly detection [1]. Anomalies, defined as data points that deviate from established patterns or behaviors, often indicate the presence of systemic issues, security vulnerabilities, or emerging opportunities that require immediate attention and action [2]. These anomalies, if not identified and addressed in a timely manner, can lead to critical errors in decision-making, financial loss, or even pose security risks.

Anomaly detection plays a central role in big data systems by enabling organizations to identify irregularities that might otherwise remain hidden. Effective anomaly detection supports decision-making by providing timely insights into issues that may impact business operations, financial stability, or public safety. Conventional anomaly detection methods, which typically rely on statistical analysis or rule-based approaches, often struggle to cope with the sheer scale and complexity of big data [3]. These traditional methods tend to be limited in their ability to process vast datasets quickly and accurately, especially when dealing with high-dimensional, unstructured, or noisy data.

Artificial Intelligence (AI) has emerged as a powerful tool to overcome these limitations. AI, particularly through machine learning (ML) and deep learning (DL) techniques, offers the potential to automatically analyze vast datasets, identify complex patterns, and predict potential anomalies with a high degree of accuracy [4]. Unlike conventional methods, AI-driven approaches can continuously learn from new data, adapt to evolving patterns, and improve their performance over time. This makes AI an ideal solution for anomaly detection in dynamic, large-scale environments where traditional techniques fall short. The integration of AI into anomaly detection systems brings several advantages. First, it significantly enhances the efficiency of data processing by automating the detection of anomalies, thereby reducing the time and effort required by human analysts. This is particularly important in sectors such as finance and cybersecurity, where real-time anomaly detection is critical. In the financial sector, for instance, AI can help identify fraudulent transactions or money laundering activities as they occur, enabling swift responses and mitigating financial risks [5]. In healthcare, AI-based anomaly detection can help identify irregularities in patient data, such as early signs of diseases or medical errors, potentially saving lives and improving patient outcomes [6].

Despite its promise, the application of AI in anomaly detection is not without challenges. One of the primary hurdles is the optimization of algorithms to ensure they operate efficiently and effectively at scale. The computational demands of processing large datasets, particularly with complex algorithms like deep learning, can be substantial and may require high-performance computing resources [7]. Moreover, ensuring the interpretability of AI models remains a significant concern. While AI can achieve high levels of accuracy, understanding the reasoning behind its predictions is often difficult, which may reduce trust and hinder its adoption in sectors where transparency and accountability are paramount.

Furthermore, the success of AI-based anomaly detection depends heavily on the quality and diversity of the data it is trained on. Biased, incomplete, or unrepresentative data can lead to inaccurate predictions and missed anomalies, which could have serious implications in critical fields such as healthcare or security [7]. Therefore, careful attention must be paid to the design and implementation of AI models, including the validation of training datasets, to ensure that they are robust, unbiased, and capable of handling a wide range of data variations. This study seeks to explore the application of AI in detecting anomalies within big data systems to support more effective and data-driven decision-making [8]. By focusing on the integration of efficient and adaptive algorithms, the research aims to identify best practices for implementing AI in real-world anomaly detection scenarios. Additionally, this study will examine the potential limitations and challenges of AI applications, offering solutions to address these issues and improve the overall effectiveness of anomaly detection systems in diverse sectors. Ultimately, this research is expected to make a significant contribution to advancing AI technology and provide practical, scalable solutions to the growing challenges associated with big data processing and anomaly detection [9].

2. Research Method

This study employs a quantitative approach with an experimental method to assess the effectiveness of applying Artificial Intelligence (AI) in anomaly detection within big data systems [10]. The research process involves several key stages: system design, data collection, data processing, AI model implementation, testing,

and performance evaluation. Each stage is elaborated to ensure the successful application and testing of the anomaly detection model. In the system design phase, the architecture for anomaly detection in big data is developed. The designed system includes key components such as a big data management platform, utilizing technologies like Apache Hadoop or Apache Spark to handle large-scale data. Additionally, AI models are integrated using machine learning frameworks like TensorFlow or PyTorch to build anomaly detection models. The system architecture is designed with a focus on data processing efficiency and the ability to handle real-time data.

The data collection phase involves two main sources. First, publicly available datasets are used, including large-scale datasets that contain real anomalies, such as the KDD Cup dataset or financial transaction datasets. Second, synthetic data is generated through simulation by inserting anomaly patterns into a normal dataset to test the model's generalization ability. The collected data is then split into three parts: training data (70%), validation data (15%), and testing data (15%). In the data processing stage, several critical steps are performed. First, data preprocessing is carried out, including data cleaning, handling missing values, and normalization to ensure the data is in a suitable format for analysis. Second, feature extraction is performed using techniques such as Principal Component Analysis (PCA) to reduce data dimensions without losing important information. The next stage is the implementation of AI models, where various machine learning and deep learning algorithms are applied for anomaly detection. These include K-Means Clustering for group-based anomaly detection, Isolation Forest for outlier identification, and Autoencoder Neural Networks to detect abnormal patterns in complex data. The parameters for each model are optimized using techniques such as grid search and cross-validation to achieve the best performance in anomaly detection.

3. Result and Discussion

3.1 Results

This study provides an evaluation of the application of Artificial Intelligence (AI) in detecting anomalies within big data systems. The assessment is based on the performance of several AI algorithms: K-Means Clustering, Isolation Forest, and Autoencoder Neural Networks. The test results demonstrate the effectiveness of each algorithm in anomaly detection using evaluation metrics: accuracy, precision, recall, F1-score, and processing time.

Table 1. Algorithm Testing Using Dataset

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
K-Means Clustering	85.3	80.5	78.6	79.5
Isolation Forest	92.4	90.7	89.3	90.0
Autoencoder	94.1	92.5	91.8	92.1

K-Means Clustering showed a fairly good result with an accuracy of 85.3%. However, this algorithm has limitations in detecting anomalies in high-dimensional datasets, as indicated by the relatively lower recall value (78.6%). This is due to the algorithm's limitations in handling unstructured data with complex patterns. Isolation Forest outperforms K-Means Clustering, achieving an accuracy of 92.4% and an F1-score of 90.0%. This algorithm performs well in detecting outliers with relatively fast processing time (12 seconds). It is well-suited for big data with clear anomaly patterns.

Table 2. K-Means Clustering Performance

Parameter	Value (%)	Analysis
Accuracy	85.3%	Indicates the overall ability of the algorithm to classify data correctly.
Recall	78.6%	Relatively low, indicating that the algorithm is less effective in detecting all anomalies.
Limitation	High dimensionality	Performance decreases with high-dimensional datasets due to more complex data patterns.
Cause of Limitation	Complex patterns	The algorithm's limitation in handling unstructured data with complex patterns.
Context of Use	Simple datasets	Better suited for datasets with simple structures and patterns.

Autoencoder demonstrated the best performance with an accuracy of 94.1% and an F1-score of 92.1%. The strength of Autoencoder lies in its ability to understand complex patterns through deep neural networks.

However, its processing time is higher (18 seconds), which can be a challenge when applied in real-time systems.

Table 3. Autoencoder Neural Networks Performance

Parameter	Value (%)	Analysis
Accuracy	94.1%	Shows the best performance in classifying data correctly overall.
F1-Score	92.1%	Indicates high capability in detecting anomalies without sacrificing precision or recall.
Strength	Complex patterns	Ability to understand complex data patterns through deep neural networks.
Processing Time	18 seconds	Relatively longer compared to other algorithms, posing a challenge for real-time applications.
Context of Use	Complex datasets	Well-suited for datasets with unstructured or highly complex patterns.
Limitation	Time efficiency	Higher processing time requires optimization for real-time systems.

Based on the research findings, Autoencoder Neural Networks is the most effective algorithm for anomaly detection in big data systems, particularly for datasets with complex patterns. Isolation Forest excels in processing efficiency, making it a good choice for systems requiring quick response times. K-Means Clustering, while efficient for smaller datasets, is less optimal for large datasets with high complexity.

Table 4. Algorithm Implementation Recommendations

System Criteria	Recommended Algorithm	Reason
Datasets with complex patterns	Autoencoder Neural Networks	Ability to understand data with high complexity.
Need for quick response times	Isolation Forest	Fast processing time with good performance.
Simple or low-dimensional datasets	K-Means Clustering	Efficient for datasets with simple structures.

3.2 Discussion

The results of this study highlight the varying effectiveness of three AI algorithms—K-Means Clustering, Isolation Forest, and Autoencoder Neural Networks—in anomaly detection within big data systems. Each algorithm demonstrated specific strengths and weaknesses depending on the complexity of the data and the application requirements, such as processing time, accuracy, and the ability to handle high-dimensional data. K-Means Clustering, with an accuracy of 85.3%, proved to be effective for simpler, low-dimensional datasets but struggled when applied to high-dimensional data. Its relatively low recall (78.6%) indicates that it missed some anomalies, especially in datasets with complex or unstructured patterns. This is due to K-Means' reliance on Euclidean distance, which is less effective for non-linear data relationships, thus limiting its ability to detect anomalies in more complex big data environments. Therefore, while K-Means is efficient and quick for small datasets with simple patterns, its application in large-scale or complex datasets is limited.

In comparison, Isolation Forest outperformed K-Means with an accuracy of 92.4% and an F1-score of 90.0%. This algorithm demonstrated high efficiency in detecting outliers, particularly in large, high-dimensional datasets. Isolation Forest works by isolating anomalies instead of profiling normal data, which makes it particularly effective for anomaly detection in big data with clear patterns. Additionally, it required only 12 seconds of processing time, making it a suitable choice for applications where fast decision-making is necessary, such as fraud detection in financial transactions or intrusion detection in cybersecurity. Its ability to deliver high performance with relatively low computational cost makes it a strong candidate for real-time applications where speed is critical.

Autoencoder Neural Networks achieved the highest performance in this study, with an accuracy of 94.1% and an F1-score of 92.1%. Autoencoders are particularly effective in detecting anomalies in complex, high-dimensional datasets, as they are capable of learning non-linear patterns through deep neural networks. This makes them ideal for applications involving unstructured or intricate data patterns. However, the main drawback of Autoencoders is their relatively high processing time, taking 18 seconds per dataset, which could be a limitation for real-time anomaly detection systems. Despite this, Autoencoders are particularly useful in fields like healthcare, where early detection of anomalies in medical data is crucial for patient safety and outcomes. The increased processing time, however, would require optimization for real-time deployment.

In terms of practical application, the study indicates that the choice of algorithm should be based on the specific needs of the system. For datasets with simple structures and low dimensionality, K-Means Clustering remains an efficient choice. However, for systems requiring high accuracy and the ability to handle complex, high-dimensional data, Autoencoder Neural Networks are the most effective option, though further optimization would be needed for real-time applications. Isolation Forest strikes a balance between accuracy and processing time, making it ideal for real-time anomaly detection in big data systems, particularly in dynamic fields such as finance or cybersecurity. Thus, depending on the characteristics of the dataset and the operational requirements, each algorithm offers valuable strengths for anomaly detection.

4. Related Work

Anomaly detection in big data systems has become an essential research area, driven by the rapid expansion in data volume, variety, and velocity. The challenges posed by the complexity of big data demand advanced techniques for identifying anomalies, as traditional methods such as K-Means Clustering and statistical approaches often fail in high-dimensional or unstructured datasets. While these conventional techniques are effective in environments with clearly defined patterns, they struggle when applied to noisy or complex data, where patterns may not be easily recognized [11][12][13]. Recent advances in machine learning and deep learning have opened new avenues for tackling these challenges. Deep learning models, particularly Convolutional Neural Networks (CNNs) and Generative Adversarial Networks (GANs), have demonstrated strong performance in detecting anomalies across various domains, such as industrial systems and Internet of Things (IoT) applications. GANomaly, for instance, uses adversarial training in a semi-supervised manner to improve anomaly detection performance, achieving promising results on diverse datasets [14]. Likewise, Lai *et al*. applied CNNs to detect industrial anomalies, specifically network attacks in SCADA systems, and obtained high accuracy [15]. These methods highlight the trend of leveraging deep learning to automatically extract critical features from data, leading to improved anomaly detection in complex, high-dimensional environments [16].

Data preprocessing plays a significant role in enhancing the effectiveness of anomaly detection systems. Kaya *et al*. (2021) pointed out the importance of smart data filtering, particularly in managing large health data, which reduces computational load while improving accuracy [18]. Kulanuwat *et al*. (2021) combined statistical anomaly detection with machine learning-based data imputation to address challenges in hydrological time series data [19]. These hybrid methods have proven effective in improving anomaly detection and supporting better decision-making in big data systems. Hybrid and ensemble approaches have also gained attention in recent years. By combining multiple algorithms, these methods leverage the strengths of different models to improve overall performance. Integrating K-Means clustering with Isolation Forest can combine the benefits of both clustering and outlier detection. Furthermore, integrating Autoencoders with clustering techniques or ensemble classifiers helps to overcome the limitations of individual models, particularly in datasets with complex patterns.

Real-time anomaly detection is another growing focus, especially for applications that require fast decision-making, such as fraud detection, intrusion detection, and health monitoring. Algorithms for real-time detection must not only be accurate but also operate with low latency. Incremental learning techniques, which allow models to adapt as new data arrives, are particularly useful in these scenarios. These methods are suited for streaming data applications, where continuous updates and quick response times are critical. The use of artificial intelligence for anomaly detection in big data systems continues to evolve. While traditional methods like K-Means and Isolation Forest remain useful in certain applications, deep learning models such as Autoencoders and hybrid approaches are proving more effective for complex, high-dimensional data. Challenges related to computational efficiency and real-time processing still persist, but ongoing research aims to address these issues by enhancing scalability, speed, and overall performance [20][17]. Developing more efficient models will be crucial in improving anomaly detection systems, making them suitable for a wide range of big data applications [19][20].

5. Conclusion and Recommendations

Based on the research results, several main conclusions can be summarized as follows. The three AI algorithms tested in this study, namely K-Means Clustering, Isolation Forest, and Autoencoder Neural Networks, showed differences in performance in detecting anomalies. Autoencoder Neural Networks showed

the best performance with an accuracy of 94.1% and an F1-score of 92.1%, making it a very effective choice for detecting anomalies in datasets with complex patterns. Meanwhile, Isolation Forest recorded an accuracy of 92.4% and a relatively fast processing time, making it suitable for applications that require real-time responses. On the other hand, K-Means Clustering showed lower accuracy (85.3%) and had limitations in handling datasets with high dimensions or unstructured data patterns. Each algorithm has its advantages and limitations. Autoencoder Neural Networks excels in handling complex data patterns thanks to its ability to understand non-linear patterns through deep neural networks. However, the higher processing time (18 seconds) is a challenge in real-time applications, so further optimization is needed to improve time efficiency. Isolation Forest, with its time efficiency advantage (12 seconds), offers high accuracy, making it a good choice for applications with fast processing needs. However, this algorithm may be less effective if the anomalous pattern in the dataset is very complex. K-Means Clustering, although simple and efficient for datasets with clear and structured patterns, is less suitable for handling datasets with high dimensions or more complicated data patterns, due to its limitations in capturing non-linear relationships.

Based on these findings, recommendations for the use of algorithms for anomaly detection in big data systems can be adjusted to the characteristics and needs of the application. Autoencoder Neural Networks are recommended for use on datasets that have complex patterns and require in-depth analysis. Isolation Forest is more suitable for applications that prioritize fast response time with relatively simple data patterns. Meanwhile, K-Means Clustering can be applied to datasets with simple structured patterns to achieve computational efficiency.

References

[1] Arie, A. P. P. (2024). Transformasi akuntansi di era big data dan teknologi artificial intelligence (AI). *Jurnal Cahaya Mandalika*, 5(2), 937–943. https://doi.org/10.36312/JCM.V5I2.3279

[2] Caseba, F. L., & Dewayanto, T. (2024). Penerapan artificial intelligence, big data, dan blockchain dalam fintech payment terhadap risiko penipuan komputer (computer fraud risk): A systematic literature review. *Diponegoro Journal of Accounting, 13*(3).

[3] Dewi, F. S., & Dewayanto, T. (2024). Peran big data analytics, machine learning, dan artificial intelligence dalam pendeteksian financial fraud: A systematic literature review. *Diponegoro Journal of Accounting, 13*(3).

[4] Syamsu, M., Terisia, V., & Yusuf, D. (2022). Penerapan model infrastruktur artificial intelligence sebagai penggerak industri 4.0. *Jurnal Teknologi Informasi (JUTECH)*, 3(1), 1–14. https://doi.org/10.32546/JUTECH.V3I1.2375

[5] Abirami, S., Pethuraj, M., Uthayakumar, M., & Chitra, P. (2024). A systematic survey on big data and artificial intelligence algorithms for intelligent transportation system. *Case Studies in Transport Policy*, 17, 101247. https://doi.org/10.1016/J.CSTP.2024.101247

[6] Reka, S. S., Dragicevic, T., Venugopal, P., Ravi, V., & Rajagopal, M. K. (2024). Big data analytics and artificial intelligence aspects for privacy and security concerns for demand response modelling in smart grid: A futuristic approach. *Heliyon*, 10(15), e35683. https://doi.org/10.1016/J.HELIYON.2024.E35683

[7] Kamyab, H., *et al*. (2023). The latest innovative avenues for the utilization of artificial intelligence and big data analytics in water resource management. *Results in Engineering*, 20, 101566. https://doi.org/10.1016/J.RINENG.2023.101566

[8] Jiao, Z., Ji, H., Yan, J., & Qi, X. (2023). Application of big data and artificial intelligence in epidemic surveillance and containment. *Intelligent Medicine*, 3(1), 36–43. https://doi.org/10.1016/J.IMED.2022.10.003

[9] Papachristou, N., *et al*. (2023). Digital transformation of cancer care in the era of big data, artificial intelligence, and data-driven interventions: Navigating the field. *Seminars in Oncology Nursing*, 39(3), 151433. https://doi.org/10.1016/J.SONCN.2023.151433

[10] Lasisi, M., Kolade, K., & Rotimi, O. (2025). Big data. In *Encyclopedia of Libraries, Librarianship, and Information Science* (pp. 19–25). https://doi.org/10.1016/B978-0-323-95689-5.00269-8

[11] Hang, F., Xie, L., Zhang, Z., Guo, W., & Li, H. (2024). Research on the application of network security defense in database security services based on deep learning integrated with big data analytics. *International Journal of Intelligent Networks*, 5, 101–109. https://doi.org/10.1016/J.IJIN.2024.02.006

[12] Habeeb, R., Nasaruddin, F., Gani, A., Hashem, M., Ahmed, E., & Imran, M. (2019). Real-time big data processing for anomaly detection: A survey. *International Journal of Information Management*, 45, 289-307. https://doi.org/10.1016/j.ijinfomgt.2018.08.006

[13] Khlevna, I., & Koval, B. (2022). Development of infrastructure for anomalies detection in big data. *Applied Aspects of Information Technology, 5*(4), 348-358. https://doi.org/10.15276/aait.05.2022.23

[14] Akçay, S., Atapour-Abarghouei, A., & Breckon, T. (2019). GANomaly: Semi-supervised anomaly detection via adversarial training. In *Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision* (pp. 622-637). https://doi.org/10.1007/978-3-030-20893-6_39

[15] Lai, Y., Zhang, J., & Liu, Z. (2019). Industrial anomaly detection and attack classification method based on convolutional neural network. *Security and Communication Networks, 2019*, 1-11. https://doi.org/10.1155/2019/8124254

[16] Pang, G., Shen, C., Cao, L., & Hengel, A. (2021). Deep learning for anomaly detection. *ACM Computing Surveys*, 54(2), 1-38. https://doi.org/10.1145/3439950

[17] Munir, M., Siddiqui, S., Dengel, A., & Ahmed, S. (2019). DeepAnt: A deep learning approach for unsupervised anomaly detection in time series. *IEEE Access*, 7, 1991-2005. https://doi.org/10.1109/access.2018.2886457

[18] Kaya, S., Erdem, A., & Gunes, A. (2021). A smart data pre-processing approach to effective management of big health data in IoT edge. *Smart Homecare Technology and Telehealth*, 8, 9-21. https://doi.org/10.2147/shtt.s313666

[19] Kulanuwat, L., Chantrapornchai, C., Maleewong, M., Wongchaisuwat, P., Wimala, S., Sarinnapakorn, K., & Boonya-aroonnet, S. (2021). Anomaly detection using a sliding window technique and data imputation with machine learning for hydrological time series. *Water, 13*(13), 1862. https://doi.org/10.3390/w13131862

[20] Maurya, C. (2022). Anomaly detection in big data. https://doi.org/10.48550/arxiv.2203.01684.

