

Real-Time Anomaly Detection Business Process for Industrial Equipment Using Internet of Things and Unsupervised Machine Learning Algorithms

Emrullah Gultekin(✉) and Mehmet S. Aktas

Computer Engineering Department, Yildiz Technical University, Istanbul, Türkiye
emrullah.gultekin@std.yildiz.edu.tr, aktas@yildiz.edu.tr

Abstract. Today, it is evident that the Internet of Things, real-time data processing, and artificial intelligence technologies are essential in industrial settings to enable early warning autonomous anomaly detection systems. Such systems can detect anomaly situations that could cause failures shortly after they occur, allowing necessary maintenance to be performed promptly. In this research, a software platform has been designed and developed to collect data from sensors placed on industrial equipment to monitor their condition using the required IoT infrastructure. The real-time data collected from this platform is analyzed using real-time data processing techniques. Here, a business process is introduced for instant anomaly detection using real-time clustering analysis methods. To validate the proposed business process architecture, a prototype software has been developed, and its ability to detect anomaly situations has been evaluated. The results show that the proposed business process architecture is effective in real-time anomaly detection and can successfully detect anomalies that can lead to industrial equipment failures.

Keywords: internet of things · anomaly detection · artificial intelligence · clustering · real-time unsupervised machine learning · real-time streaming-data prediction

1 Introduction

The way we connect with the world around us has changed as a result of the Internet of Things (IoT). IoT devices provide enormous amounts of data in real-time, which may be leveraged to get insightful knowledge and improve decision-making. Anomalies or outliers in this data may, however, also be a sign of issues or security breaches. Real-time detection of these anomalies is essential to avoiding infrastructure damage, income loss, and other undesirable effects.

In today's highly competitive industrial landscape, businesses are under increasing pressure to maintain their operational efficiency and profitability.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2023
O. Gervasi et al. (Eds.): ICCSA 2023 Workshops, LNCS 14108, pp. 16–31, 2023.
https://doi.org/10.1007/978-3-031-37117-2_2

Industrial equipment such as machines, pumps, and turbines play a critical role in the smooth functioning of manufacturing processes, and any downtime or malfunction can have a significant impact on productivity and profitability. Real-time anomaly detection is an essential tool for businesses to proactively detect and address issues before they result in equipment failure or downtime.

The advent of the IoT and Artificial Intelligence (AI) technologies has transformed the industrial equipment landscape, allowing businesses to monitor and analyze equipment data in real-time. By leveraging IoT sensors, businesses can collect vast amounts of data on equipment performance, which can then be analyzed using AI algorithms to detect anomalies and predict potential issues.

Unsupervised learning techniques have become a promising answer to these problems. Unsupervised learning techniques may adapt to changes in the data distribution without needing labeled data, which is crucial in dynamic contexts where the data's features can vary over time.

There are various difficulties with real-time anomaly detection on streaming IoT data. First, it is challenging to analyze IoT data using conventional statistical techniques since it is frequently noisy, high-dimensional, and missing values. Second, algorithms that can process massive amounts of data rapidly and effectively are needed since streaming data must be analyzed in real-time. Finally, as anomalies are uncommon occurrences, it may be challenging to train a supervised learning model to recognize them because the training data may not adequately represent them.

This paper presents a real-time anomaly detection business process for industrial equipment using IoT and unsupervised machine learning algorithms. The proposed business process involves collecting data from IoT sensors on equipment performance, preprocessing the data to extract relevant features, and using clustering algorithms to detect anomalies in real-time. The detected anomalies are then sent to a notification system for further analysis and action.

The subsequent sections of this paper are organized as follows: We initiated our investigation by identifying pertinent research issues. In Sect. 3, we conducted a thorough review of the relevant literature. The proposed methodology was introduced in Sect. 4. In Sect. 5 we present the prototype application and evaluation. Finally, in Sect. 6, we provide a succinct summary of the research findings.

2 Research Problem

The task of real-time anomaly detection on streaming IoT data is critical and requires advanced techniques that can handle the vast and rapidly changing data. Real-time unsupervised machine learning methods have shown promise in this area and can be used to improve decision-making and prevent negative outcomes in IoT applications. However, further research is needed to develop more sophisticated real-time unsupervised learning techniques that can handle the complexity and diversity of IoT data, as well as to evaluate their effectiveness in practical settings.

Four key research questions that need to be addressed in this context are:

(1) What is a good software architecture design of a unsupervised machine learning based business process for detecting anomalies, which can lead to failures of industrial machines, on the streaming data coming from IoT devices?
(2) Can clustering algorithms be used effectively for real-time anomaly detection in streaming IoT data?
(3) How can we perform real-time anomaly detection on streaming unlabeled data in an IoT system?
(4) Which real-time unsupervised machine learning algorithms are best suited for detecting anomalies in streaming IoT data?

Answering these questions will enable researchers to develop more effective and efficient anomaly detection systems for IoT applications.

3 Related Work

Real-time anomaly detection on streaming IoT data has grown in significance due to the proliferation of IoT devices and the increasing amount of data they generate. The identification of anomalies in streaming IoT data has been successfully accomplished through the use of supervised and unsupervised learning techniques. This paper focuses on the significance of real-time anomaly detection in IoT, the challenges it poses, and potential solutions based on unsupervised learning.

Real-time anomaly detection in unsupervised machine learning, particularly in the context of clustering, has been an active research area in recent years. Several methods and algorithms have been proposed to address the challenges of real-time anomaly detection in clustering. In this section, we review some of the relevant literature on real-time anomaly detection methods for unsupervised machine learning, specifically in the context of clustering.

The utilization of streaming data in numerous real-world applications, such as e-commerce, banking, sensor data, and telecommunications logs, has led to an increasing interest in the field of data stream mining [4]. In contrast to batch classification, the classification of data streams differs in terms of how learning and predictions occur. The classifier in data stream classification uses the existing model to predict at any time with continuously streaming data [5].

In our previous work, we proposed a workflow architecture utilizing stream-based machine learning algorithms to provide predictive maintenance in IoT systems [1]. The proposed workflow employs various supervised machine learning algorithms, including Adaptive Random Forest, Hoeffding Tree, Leveraging Bagging, SPegasos, and Single Drift Classifier classification algorithms. A prototype application is provided to demonstrate the utility of the proposed workflow, and a prototype experiment is presented to investigate the efficacy of the prediction. In this study, real-time anomaly detection was performed in IoT streaming data using unsupervised machine learning techniques.

Mollaoglu et al. [2] suggested an approach for identifying fraud situations using tagged customer data in telecommunications, utilizing machine learning

algorithms and big data analytics platforms. In contrast, we provide a full solution that leverages machine learning for anomaly detection in the IoT space, with distinct dataset, algorithm, and success criteria.

Kayacik et al. [3] presented a methodology for anticipating anomalous behavior in real-time streaming customer data, utilizing supervised machine learning techniques in the communications field. In comparison, we present a comprehensive business process that applies unsupervised machine learning to IoT real-time anomaly detection, with distinct dataset, techniques, and success criteria.

Liang et al. [12] suggested a bearing health indicator using a recurrent neural network to estimate the residual life values of bearings. Our research has a similar objective but with a distinct methodology and structure, focusing on preventive maintenance using IoT systems and employing unsupervised machine learning techniques to detect anomalies in real-time unlabeled data.

Akbar et al. [6] proposed a proactive IoT architecture for predicting complex events using real-time data through Complex Event Processing (CEP) and machine learning. In contrast, our work focuses on a methodology for predictive maintenance of IoT systems using unsupervised machine learning algorithms to ensure the system's reliability.

Certain studies concentrate on the analysis of provenance data, which encompasses the interactions between user agents, data, and processes. Such data is employed to extract valuable insights from these interactions in diverse domains, including social computing [11], social networking [13,14], numerical weather prediction [10], and others. However, our investigation diverges from these studies as we analyze streaming IoT data for the purpose of anomaly detection using unsupervised machine learning algorithms.

Online clustering algorithms update the cluster model continuously as new data points arrive, while also allowing for the removal of outdated data points to adapt to changing data patterns. CluStream is a popular online clustering algorithm that uses micro-clusters to represent the approximate distribution of data points and identifies anomalies based on the deviation of data points from the micro-cluster centroids. Aggarwal et al. proposed a framework for clustering evolving data streams, which includes online clustering algorithms, for real-time anomaly detection in data streams [14].

In the context of this research, a system design has been proposed that enables protective maintenance to be implemented utilizing Internet of Things technology for large physical machinery (such as motors). The proposed system uses real-time clustering analysis methods on data from IoT sensors to identify abnormal operating conditions that could result in the motors failing. We observe metadata management solutions for web service-based systems in the literature [16,17]. However, in this study, real-time system data analysis is used to find abnormal operating conditions. Additionally, we also observe studies that focus on distributed systems built on web services developed for a number of study fields [10,18–22,24,30,31,33]. In contrast to these studies, a restful web service-based system is being developed in this research to detect unexpected conditions in the operation of motors. Different data labeling methods have been developed

for usage data obtained during the use of systems in the literature [25,28,29]. In this research, the focus is on real-time analysis and interpretation of data obtained from sensors, and data labeling methods are not included. Among the studies in the literature, there are also those that focus on software quality for developed software [26,27]. However, software quality examination has not been performed in this study. When the literature is examined, it is also observed that studies have been conducted on real-time data analysis and detecting compliance with rules [15,32]. In this research, real-time data analysis is performed on data from IoT sensors to detect unexpected operational behaviors.

4 Proposed Methodology

In our study, we aim to detect and intervene in the errors of large industrial machines without causing any damage to the machines by using real-time unsupervised machine learning methods on unlabeled data. The goal is to detect and intervene in errors in real-time through the proposed process as shown in Fig. 1. Here, sensor environment data and sensor device data are first obtained from IoT devices equipped with a mobile operating system (such as Contiki NG). Then, the streaming data undergoes various operations and is clustered using clustering algorithms within a streaming machine learning library (such as the MOA framework). Data that does not belong to any cluster is considered anomalous after the clustering process.

In the current study, we aimed to complement our prior work by investigating a methodology for real-time anomaly detection using unsupervised machine learning methods. Specifically, the effectiveness of unsupervised machine learning algorithms was evaluated to detect anomalies in IoT systems. This stands in contrast to the previous work, which utilized supervised machine learning methods.

The proposed process consists of various modules, which are detailed below.

4.1 Feature Extraction Module

Real-time anomaly detection in IoT systems requires the extraction of relevant features from a variety of sensor data streams. These features are used to train a machine learning model to detect anomalies in the data and trigger alerts or responses in real-time.

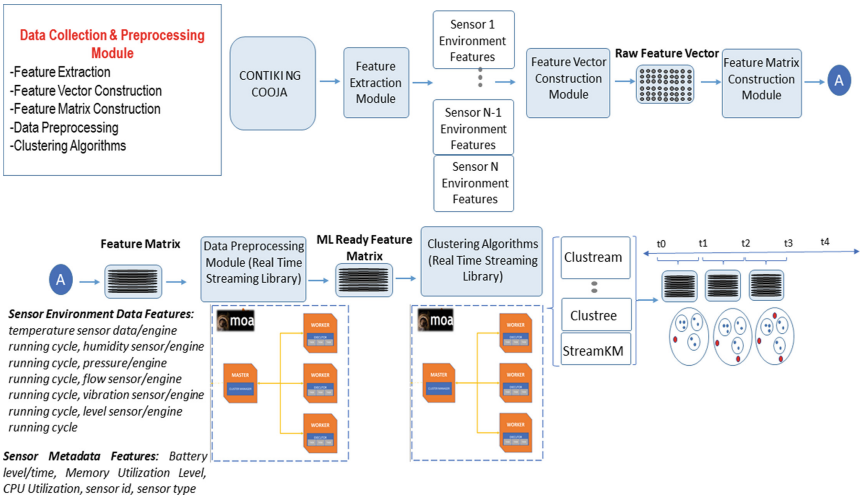


Fig. 1. Proposed Business Process

Domain-specific features such as temperature gradients, pressure differentials, humidity sensor values, flow sensor values, level sensor values and vibration patterns can be used to extract information specific to the IoT system being monitored. These features can be used to identify anomalies that are unique to the system or the environment it is operating in. In addition to these, metadata information related to the specific features of IoT devices can be added. Information such as battery level, memory status, CPU status, sensor ID, and sensor type can also be extracted.

4.2 Feature Vector Construction Module

In IoT systems, real-time anomaly detection requires the construction of a feature vector that captures the relevant features of the data. The feature vector is used as input to a machine learning algorithm to detect anomalous behavior in real-time.

The construction of a feature vector involves selecting and extracting the most relevant features from the data and combining them into a single vector. The feature vector should be constructed in a way that maximizes the accuracy and efficiency of the anomaly detection algorithm.

At this stage, the relevant features are identified, normalized, and combined. Additionally, dimensionality reduction and updating the feature vector are performed in this step.

4.3 Feature Matrix Construction Module

The construction of a feature matrix involves selecting and extracting the most relevant features from the data and organizing them into a matrix. The feature

matrix should be constructed in a way that maximizes the accuracy and efficiency of the anomaly detection algorithm. Once the feature matrix has been constructed, it can be used as input to a machine learning algorithm to detect anomalous behavior over time. The accuracy and efficiency of the algorithm will depend on the quality of the feature matrix, so it is important to carefully select and construct the features and organize them in a way that maximizes the performance of the algorithm.

4.4 Data Preprocessing Module

To achieve high accuracy in real-time anomaly detection, the data preprocessing step is crucial. Such systems require careful data preprocessing to ensure that the algorithm can accurately identify anomalous behavior. The data preprocessing step involves cleaning, transforming, and normalizing the data to prepare it for analysis.

4.5 Clustering Algorithms Module

For clustering we propose the MOA (Massive Online Analysis) framework which is a popular open-source platform for real-time machine learning and data stream mining. MOA provides a range of clustering algorithms that can be used for anomaly detection in IoT systems.

Here are some real-time clustering algorithms that can be used for anomaly detection using streaming based machine learning library (i.e. the MOA library):

CluStream Clustering: CluStream is a clustering algorithm specifically designed to handle data streams that exhibit concept drift. It operates on two levels of clustering, macro-clusters, and micro-clusters, where macro-clusters represent the overall data distribution, and micro-clusters represent local regions in the data distribution. By maintaining micro-clusters, CluStream can detect changes in the data distribution and update macro-clusters accordingly, making it suitable for handling concept drift.

ClusTree Clustering: ClusTree is a hierarchical clustering algorithm that performs online, incremental hierarchical clustering of evolving data streams. It handles noisy, high-dimensional data by using a projection-based clustering approach that maps data into a lower-dimensional subspace. ClusTree maintains a tree structure of clusters, where each node represents a cluster of data points. The tree is incrementally updated by merging nodes as new data arrives, and the height of the tree can be adjusted to control the granularity of the clustering.

StreamKM Clustering: StreamKM is an assignment-based streaming clustering algorithm and a streaming variant of the k-means clustering algorithm. It updates the cluster centers incrementally and maintains a set of potential outliers, which can detect concept drift or outliers in the stream. StreamKM is efficient in the data stream setting, where data arrives continuously and can only be processed once.

These clustering algorithms can be used in combination with the streaming based machine learning libraries (such as the MOA library). In turn, this will lead

to incremental learning capabilities to detect anomalous behavior in real-time. The choice of clustering algorithm will depend on the nature of the data and the requirements of the anomaly detection task. It is important to carefully evaluate the performance of different clustering algorithms and tune their parameters to achieve the best results.

In the proposed software architecture, we indicate that any streaming based machine learning library that can support aforementioned streaming based clustering algorithms can be utilized. In the prototype implementation of the proposed architecture, we use the MOA framework for the streaming ML library to be able to cluster real-time data and detect anomalies in the running observation data of the industrial equipment. We use the Clustream, Clustree and StreamKM clustering algorithms available in the MOA library to cluster the data and identify outliers that do not belong to any cluster as anomalies. This approach will enable us to detect anomalies in real-time and take appropriate actions to prevent any potential damage. The proposed business process described in this section, addresses research problem-1 presented in Sect. 2.

5 Prototype Application and Evaluation

We developed a prototype version of the proposed methodology to assess its effectiveness. We performed performance testing on the prototype and evaluated the suggested methodology using an actual dataset. Specifically, we utilized the NASA dataset [7], which leveraged C-MAPSS to simulate engine degradation under different operating conditions and failure modes. The dataset comprises four subsets, each of which represents a unique combination of operating conditions and failure modes. The dataset does not have any labeled data.

To construct a model in a big data platform environment, we first labeled the data using clustering algorithms. By passing through this model, incoming data in the streaming data is determined whether it belongs to any existing clusters. If the data does not belong to any cluster, it is considered an anomaly.

We employed the MOA framework (version 21.07), which is a big data processing framework, to train and test real-time streaming IoT data. We implemented unsupervised machine learning algorithms such as Clustream, Clustree, and StreamKM. This procedure was performed on a machine with 4 Core CPUs, 16GB RAM and 2.9 GHz.

In this study, we focus on detecting anomalies in streaming IoT data using unsupervised machine learning techniques, which do not require labeled data and are well-suited for real-time detection. These techniques are particularly useful in dynamic contexts where the features of the data can change over time and require adaptation to changes in data distribution.

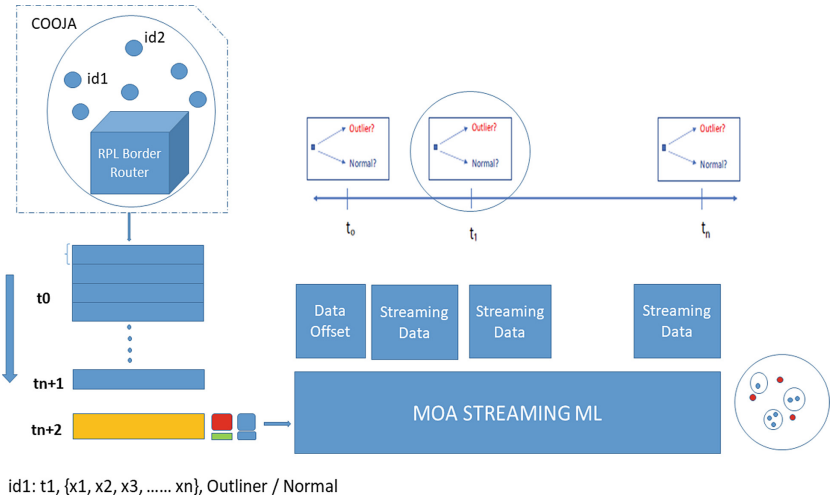


Fig. 2. Proposed Data Processing and Analysis Process

In alignment with our previous work, the data used in this study originates from IoT devices that run on the Contiki NG operating system. To simulate the system, we used Cooja, a network simulator. Contiki NG is specifically designed for wireless sensor networks. For hosting the Contiki NG operating system, we employed Docker containers. NodeJS and REST architecture services were utilized to present data to the external environment. We saved the NASA Turbofan Jet Engine data as if it originated from this system in a PostgreSQL database. The application retrieves this data from the database and sends it to the model as streaming data.

5.1 Dataset

For testing our proposed process, we utilized the Nasa Turbofan Jet Engine dataset, which contains a larger number of records and presents a challenging problem for anomaly detection. The dataset is part of NASA's C-MAPSS simulation tool for turbofan engines and includes fourteen input parameters that allow simulation of a variety of operating behaviors [8]. The dataset comprises 61249 running observation data of industrial equipment (i.e. NASA turbo fan engine). Each running observation data consists of attributes, IoT sensor environment data, describing the running performance of the motor. The data includes anomalies (varying levels of deterioration over time) until the system failure [9].

The data are provided in a compressed text file containing 26 columns of separated integers, with each row representing a sampling of data collected during a single operational cycle and each column representing a unique variable.

5.2 Test Design

In order to test the prototype system on the aforementioned dataset, we first conducted data preprocessing (i.e. filtering) on the data. By filtering unnecessary columns and using clustering algorithms, we aim to identify anomalous patterns and improve the accuracy of predictive models.

The NASA Turbofan Jet Engine data is originally stored in text files, which are presented as four separate train and test data. In order to easily access and process the data, we transferred it to a relational database, specifically PostgreSQL. Some columns in the dataset do not experience significant changes and, thus, do not have much effect on the result. By filtering these columns, only the necessary fields were transferred to the model for improved efficiency.

We started the realtime clustering process with an offset dataset. First we feed the system with 35% offset data as shown in Fig. 2. As time progressed, the system was fed with subsequent data with 15% offset, and unsupervised machine learning algorithms were used to analyze the data. These algorithms separated the data into clusters, and any data not included in a cluster was considered an outlier. The number of outliers that occurred over time was calculated to evaluate the effectiveness of the algorithms.

To determine outliers for each clustering type, we applied the following methodology. For each data point, we tested whether it belongs to the current clusters by performing cluster membership testing for all clusters. We did this by measuring the distance between the data point and the centroid of the relevant cluster. If the distance value was above a certain threshold value, we considered the data point as not a member of the tested cluster. A data point that did not belong to any cluster using this membership test was classified as an outlier. Different threshold values were used for the distance-based membership test for different clustering algorithms. For CluStream, ClusTree, and StreamKM, these threshold values were chosen to be 35, 35, and 350, respectively. For outlier detection, the euclidean distance formula was used to determine the distances of the data to the clusters and compared with a threshold value. The v-measure values of these clusters were calculated to measure the quality of the clustering process.

With this experimental study test design, we aim to answer research problem 3, which we presented in Sect. 2.

5.3 Evaluation Result and Discussion

In this study, we evaluated the real-time anomaly detection process we proposed for industrial equipment using IoT and AI technologies. We worked with data from the NASA dataset, specifically the data numbered 4. In this dataset, the engines operate until they experience a fault. The main idea of our study is that by analyzing data from industrial machines, possible faults can be detected in advance and addressed through predictive maintenance. We aimed to detect anomalies before faults occur by analyzing unlabeled data using unsupervised machine learning techniques. Machines can exhibit many anomaly indicators

before experiencing a fault, and we expect these anomaly situations to exhibit an increasing trend close to the occurrence of a fault. In our study, we first train our model using the Clustream, Clustree, and StreamKM clustering algorithms in the MOA framework on the initial 35% offset section of 61,249 data points. Then, we input data in 15% sections into the system to detect anomalies. As the data comes in, we observe that anomalous situations not initially present in the system increase towards the end.

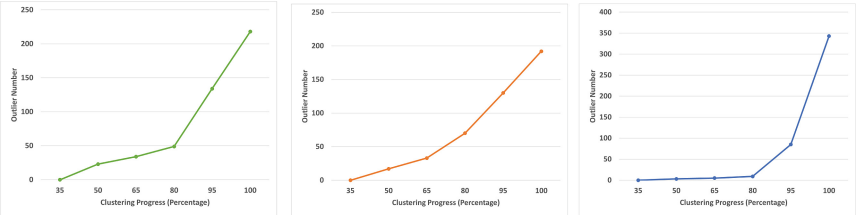


Fig. 3. The figure on the left indicates the clustream clustering algorithm results. The figure in the middle represents the clustree clustering algorithm result. The figure on the right shows the streamkm clustering algorithm results

Table 1 shows the anomaly detection data for the evaluated Clustream algorithm. As seen in the figure, while there are no outliers in the initial 35% offset data, 23, 34, 49, 134, and 218 outliers were detected in the 50%, 65%, 80%, 95%, and 100% data, respectively.

Table 2 contains the data for the Clustree algorithm. Similarly, in the results of this algorithm, although there were no outliers initially, 17, 33, 70, 130, and 192 outliers were detected in sequence. An increase in the number of outliers close to the occurrence of a fault was also observed here.

Table 3 contains the data for the Stream KM algorithm. Unlike other algorithms, it is necessary to determine the cluster count. For the detection of K cluster count, the Silhouette method was used, and the cluster number was determined as 3 for this data. As seen in the table, although there were no outliers initially, 3, 5, 9, 85, and 343 outliers were detected over time. An increase in the number of outliers towards the fault time is also noteworthy here. The results for the clustering algorithms presented in Fig. 3.

We used the V-Measure method to measure the quality of clustering results. It can be seen that the clustering performed with all three algorithms is of high quality according to the V-Measure values.

This results indicates that algorithms using IoT and AI technologies may be more suitable for real-time anomaly detection in industrial equipment. This section addresses research problems 2 and 4 presented in Sect. 2.

Overall, the proposed business process demonstrated good performance in detecting anomalies in real-time. The combination of IoT sensors, data preprocessing techniques, and unsupervised machine learning algorithms enabled the system to detect anomalies with a high degree of accuracy, minimizing the risk of equipment failure and downtime.

Table 1. Realtime Analysis Results of Clustream Clustering on IoT Sensor Data

Observation Number	Progress (Percentage)	Cluster Count	Outlier Number	V-Measure
21311	35	100	0	0.99460
30557	50	100	23	0.99557
39697	65	100	34	0.99704
48903	80	100	49	0.99630
58059	95	100	134	0.99703
61249	100	100	218	0.99737

Table 2. Realtime Analysis Results of Clustree Clustering on IoT Sensor Data

Observation Number	Progress (Percentage)	Cluster Count	Outlier Number	V-Measure
21311	35	352	0	0.98702
30557	50	314	17	0.99037
39697	65	330	33	0.99016
48903	80	316	70	0.99211
58059	95	331	130	0.99373
61249	100	280	192	0.99575

Table 3. Realtime Analysis Results of StreamKM Clustering on IoT Sensor Data

Observation Number	Progress (Percentage)	Cluster Count	Outlier Number	V-Measure
21311	35	3	0	0.96640
30557	50	3	3	0.99336
39697	65	3	5	0.96265
48903	80	3	9	0.99894
58059	95	3	85	0.99970
61249	100	3	343	0.99153

However, there were some limitations to our study. First, we used the NASA dataset, and the performance of the system may vary with different datasets. Second, we only evaluated a limited number of clustering algorithms, and there may be other algorithms that could perform better in detecting anomalies.

We also observe that as the number of outliers in the running observation data increases, the industrial equipment is more likely to fail. Hence, such real-time anomaly detection systems can prevent the system from failures by keeping track of the increasing anomaly behaviours.

The system can be customized for specific equipment types and integrated into existing equipment monitoring systems to improve operational efficiency and minimize downtime. Further research is needed to evaluate the system's performance with larger datasets and to explore the use of other clustering algorithms for anomaly detection.

6 Conclusion and Future Work

In this study, we proposed a real-time anomaly detection business process for predictive maintenance of industrial equipment. The proposed process utilizes IoT sensors and real-time unsupervised machine learning techniques, which do not require labeled data and are well-suited for real-time detection. The system uses IoT sensors to collect data from equipment in real-time and applies data preprocessing techniques and streaming-based clustering algorithms to detect anomalies. To facilitate testing of the proposed business process, we implemented a prototype application. The protoype application was evaluated using the NASA Turbofan Engine Degradation Simulation dataset, which contains sensor readings from aircraft engines that can be used to predict when an engine is likely to fail, in a streaming data fashion. The experimental study results are promising in detecting anomalies in real-time. The results also indicate that as the number of outliers in the running observation data of the industrial equipment increases, they are likely to fail. Hence, we claim that such real-time anomaly detection systems can prevent the industrial equipment from failures by keeping track of the increasing anomaly behaviours. We calculate the number of outliers over time using Clustream, Clustree, and StreamKM streaming clustering algorithms. The results show that the increase in the number of outliers over time is correlated with the trend of approaching an error condition. These anomaly situations can be detected before an error occurs, allowing for early intervention. These results demonstrate the potential of unsupervised machine learning techniques for real-time anomaly detection in IoT applications, and highlight the need for further research to develop more sophisticated techniques that can handle the complexity and variety of IoT data.

There are several areas for future work to improve the proposed real-time anomaly detection business process. First, the system could be evaluated with different datasets. Second, additional clustering algorithms could be evaluated to determine if they perform better than the algorithms used in this study. Third, the proposed system could be integrated with other AI technologies, such as deep learning or reinforcement learning, to further improve the accuracy of anomaly detection. Deep learning algorithms could be used to learn complex patterns in the data, while reinforcement learning could be used to optimize the clustering algorithms in real-time.

References

1. Gultekin, E., Aktas, M.S.: A business workflow architecture for predictive maintenance using real-time anomaly prediction on streaming IoT data. In: 2022 IEEE International Conference on Big Data (Big Data), Osaka, Japan (2022). https://doi.org/10.1109/bigdata55660.2022.10020384
2. Mollaoğlu, A., Baltaoğlu, G., Çakir, E., Aktaş, M.S.: Fraud detection on streaming customer behavior data with unsupervised learning methods. In: 2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), Kuala Lumpur, Malaysia, pp. 1–6 (2021). https://doi.org/10.1109/ICECCE52056.2021.9514152

3. Kayacik, A.F., Özcan, B., Baltaoğlu, G., Çakir, E., Aktaş, M.S.: Real-time fraud prediction on streaming customer-behaviour data. In: 2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), Kuala Lumpur, Malaysia, pp. 1–6 (2021). https://doi.org/10.1109/ICECCE52056.2021.9514169
4. Homayoun, S., Ahmadzadeh, M.: A review on data stream classification approaches. J. Adv. Comput. Sci. Technol. **5**, 8 (2016). https://doi.org/10.14419/jacst.v5i1.5225
5. Gomes, H.M., Bifet, A., Read, J., et al.: Adaptive random forests for evolving data stream classification. Mach. Learn. **106**, 1469–1495 (2017). https://doi.org/10.1007/s10994-017-5642-8
6. Akbar, A., Khan, A., Carrez, F., Moessner, K.: Predictive analytics for complex IoT data streams. IEEE Internet Things J. **4**(5), 1571–1582 (2017). https://doi.org/10.1109/JIOT.2017.2712672
7. Saxena, A., Goebel, K., Simon, D., Eklund, N.: Damage propagation 893 modeling for aircraft engine run-to-failure simulation. In: Proceedings of the International Conference 894 Prognostics Health Management, pp. 1–9 (2008)
8. Asif, O., Haider, S.A., Naqvi, S.R., Zaki, J.F.W., Kwak, K.-S., Islam, S.M.R.: A deep learning model for remaining useful life prediction of aircraft turbofan engine on C-MAPSS dataset. IEEE Access **10**, 95425–95440 (2022). https://doi.org/10.1109/ACCESS.2022.3203406
9. Muneer, A., Taib, S.M., Fati, S.M., Alhussian, H.: Deep-learning based prognosis approach for remaining useful life prediction of turbofan engine. Symmetry **13**, 1861 (2021). https://doi.org/10.3390/sym13101861
10. Tufek, A., Gurbuz, A., Ekuklu, O.F., Aktas, M.S.: Provenance collection platform for the weather research and forecasting model. In: 2018 14th International Conference on Semantics, Knowledge and Grids (SKG), Guangzhou, China, pp. 17–24 (2018). https://doi.org/10.1109/SKG.2018.00009
11. Riveni, M., Baeth, M.J., Aktas, M.S., Dustdar, S.: Provenance in social computing: A case study. In: 2017 13th International Conference on Semantics, Knowledge and Grids (SKG), Beijing, China, pp. 77–84 (2017). https://doi.org/10.1109/SKG.2017.00021
12. Guo, L., Li, N., Jia, F., Lei, Y., Lin, J.: A recurrent neural network based health indicator for remaining useful life prediction of bearings. Neurocomputing **240C**(2017), 98–109 (2017). https://doi.org/10.1016/j.neucom.2017.02.045
13. Tas, Y., Baeth, M.J., Aktas, M.S.: An approach to standalone provenance systems for big social provenance data. In: 2016 12th International Conference on Semantics, Knowledge and Grids (SKG), Beijing, China, pp. 9–16 (2016). https://doi.org/10.1109/SKG.2016.010
14. Aggarwal, C.C., Han, J., Wang, J., Yu, P.S.: A framework for clustering evolving data streams. In: Proceedings of the 29th International Conference on Very Large Data Bases - Volume 29 (VLDB '03), vol. 29. VLDB Endowment, pp. 81–92 (2003)
15. Baeth, M.J., Aktas, M.S.: An approach to custom privacy policy violation detection problems using big social provenance data. Concurr. Computat. Pract. Exp. **30**, e4690 (2018). https://doi.org/10.1002/cpe.4690
16. Aktas, M.S., Fox, G.C., Pierce, M., Oh, S.: XML metadata services. Concurr. Computat. Pract. Exp. **20**, 801–823 (2008). https://doi.org/10.1002/cpe.1276
17. Aktas, M.S., Pierce, M.: High-performance hybrid information service architecture. Concurr. Computat. Pract. Exp. **22**, 2095–2123 (2010). https://doi.org/10.1002/cpe.1557

18. Fox, G.C., et al.: Real time streaming data grid applications. Distributed Cooperative Laboratories: Networking, Instrumentation, and Measurements, pp. 253–267. Springer, Boston (2006). https://doi.org/10.1007/0-387-30394-4_17
19. Aktas, M.S., Fox, G.C., Pierce, M.: Fault tolerant high performance information services for dynamic collections of grid and web services. Future Gen. Comput. Syst. **23**(3), 317–337 (2007). https://doi.org/10.1016/j.future.2006.05.009
20. Aydin, G., Sayar, A., Gadgil, H., Aktas, M.S., Fox, G.C., Ko, S., Bulut, H., Pierce, M.E.: Building and applying geographical information system grids. Concurr. Computat. Pract. Exp. **20**, 1653–1695 (2008). https://doi.org/10.1002/cpe.1312
21. Fox, G.C., Aktas, M.S., Aydin, G., et al.: Algorithms and the grid. Comput. Visual Sci. **12**, 115–124 (2009). https://doi.org/10.1007/s00791-007-0083-8
22. Aktas, M., Aydin, G., Donnellan, A., et al.: iSERVO: Implementing the international solid earth research virtual observatory by integrating computational grid and geographical information web services. Pure Appl. Geophys. **163**, 2281–2296 (2006). https://doi.org/10.1007/s00024-006-0137-8
23. Aydin, G., Aktas, M.S., Fox, G.C., Gadgil, H., Pierce, M., Saya, A.: SERVOGrid complexity computational environments (CCE) integrated performance analysis. In: The 6th IEEE/ACM International Workshop on Grid Computing, 2005, Seattle, p. 6 (2005). https://doi.org/10.1109/GRID.2005.1542750
24. Pierce, M.E., et al.: The QuakeSim Project: Web services for managing geophysical data and applications. In: Tiampo, K.F., Weatherley, D.K., Weinstein, S.A. (eds) Earthquakes: Simulations, Sources and Tsunamis. Pageoph Topical Volumes. Birkhäuser Basel (2008). https://doi.org/10.1007/978-3-7643-8757-0_11
25. Uygun, Y., Oguz, R.F., Olmezogullari, E., Aktas, M.S.: On the large-scale graph data processing for user interface testing in big data science projects. In: 2020 IEEE International Conference on Big Data (Big Data), Atlanta, pp. 2049–2056 (2020). https://doi.org/10.1109/BigData50022.2020.9378153
26. Sahinoglu, M., Incki, K., Aktas, M.S.: Mobile application verification: A systematic mapping study. In: Computational Science and Its Applications - ICCSA 2015. ICCSA 2015. Lecture Notes in Computer Science, vol. 9159. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-21413-9_11
27. Kapdan, M., Aktas, M., Yigit, M.: On the structural code clone detection problem: A survey and software metric based approach. In: Computational Science and Its Applications - ICCSA 2014. ICCSA 2014. Lecture Notes in Computer Science, vol. 8583. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-09156-3_35
28. Olmezogullari, E., Aktas, M.S.: Pattern2Vec: Representation of clickstream data sequences for learning user navigational behavior. Concurr. Computat. Pract. Exp. **34**(9), e6546 (2022). https://doi.org/10.1002/cpe.6546
29. Olmezogullari, E., Aktas, M.S.: Representation of click-stream datasequences for learning user navigational behavior by using embeddings. In: 2020 IEEE International Conference on Big Data (Big Data), Atlanta, pp. 3173–3179 (2020). https://doi.org/10.1109/BigData50022.2020.9378437
30. Nacar, M.A., et al.: VLab: collaborative grid services and portals to support computational material science. Concurr. Computat.: Pract. Exper. **19**, 1717–1728 (2007). https://doi.org/10.1002/cpe.1199
31. Dundar, B., Astekin, M., Aktas, M.S.: A big data processing framework for self-healing Internet of Things applications. In: 2016 12th International Conference on Semantics, Knowledge and Grids (SKG), Beijing, China, pp. 62–68 (2016). https://doi.org/10.1109/SKG.2016.017

32. Baeth, M.J., Aktas, M.S.: Detecting misinformation in social networks using provenance data. In: 2017 13th International Conference on Semantics, Knowledge and Grids (SKG), Beijing, China, pp. 85–89 (2017). https://doi.org/10.1109/SKG.2017.00022
33. Aktas, M., et al.: Implementing geographical information system grid services to support computational geophysics in a service-oriented environment. In: NASA Earth-Sun System Technology Conference, University of Maryland, Adelphi, Maryland (2005)