Unsupervised Anomaly Detection in Data Quality Control

Lex Poon
Multiscale Networked Systems
University of Amsterdam
Amsterdam, The Netherlands
l.poon@uva.nl

Siamak Farshidi
Multiscale Networked Systems
University of Amsterdam
Amsterdam, The Netherlands
s.farshidi@uva.nl

Na Li
Multiscale Networked Systems
University of Amsterdam
Amsterdam, The Netherlands
n.li@uva.nl

Zhiming Zhao
Multiscale Networked Systems
University of Amsterdam
Amsterdam, The Netherlands
z.zhao@uva.nl

2021 IEEE International Conference on Big Data (Big Data) | 978-1-6654-3902-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/BigData52589.2021.9671672

***Abstract*—Data is one of the most valuable assets of an organization and has a tremendous impact on its long-term success and decision-making processes. Typically, organizational data error and outlier detection processes perform manually and reactively, making them time-consuming and prone to human errors. Additionally, rich data types, unlabeled data, and increased volume have made such data more complex. Accordingly, an automated anomaly detection approach is required to improve data management and quality control processes. This study introduces an unsupervised anomaly detection approach based on models comparison, consensus learning, and a combination of rules of thumb with iterative hyper-parameter tuning to increase data quality. Furthermore, a domain expert is considered a human in the loop to evaluate and check the data quality and to judge the output of the unsupervised model. An experiment has been conducted to assess the proposed approach in the context of a case study. The experiment results confirm that the proposed approach can improve the quality of organizational data and facilitate anomaly detection processes.**

***Index Terms*—data quality control, data quality assessment, unsupervised learning, anomaly detection, automated data quality control**

I. Introduction

Data-driven decision-making is at the center of modern enterprises and institutions. In other words, data guides business processes and decisions so that working with incorrect or missing data can lead to bad decisions and potentially harm the organization. The quality of data has effects across teams, and organizational boundaries, especially in large organizations with complex systems that result in complex data dependencies [1], and when organizations implement forecasting and Machine Learning (ML) models in their business process. These models are reliant on the input data as the assumptions made depend on this. Additionally, in modern information infrastructures, data lives in many places and comes in different formats. These sources do not support integrity constraints and data quality checks. Therefore, every team and system involved in data processing must somehow take care of data validation, resulting in tedious and repetitive work. Commons sources of errors are bugs in external data sources and data pre-processing code (e.g., when a data engineer accidentally changes a time measurement from seconds to milliseconds in a data-producing pipeline) [2].

Data are primarily unlabeled; thus, outliers in data are hard to detect and can undermine the quality and the decision-making. Accordingly, the unsupervised learning method for automated quality control methods should be considered. The current approaches in the literature for handling errors are reactively and manually, so they are time-consuming processes that lead to human errors. Furthermore, technological advances have made data management more complex and controversial. Accordingly, it is essential to automate the quality control process to reduce the cost of data monitoring and to improve their quality. Moreover, the need for a human-based approach needs to be limited as much as possible.

The main research question in this is: *how to enable unsupervised anomaly detection in validity and quality control for data management?* The following five research questions are formulated to answer the research question: (RQ_1) What are the challenges in automated data quality control? (RQ_2) How to assess if the data quality control is correct? (RQ_3) What are the existing automated solutions to data quality control? (RQ_4) How to implement an automated data quality control? (RQ_5) How to validate the quality control method?

In this study, we followed a mixed research method, a combination of qualitative and quantitative research, to systematically capture knowledge regarding anomaly detection approaches in the literature and make it available in a reusable and extendable format. First, we conducted an extensive literature study using snowball and descriptive methods to answer the first three research questions (RQ_1, RQ_2, and RQ_3). Next, we developed an unsupervised anomaly detection approach based on models comparison, consensus learning, and a combination of rules of thumb with iterative hyper-parameter tuning to increase data quality. To evaluate the effectiveness of the proposed approach and address the last two research questions (RQ_4 and RQ_5), a case study research in the context of an organization has been conducted.

This study makes four contributions to existing research: (C_1) Review state of the art for automated quality control approaches. (C_2) Propose a framework to enhance the current manually based quality control with a (semi)-automated anomaly detection pipeline. (C_3) Compare different algorithms for automated quality control. (C_4) Exploring ensemble learning-based approach.

II. Literature Study

This section elaborates on the literature review phase of the study to identify the state-of-the-art solutions besides interpretable patterns and gaps in the literature concerning existing propositions, methodologies, and findings.

A. Data quality definition

Based on the literature study phase of this research, we realized that the term *data quality* has different definitions and can be used interchangeably with *fitness for use* [3]. Wang et al. [3] proposed that data quality judgment depends on the data consumers. Other researchers refer to data quality as *the degree to which a set of inherent fulfill the requirements* [1], [4]. These two definitions are related by deriving the concept of "data quality" from the concept of "quality" and the suitability of the data for a particular use case.

Nikiforova [5] combined the definitions above and defines *data quality* as *the suitability of a given data set and its properties for a particular usage or use case, which depends on the data consumer using them, for example, in analytics, making business decisions, planning, etc.*. Low-quality data profoundly influence the quality of business processes. It is recognized as a relevant performance issue of operating processes of decision-making activities and of inter-organizational cooperation requirements [6]. The presence of incorrect or inconsistent data can significantly distort the results of analyses, often negating the potential benefits of information-driven approaches [7]. Based on these definitions with corresponding effects of low data quality, the definition of [5] is used as it emphasizes the importance of taking a consumer viewpoint of quality. Without ignoring the effects, low data quality has across teams and organizational boundaries.

B. Data quality challenges

Before a data item ends up in a database, it typically passes through several steps involving human interaction and computation. Errors can occur in the process from the initial data acquisition to the archival storing [7], [8]. It is essential to understand the sources of data errors and the challenges that data quality faces to minimize them and develop appropriate data cleaning techniques to detect and relieve errors. To minimize the entry data errors and in developing appropriate data cleaning techniques to detect and relieve errors. The sources of errors in databases fall into the following categories: **Data entry errors -** commonly, data entries are done in a setting where a human is involved. Information can be extracted from speech or by keying in data from written or printed sources. Data entries can be corrupted by typographic errors or by a misunderstanding of the data source. **Measurement errors -** in some cases, data are intended to measure some physical process in the world. This can be the speed, the size of a population, activities' durations, etc. These measurements can be undertaken by a human process that can have errors in the design and execution. Sensor technology has increased and avoids various human errors in data acquisition, but measurement errors can still occur. The human design of sensor technology can still affect data quality by miscalibration and interference from unintended signals. **Distillation errors -** in many settings, raw data are pre-processed and summarised before they are entered in a database [7]. This is called data distillation, and it is done to reduce the complexity/noise in the raw data, perform domain-specific statistical analyses, emphasize aggregated properties of the raw data, and reduce the volume of data that has to be stored. Errors may occur during the distillation process or the techniques used to interfere with the final analysis. **Data integration errors -** databases often contain data from different sources. Data is collected and entered from multiple sources over multiple methods over time. Databases evolve and merge with others, and this requires resolving inconsistencies across databases. Procedures that integrate data from multiple sources can lead to errors.

With a large volume of data, it is not easy to judge the data quality within a reasonable amount of time [9]. The challenge lies in collecting, cleaning, and finally securing high-quality data within a given time frame. This is especially the case with unstructured data. Transforming this to further processing this data will take much time. It demands a lot from the existing techniques. Finally, data can change very fast, and the timeliness of data is short. Companies have to consider that the data has to be dealt with within a given time. Otherwise, it can be outdated and invalid information. This can harm the business process and analysis, which is based on these data. It leads to worthless and wrong conclusions, eventually leading to mistake in decision making.

C. Data quality dimensions

The literature defines the qualities, dimensions, and metrics to assess data as a critical activity. Quality can be described as a concept with multiple dimensions, and every dimension refers to an aspect of the quality of data. These dimensions correspond with the operational definition of the corresponding and unique metrics. The dimensions are mostly referred to as an extension of data-data values [10]. Views on what features increase data qualities can differ across organizations and industrial companies. The need to evaluate the quality of the data based on the aspects and features by measurable categories are called data quality dimensions [11].

Farshidi et al. [12]–[14] presented a framework for assisting decision-makers, such as software developers and architects, with their decision-making processes in software production. They suggested using standard software quality models, such as ISO/IEC 25010, to assess software and data quality. They asserted that the quality of a system is the degree to which the system meets its requirements (functionality, performance, security, maintainability, etc.). It is necessary to find quality attributes widely recommended by other researchers to measure the system's characteristics.

Batini et al. [6] introduced a framework based on classifications in clusters of dimensions, where dimensions are included in the same cluster according to their similarity. The clusters are defined as follows to represent the dimension of the cluster, followed by other member dimensions: (1) **Accuracy**, correctness, validity, and precision. (2) **Completeness**, pertinence, relevance. (3) **Redundancy**, minimality, compactness, and conciseness. (4) **Readability**, comprehensibility, clarity, and simplicity. (5) **Accessibility** and availability. (6) textbfConsistency, cohesion, and coherence.

This six-dimension classification described above is a selection based on the analysis of [10]. According to this analysis, it is possible to define a basic set of dimensions that consist of *accuracy, completeness, consistency, and timeliness*. These four dimensions constitute the focus of the majority of the publications in the literature. Note that there is a broad agreement on data quality dimensions in general with corresponding research in the last decades, but there is still no consensus on a standardized list of dimensions and metrics for data quality measurements [15].

D. *Data quality control methods*

Batini et al. [10] presented a subset of data quality assessment methods and categorized them into the following categories:

(1) **Total Data Quality method** (TDQM) is to deliver high quality information products (IP) to information consumers. The aim is to facilitate the implementation of an organization's overall data quality policy formally expressed by top management [16]. [16] uses the term ”information” interchangeably with ”data”. The TDQM cycle consists of four phases to continuously ensure and improve quality, defining, measuring, analyzing, and improving. Identifying information quality (IQ) dimensions and the corresponding IQ requirements with the definition phase is important. These IQ dimensions are related to the data quality dimensions discussed in section II-C. After defining the dimensions, the measure phase starts, which produces IQ metrics. An analysis is then applied to identify root causes for IQ problems to calculate the effects of low-quality information subsequently. Finally, the improvement starts, to make a selection of strategies and techniques to improve information quality.

(2) **Total Information Quality Management** The TIQM focuses on the management activities of data and focuses on the integration of operational data sources. This is done via strategy and organizational order to make effective technical choices. The analysis made is based on cost-benefit. This is supported from a managerial perspective. TIQM consists of three phases: assessment, improvement, and improvement management and monitoring. In the assessment phase, the identification of data, users, and requirements are made. With the improvement phase, data defects are analyzed—the identification of data sources that require cleaning. Lastly, in the improved management and monitoring phase, an improvement plan will be initialized to improve to resolve root causes for poor data quality [17]. This last phase is one of the valuable contributions of the TIQM methodology, as it provides guidelines to manage the change in the organization's structure according to data quality management requirement [10]. The TIQM is a detailed methodology that makes the implementation of all processes unnecessary. Detailed sub-tasks of the three phases are selected in specific settings as not all options are applicable to specify certain settings. The TIQM focuses on the management activities of data and the integration of operational data sources. This is done via strategy and organizational order to make effective technical choices. The analysis made is based on cost-benefit. This is supported from a managerial perspective. TIQM consists of three phases: assessment, improvement, and improvement management and monitoring. In the assessment phase, the identification of data, users, and requirements are made. With the improvement phase, data defects are analyzed—the identification of data sources that require cleaning. Lastly, in the improved management and monitoring phase, an improvement plan will be initialized to improve to resolve root causes for poor data quality [17]. This last phase is one of the valuable contributions of the TIQM methodology, as it provides guidelines to manage the change in the organization's structure according to data quality management requirement [10]. The TIQM is a detailed methodology that makes the implementation of all processes unnecessary. Detailed sub-tasks of the three phases are selected in specific settings as not all options are applicable particular setting.

(3) **Comprehensive Data Quality Method** (CDQ), in which the cost-benefit analysis is intensively used in different steps to define the data quality targets based on available funds and to give the selection of the most suitable improvement process. It is conceived to be complete, flexible, and simple to apply to all types of data. Completeness derives from the fact that it works in intra-organizational and inter-organizational contexts and according to the data type. It is flexible as the data quality techniques can change within each phase and in any context. Lastly, the CDQ is simple. It is set up in phases where specific goals with corresponding techniques characterize each phase. CDQ consists of three main phases, state reconstruction, assessment, and improvement. The state reconstruction identifies the role of organizational units in data usage, processes, services. This is reconstructed and documented. Second, the assessment is where data quality dimensions are measured and assessed in order to set new data quality targets [18], [19]. Finally, in the optimal improvement process, a selection of strategies and techniques is made to evaluate the cost/benefit ratio. In the state reconstruction phase, a complete picture of data providers and users is made, of the data flow among them and of the use of data [18]. This is modeled by using matrices which describe the use of data of organizational units and the roles in different business processes. The link to data and data flows is shown. The assessment phase consists of two steps. First, the data quality issues are identified via interviews with users. This highlights the quality problems and understands the consequences of low-quality data on the work. This is done according to data quality dimensions. This is the basis of the analysis of the process identified in phase one and the identification of the causes of low data quality [18]. Second, a quantitative evaluation of the quality issues is done. The related metrics are selected to apply to data and data flow in

the previous step. The final phase comprises five steps. The goal is to identify the best improvement process, the optimal cost/benefit ratio. First, target quality values have to be set to consider the cost and benefits. Next, different improvement techniques are executed to attain the target quality values. Next, the best technique for each activity is selected. It is necessary to analyze the techniques and compare their costs, and technical characteristics [18]. The improvement process is then made. Finally, the optimal improvement processes are compared via a cost-benefit analysis and selected.

*E. **Data curation***

Approaching data is a multi-step and iterative process involving collection, transformation, storing, auditing, cleaning, and analysis [7], [20]. The process includes people and equipment from multiple organizations within or across departments. In each step in the process, data quality improvement designs can be implemented. Data curation is at the center of data quality assessment. It aims to clean and transform the data to meet the quality criteria set by the user. Data curation is also referred to as data cleaning [21].

The curation task can be implemented manually or automated. Data standardization, de-duplication, and matching are examples of automated tasks [22]. The focus of this research lies in automated data curation, within particular anomaly detection. Several computational techniques are developed by research and industry for identifying and, at times fixing errors in data. Automating data curation methods are valuable for large amounts of data, as errors are hard to find. Although data curation is at the center of data quality assessment, automated data curation is not discussed extensively in the methodologies discussed in section II-E. The focus of this research lies in automated data curation, within particular anomaly detection. Several computational techniques are developed by research and industry for identifying and, at times fixing errors in data. Interruptions in the pipeline can introduce troublesome anomalies. Models that are derived from it will perform poorly and generate unreliable conclusions. According to [22], the key to high-quality ML are the three principles of data quality: prevention, detection, and correction. Anomaly detection aims to detect abnormal patterns deviating from the rest of the data, called anomalies or outliers [23]. These terms are considered synonymous and will be used in this research as items or events that do not conform to an expected pattern or other items in a data set. The types of data can categorize the space of techniques and products that they target.

F. Gap Analysis

The methodologies TDQM and TIQM, CDQ share the same objective, improving the quality of data. They have overlapping ideas but differ in the details. The same core principles are: (1) Identifying the importance of the quality of data. (2) Assessment with dimensions to measure the quality of data. (3) Analysing the quality of the data. (4) Improving the data quality.

Based on these similarities, a broad application of data quality control methodology can be defined. However, there are differentiating characteristics in these three methods. The TDQM follows the identification of roles of information to support each phase. The TIQM handles the tasks with specific detailed sub-tasks. Also, there is a distinction between the improvement of data itself and the improvement of the process. Additionally, the CDQ supports the user in the selection of techniques to apply. Based on the literature, a combination of the three methods will investigate how to enable unsupervised anomaly detection in validity and quality control for data management. The core principles of the three methods above will be used as guidelines with specific characteristics of the operational methodologies as these focus more on technological issues in the application of data-oriented techniques. The methods used to assess the data quality depend on the task at hand.

While there has been much research on data quality control methods and how to assess them, little research has considered automated anomaly detection. This research investigates the potential of using automated anomaly detection during the validation process regarding the data quality control for data management. In particular, the potential of using unsupervised methods because likely, just collected data has not gone through a data quality control. The quality of data is essential for the data consumer using them. It influences the quality of business processes and decision-making. The problem regarding the quality of data is the possibility of having anomalies in data. Before consumers can use data, it is helpful to prevent, detect and correct the data. The improvement area lies in increasing the quality of the data by removing these anomalies.

The operational methodology, as the TIQM and CDQ methodology by [24] and [18], do give guidelines in how to potentially position the automated unsupervised anomaly detection into the data quality control monitoring. Moreover, there is sub task given in the TIQM in the improvement phase, as the identification of data sources that require data cleaning or the extraction and analysis of relevant source data for anomalies [17]. However, it is lacking in the actual practical execution of trying to automate this.

In short, the explicit gaps of the current literature are: (a) A few research efforts have been performed in automated anomaly detection. (b) There is no standardized method in existing data quality control for detecting anomalies. (c) The current state of implementing unsupervised learning is unknown in data quality control. (d) There is a lack of practical execution. The action steps to be taken are to map the current state of data quality control and compare it to both the TIQM and CDQ methodology. Review the state and investigate how automated unsupervised anomaly detection can be implemented in the data quality control process.

III. A Semi-Automated Data Quality Pipeline

For this research, we will implement the application of automated anomaly detection for data quality control. As stated earlier, current research lacks the support of automated unsupervised anomaly detection during the data quality control process. However, the methodologies do give guidelines in how to assess data quality. TDQM, TIQM, and CDQ methodologies were used to synthesize the methodology used in this research.

The method proposed should be an iterative process and ideally automated, but a human in the loop has to be implemented to evaluate the results of the outliers. The main similarities of TDQM, TIQM, and CDQ are considered. These requirements are included in creating the methodology. A combination is made, and it resulted in the pipeline as seen in figure 1, with the addition of the implementation of automated anomaly detection to the data quality control.

The data validation workflow is a proof-of-concept and consists of the following five components: **Database**, the database needs no real explanation. It is a storage of the data with tables that can be accessed. **Data profiling and quality assessment**, which is similar to identifying the importance of the quality of data with the assessment with the dimensions to measure the quality. The **aggregation of the tables** is done in combination with pre-processing in order to get the data ready for the next step. **Automated anomaly detection** is a process that consists of sub-processes. Dependent on the task at hand and data types, the statistical or unsupervised clustering techniques can assess to detect outliers. After the selection, the process will be iterative as the model tries several different and tune the parameters. A score will be calculated to determine if the unsupervised clustering technique performed correctly. The best parameters and model will be selected before outputting the results. The output results will be executed and shown via a table and visualization. **Evaluate outliers** this is done by data management. It is a human in the loop, which makes the proposed methodology semi-automated. The human in the loop evaluates the output results to determine if the automated anomaly detection performed accordingly. To evaluate the pipeline, we have employed it in the context of a use case of LOGEX, a healthcare analytics company.

A. Design of the Pipeline

Before designing the pipeline, it should meet design requirements. These are functional and non-functional. Functional requirements describe a system or the components, while non-functional define the attributes of a pipeline.

As a functional requirement, the pipeline should allow the user to interact and tune the models selected—all dependent on the task at hand. The pipeline should show the user output for the human in the loop to understand the detected outliers. The non-functional requirement is that the pipeline should be flexible because the models selected should be adaptable. Furthermore, the pipeline is explained in detail. The data quality assessment component is required to identify the importance of the data. This is done to understand the importance of data with the specific need of the user. This enables initial input from the user of domain-specific knowledge to apprehend relevant dimensions to assess the data.

Once the mapping of identifying the importance of data quality is completed, specific data aggregations can be made to analyze the data. The data must be pre-processed as it is possible that the data provided cannot run through specific models. This is done to prevent certain errors. Subsequently, automated anomaly detection starts. This is a new component compared to existing methodologies. With automation, it is possible to detect outliers in complex data. This is a fast solution that saves the users data time because of moving from a manual to an automated approach.

Dependent on the data, a model is selected. Model selection is dependent on features that are entered in the input. Once a model is selected, the idea is the same. The model will be optimized based on the features and the parameters. This can go according to a specific rule of thumbs of the chosen model or via some hyper-parameter tuning. This process will be iterative to optimize the model and thus to find the best performing anomaly detection model. It can be done according to key performance indicators as the silhouette metric. To determine if the unsupervised clustering technique was executed correctly. Once the automated anomaly detection is executed, an outliers report is made for the analyst at a data management as output.

Lastly, the evaluation of the report is the human in the loop. This should be done manually to verify how the model performed to meet the non-functional requirements. Visualization of data and outliers are presented to the user. This is a method for the human in the loop to understand how the model performs. It is important to note that the human in the loop should be critical with the generated outliers, as they are dependent and thus vary based on the type of data.

B. Implementation

The automated anomaly detection component consists of two different techniques for anomaly detection. The first one is a more traditional statistical method, and the other is based on unsupervised clustering. In the current pipeline, we implement statistical z-score and IQR as approaches for statistical methods. Outliers can be detected with statistical z-score and IQR if data points do not lie within a certain threshold. The method is powerful when dealing with Gaussian, uni-variate, and numerical data. The statistical methods are also easy to understand with statistics.

For the unsupervised clustering approach, there is a broader selection of anomaly detection models. The models attempted for usage are based on the packages scikit-learn and PyOD (Python toolkit for detecting outlying objects) of [25]. The packages support different types of clustering methods. However, a density-based approach is eventually chosen for the pipeline and, in particular, the DBSCAN model. The DBSCAN is selected as the model can be paired with any data, making it suitable for different types of data sets. It is noted that the model selection is dependent on the aggregation of the data set and use case, which happens before the model selection.

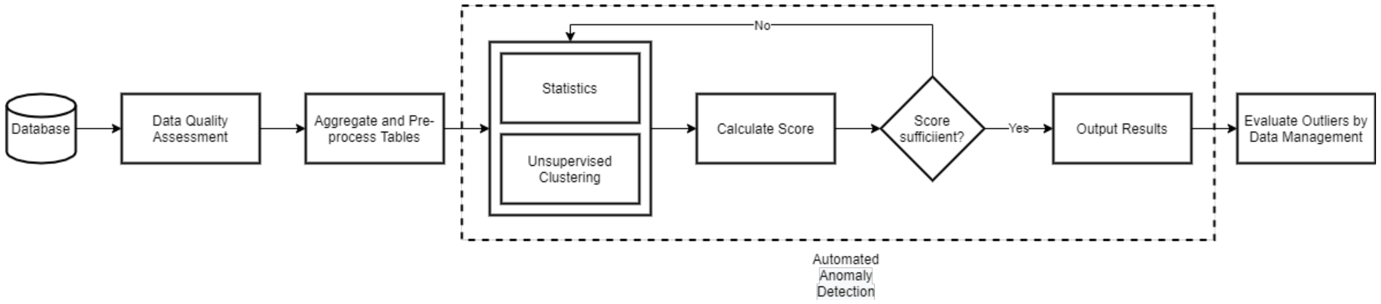

Fig. 1: Pipeline of Automated Anomaly Detection in Data Quality Control

Finally, the pipeline itself can be implemented in real-life scenarios where data quality will be assessed. It is a later step in the data quality process, as it is recommended that essential quality checks be executed before starting the pipeline. It is done to get the best results for anomaly detection. It is also recommended that data is pre-processed before implementing it into the proposed semi-automated data quality pipeline.

IV. Experimental Results

A. Current Experiment

For the experiment, the focus lies on automated anomaly detection in the data quality process. LOGEX performs their data validation process according to input and output validation as seen in figure 2. The scope of the experiment is on LOGEX's ”output” validation. Their input validation is processed according to basic cleaning rules and transformed according to LOGEX business rules. The output validation is the final stage before other business units can use data.

The data consists of all the medical activities that a specific hospital has executed. This is provided by their clients, which are hospitals in the UK. The data is mixed data which consists of quantitative and categorical data. Before analysis, the data set was aggregated to every activity encounter to understand the data. The data used are API.table_products and consists of two relational tables. The two tables are the API.encounter table, which links to the API.activity tables, consisting of 61 and 43 columns. Not all columns are relevant for the use case, and a selection of the tables is made. Besides, most columns consist of null values and are not in use yet.

To start the experiment, first, the understanding of the users of the data is made. Second, specific aggregation is applied to analyze the data. Third, once the aggregation is executed, the anomaly detection technique is selected based on the uni-variate or multi-variate approach. For the experiment, a multi-variate approach is chosen as most data are related to each other, which makes it challenging to detect outliers via a uni-variate approach complex. Most data in the data set is of a categorical type. Furthermore, the data provided is not labeled, which means unsupervised learning techniques are explored. There was also no example data set available as a reference. To select the anomaly detection models, first, a comparison between a selection of unsupervised clustering anomaly detection techniques is made. The models are implemented via the packages of scikit-learn and PyOD.

Because the data is not labeled, the results are compared with fake outliers injected into the data set. The results are evaluated via a classification report. The primary metric depended on the business objective of the user. For this use, case precision is the primary metric, this measures when a positive value is predicted, how often the prediction is correct. The positive value is for LOGEX, the true positive predicted outliers. The business objective of LOGEX is that outliers should be detected, but false negatives are acceptable as outliers do exist in real-life scenarios and thus is less of a concern. The performance is also based on the speed of the algorithm.

As a different experiment, the method of ensemble learning is also taken into consideration and explored. With ensemble learning, the approach is taken to harmonize the anomaly detection approach. For unsupervised learning, the consensus technique is used. With consensus, several clustering algorithms are executed for a given data set. The clustering ensemble then computes a set of clusters and selects what the consensus is based on the most occurring value `outlier = <1, 0>`.

If the model selection has been performed, the selected model is optimized. This is done via a specific rule of thumb of that particular model, with a combination of automatic hyper-parameter tuning. Both options are compared, and according to the silhouette metric, the best model is selected. The silhouette metric has values ranging from -1 to 1. -1 indicates that clusters are assigned in the wrong way. Data points should have been assigned to a different cluster when it is more similar. Scores near 0 indicate overlapping clusters, and a score closer to 1 indicates that clustering has performed well and is distinguished.

Lastly, the results are given to the human in the loop. He would evaluate the results and check if the automated quality control method performed accordingly. The output is shown via a CSV file with the ids of the encounters, with visualization to check the automated quality control method manually. The experiment is executed to assess if unsupervised anomaly detection can be implemented during data quality control by implementing an automated model in data validation for data management.

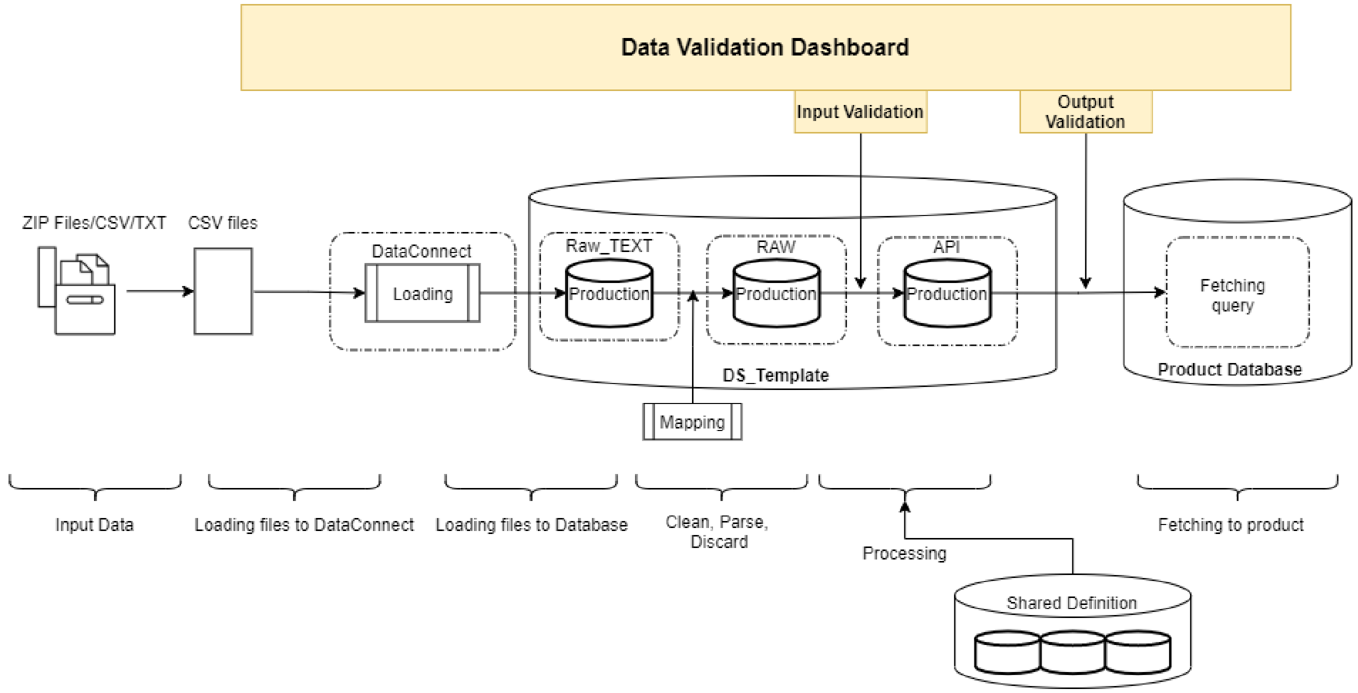

Fig. 2: Data Validation Workflow of LOGEX

B. Data Quality Assessment of LOGEX

An expert interview is conducted to understand the importance of LOGEX's data. The business objective is to understand the use of the data. LOGEX offers solutions for customers to deliver the best possible care at the lowest possible cost. Clients want to turn their data into clarity to make well-informed decisions. One of the three solutions is for the financial domain. The financial domain covers the control costs and streamlines operations, maximizing operational and financial performance. The objective of LOGEX's automated quality control is to detect outliers, to prevent clients from having a distorted view of the results. This can, for example, lead to incorrect budgeting for specific care and have severe consequences for patients if care cannot be performed. With correct data and thus information, their clients can adjust business accordingly.

The main dimension is used to assess the data accuracy, in particular, the semantic accuracy. This measures data points with a `<correct, not correct>` domain. LOGEX's focus of the automated data quality control is on detecting outliers that leads to detecting data points as `outliers = <True, False>`. The automated quality control method affects the $free\,of\,error\,rating$ formula. As errors are detected, the smaller the $free\,of\,error\,rating$.

Timeliness also affects the data quality as the data provided must be up to date and give relevant information to the client. At the moment, it is not relevant as this is done for testing. However, if the automated data quality control is implemented, it is essential to check the data for currency, and the run time must be fast or at least doable. Note, it is hard to assess the dimensions discussed as the data cannot be related to another example data set. The medical activities provided are private information and are not comparable with public information.

C. Automated Quality Control Experiment

The aggregation of the encounter tables is done on every encounter, based on the encounter_id, grouped by the summation of the los_minutes as total_los_minutes. The data consists of multi-variate data with mixed data types and high dimensionality, and pre-processing is needed. The data set consists of columns with categorical values, and these are given labels via the label encoder of scikit-learn. The null values are then dropped as these cannot be taken into account during the statistical clustering technique for anomaly detection. Next, the features are standardized by removing the mean and scaling to unit variance. Finally, the data is standardized. by scaling, all the features are aligned to a mean of zero and one standard deviation.

After the pre-processing of the data, the problem of high dimensionality is solved via principal component analysis (PCA). This reduces the features by converting the original feature set into fewer artificially derived features which still maintain most of the information encompassed in the original features. The explained variance threshold is set at 95%. For the model comparison, a selection of classifications of anomaly detection models is made, with each a specific model that represents the type of model. The models used are baseline models, i.e., run with default parameters. The selection consists of the clustering classification with the addition of linear and proximity models in the PyOD packages. The results are shown in table I with a plot in figure 3: (a) **Clustering**: KNN, AvgKNN, MedKNN (b)**Density**: DBSCAN & LOF (c) **En-**

semble: IForest (d) **Linear** (PyOD): One Class Support Vector Machine (OCSVM), PCA (e) **Proximity** (PyOD): Histogram-based Outlier Score (HBOS), Rotation-based Outlier Detection (ROD)

Model	N Outliers	Precision	Recall	F1-Score	Time Elapsed in sec.
KNN	230	0.754	**0.711**	**0.732**	0.255
AvgKNN	189	0.815	0.647	0.721	0.343
MedKNN	183	0.823	0.653	**0.732**	0.605
DBSCAN	36	**0.941**	0.168	0.286	0.084
LOF	227	0.410	0.179	0.249	0.065
IForest	256	0.773	0.626	0.692	1.377
Feature Bagging	237	0.374	0.179	0.242	0.624
OCSVM	256	0.779	0.668	0.720	1.902
PCA	256	0.878	0.379	0.529	0.013
HBOS	173	0.822	0.584	0.683	0.009
ROD	210	0.750	0.205	0.322	0.667

TABLE I: Comparison of baseline anomaly detection models

1) Ensemble learning: As a different experiment, ensemble learning is performed to explore if it is a feasible approach. This is done based on the consensus method, and with the consensus method, multiple clustering models are executed, and the output of the models is compared to find the overall consensus over all data points. For example, if data point x is chosen, and the three models detect the point as $outlier$: $x = [1, 1, 1, 0]$, the consensus is that it is an outlier as most models detect the data point as an outlier. For the selection of the models, the precision score for detecting outliers is chosen. For the use case, precision is the most important metric as it suits the business objective. The threshold for the minimum precision score for the experiment is 0.8, which leads to a combination of the following models: AvgKNN, DBSCAN, HBOS, MedKNN, and PCA. The criteria for selecting the models can be changed, depending on the business objective.

The results are shown in the table II and do show high scores based on the classification, and there is, however, a disadvantage with this method. Because the consensus ensemble learning is based on a consensus of a set of anomaly detection algorithms, the run time is much longer than only using one model.

Model	N Outliers	Precision	Recall	F1-score
consensus	135	0.922	0.558	0.695

TABLE II: Ensemble learning: consensus results

2) Model comparison: Based on the comparison between the different types of anomaly detection models, DBSCAN is selected as the automated data quality control method. A selection is made to further research into automating one model and improving a baseline model to increase the performance. In the case of LOGEX, precision is the most critical metric. It is noted that the KNN and the MedKNN approach have better recall and are better balanced with a higher F1-score.

However, because it is in their interest that as many outliers are detected as possible, LOGEX takes some outliers for granted as medical encounters can take longer in real-life scenarios. It suits their business objective better as LOGEX wants the model to filter unsuitable data points accurately. It is noted that the number of detected outliers is far less compared to other models in table I. With hyper-parameter tuning, we optimize the model accordingly to reduce type II errors.

D. Automated Data Quality Control

The DBSCAN model uses a density level estimation based on a threshold for the number of neighbors, minPts, within the radius ϵ (with an arbitrary distance measure) [26]. DBSCAN classifies data points as core, border, and noise points. Points with at least the minPts (including the point) in its surrounding areas within the ϵ radius are considered core points. Non-core points are border points; they reach a core point but do not satisfy the minPts parameter. Points that are not density reachable from any core point are considered noise and do not belong to any cluster [26].

For the parameter tuning there is a rule of thumb, [27] suggest to set the minPts to $minPts = 2 \cdot number\,of\,dimension$. The ϵ is harder to set; domain knowledge to set it would be ideal, but this is mostly not the case. However, it can be set by detecting the elbow of the distance of the k-nearest-neighbor to get the appropriate ϵ. DBSCAN's strength is that it can be paired with any data type, distance function, and indexing technique adequate for the data set to be analyzed [26]. Furthermore, it can determine arbitrary patterns, noise, and different cluster sizes accurately [28].

Two types of optimizations are applied to automate the data quality control method with DBSCAN anomaly detection. The parameters are changed via iteration and compared via the silhouette score to automate the process. Also, the rule of thumb is utilized. The parameters with the best silhouette score are chosen, and if it is sufficient, i.e., close to 1. The results for both methods shown in table III.

DBSCAN	N Outliers	Precision	Recall	F1-score	Silhouette Score
Elbow	154	0.87	0.58	0.70	0.720
Iterative	100	0.94	0.42	0.58	0.664

TABLE III: Automated DBSCAN results

Based on the silhouette metric, the experiment shows that the elbow method performs better than the iterative method. It is noted that, if only focused on the precision score, the iterative method performs better than the elbow method. Also, based on the business objective, this should be the logical choice. However, the elbow method is not far off with the precision score with the combination of having a better balanced with a higher recall and thus F1-score. Furthermore, the Silhouette metric scores better, which means that the clusters formed are better than the iterative method. We can interpret the data points as clustered correctly with a score higher than 0 and close to 1. The clusters are dense with a score of 0.720.

The experiment of selecting the best model based on the business objective and then optimizing the baseline model shows promising results for enabling unsupervised anomaly detection in quality control. Due to the results of the optimized DBSCAN with the silhouette score more significant than 0 and close to 1, it indicates that the outliers detected are executed appropriately.

E. Human in the Loop

The last stage in the pipeline is evaluating the results outputted by the automated quality control method, i.e., the

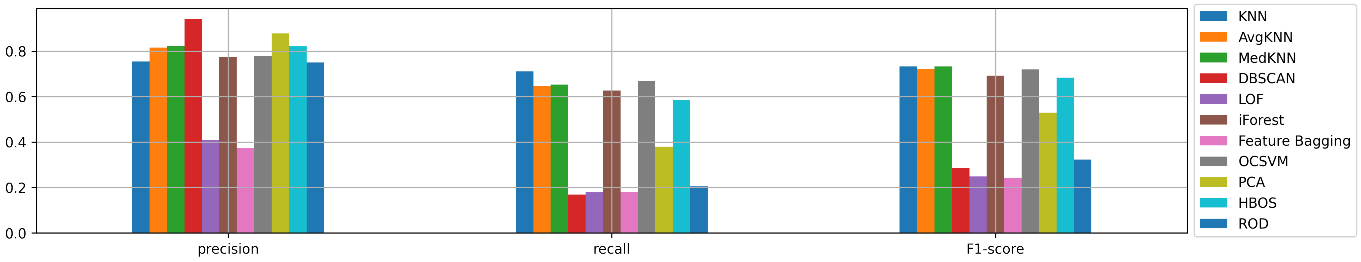

Fig. 3: Performance plot of anomaly detection models

DBSCAN algorithm. This is performed by a human in the loop, as it is essential that the output provided by the model can still lack performance. Because the results are based on unsupervised learning, and the model cannot understand the data like a human can. In the case of LOGEX, this would be performed by an analyst who is responsible for the relevant hospital or data set. This makes the pipeline semi-automated instead of fully automated.

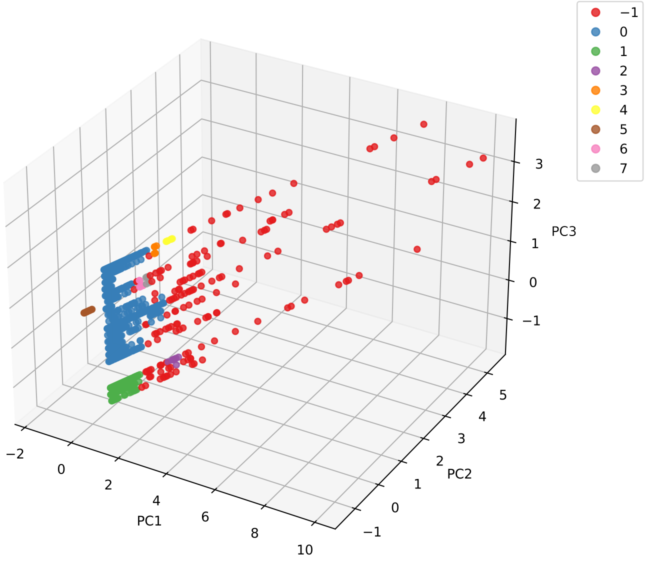

Fig. 4: Visualisation of DBSCAN elbow method results

The output can best be comprehended via a visualization. Ideally, this would be done via a dashboard to give the analyst a visual idea of how the outliers are defined. Figure 4 is an example visualisation of the DBSCAN elbow results. The red dots that are labeled ”-1” are noted as noise points and thus the outliers. The visualization is shown in a three-dimensional plot, which makes it harder to understand. However, it is still clear that, based on DBSCAN, it detects outliers correctly as the data points are not in a reachable density from the core points. After the visualization, the human in the loop has the output of the data shown in the table, the same as the original data set used as input. With an addition of a column ”outlier” with the semantic evaluation of `<true, false>` or `<1, 0>`. With these means, the human in the loop can easily find values in the data set labeled as outliers.

V. Discussion

Different anomaly detection methods are implemented to enable automated unsupervised learning into data quality control for this research. Comparing the models and eventually focusing on the DBSCAN as the automated model for quality control shows that outliers can be detected according to the proposed pipeline. Due to the nature of unsupervised learning, the validation of the results is based on the silhouette metric and the testing with injecting fake outliers.

Based on the results of the silhouette metric and the chosen metric of the use case, precision. It shows that it is feasible to enable unsupervised anomaly detection during quality control. The approach of combining the rule of thumb and iterative methods with the silhouette score creates an automated component into the pipeline. However, other models than the DBSCAN works may perform better on different aggregations. Specific situations require specific models. Implementing a dynamic selection of the models is tricky.

Several baseline anomaly detection models were run on the data set, and the performance was observed and compared. The selection is based on performance and depends on the business objective at hand. Still, the experiment proves the possibility of enabling unsupervised anomaly detection in validity and quality control for data management. Although the pipeline could not be fully automated due to the lack of labeling and knowledge by an unsupervised learning model, the human in the loop is an essential component to evaluate the results act as the final check. This does make the pipeline semi-automated.

The proposed pipeline is an addition to existing research about data quality control, as there is a gap in the literature regarding automated anomaly detection for quality control. The is a lack of practical execution, and few have considered it. Based on this research and experiment, other researchers and companies should explore the implementation of unsupervised anomaly detection into automating data quality control.

It should be noted that DBSCAN is not the final anomaly detection model. Each aggregation asks for a different model, and each business objective requires another metric to assess the performance. It remains necessary that multiple models are compared to each other to determine the selection for the automated anomaly detection algorithm. Because of this, one automated anomaly detection algorithm for all data quality problems cannot be chosen. The recommendation for further research is to explore other options of unsupervised learning in data quality control. The experiment of integrating ensemble learning shows promising results to validate the selection of outliers via a consensus learning approach. With this approach, data points can be determined as outliers with a more balanced and higher certainty.

VI. Conclusion and future work

This paper proposed a pipeline for automating data quality control by employing an unsupervised anomaly detection model. Additionally, a set of dimensions (accuracy, completeness, consistency, and timeliness) and corresponding metrics are identified to assess data quality. Fully automating the process is not (yet) feasible as there is still a need for a human in the loop. The human in the loop who evaluates the model cannot draw a conclusion based on a metric. Accordingly, a mode selection process is employed to support the human in the loop and to compare the utility of anomaly detection in data quality control. The main contribution of ensemble learning is explored to check if consensus for unsupervised learning can be implemented. The precision score of 0.922 shows promising results. The results make the unsupervised approach more balanced as the outliers detected are labeled via a "consensus". However, it should be noted that the run time is much longer than running just one model.

As the next course of action in this research, we plan to employ different anomaly detection models and evaluate their performance against each other. Automatic hyper-parameter tuning on other models could then be explored. In addition, determining outliers can be selected via `<true, false>` labels of the selected models. Some anomaly detection models can also output an anomaly score to determine the chance of a data point being an outlier. Finally, it is interesting to explore unsupervised ensemble learning further. It shows promising results with the consensus approach, and it can be an addition for anomaly detection on unlabelled data. A more balanced result via a "consensus" can give the data better assurance of labeling a data point as an outlier.

Acknowledgment

This work has been partially funded by the European Union's Horizon 2020 research and innovation programme by the ENVRI-FAIR project grant agreement No 824068, by the BLUECLOUD project grant agreement No 862409, by the ARTICONF project grant agreement No 825134, by the project CLARIFY under the Marie Sklodowska-Curie grant agreement No 860627, and by the LifeWatch ERIC.

References

[1] Sebastian Schelter, Dustin Lange, Philipp Schmidt, Meltem Celikel, Felix Biessmann, and Andreas Grafberger. Automating large-scale data quality verification. In *Proceedings of the VLDB Endowment*, volume 11, pages 1781–1794. Association for Computing Machinery, 2018.

[2] Sergey Redyuk, Zoi Kaoudi, Volker Markl, and Sebastian Schelter. *Automating Data Quality Validation for Dynamic Data Ingestion*. 2021.

[3] Richard Y Wang and Diane M Strong. Beyond Accuracy: What Data Quality Means to Data Consumers. Technical Report 4, 1996.

[4] Shazia Sadiq and Marta Indulska. Open data: Quality over quantity. *International Journal of Information Management*, 37(3):150–154, 6 2017.

[5] Anastasija Nikiforova. Definition and evaluation of data quality: User-oriented data object-driven approach to data quality assessment. *Baltic Journal of Modern Computing*, 8(3):391–432, 2020.

[6] Carlo Batini and Monica Scannapieco. Data-Centric Systems and Applications Data and Information Quality. Technical report, 2016.

[7] Joseph M Hellerstein. Quantitative Data Cleaning for Large Databases. Technical report, 2008.

[8] Zhiming Zhao, Adam Belloum, and Marian Bubak. Special section on workflow systems and applications in e-science. *Future Generation Computer Systems*, 25(5):525, 2009.

[9] Li Cai and Yangyong Zhu. The challenges of data quality and data quality assessment in the big data era. In *Data Science Journal*, volume 14. Committee on Data for Science and Technology, 2015.

[10] Carlo Batini, Cinzia Cappiello, Chiara Francalanci, and Andrea Maurino. Methodologies for data quality assessment and improvement. *ACM Computing Surveys*, 41(3), 7 2009.

[11] AltexSoft Inc. Data Quality Management: Roles, Processes, Tools — by AltexSoft Inc — DataDrivenInvestor, 2019.

[12] Siamak Farshidi, Slinger Jansen, and Sven Fortuin. Model-driven development platform selection: four industry case studies. *Software and Systems Modeling*, pages 1–27, 2021.

[13] Siamak Farshidi, Slinger Jansen, and Mahdi Deldar. A decision model for programming language ecosystem selection: Seven industry case studies. *Information and Software Technology*, page 106640, 2021.

[14] Siamak Farshidi and Slinger Jansen. A decision support system for pattern-driven software architecture. In *European Conference on Software Architecture*, pages 68–81. Springer, 2020.

[15] Lisa Ehrlinger, Elisa Rusz, and Wolfram Wöß. A Survey of Data Quality Measurement and Monitoring Tools. 7 2019.

[16] Richard Y Wang. A Product Perspective on Total Data Quality Management. Technical Report 2, 1998.

[17] Christian Fürber. *Data Quality Management with Semantic Technologies*. Springer Fachmedien Wiesbaden, 2016.

[18] Carlo Batini, Federico Cabitza, Cinzia Cappiello, and Chiara Francalanci. A comprehensive data quality methodology for web and structured data. *International Journal of Innovative Computing and Applications*, 1(3):205–218, 2008.

[19] Zhengqiu Zhu, Bin Chen, Wenbin Liu, Yong Zhao, Zhong Liu, and Zhiming Zhao. A cost-quality beneficial cell selection approach for sparse mobile crowdsensing with diverse sensing costs. *IEEE Internet of Things Journal*, 8(5):3831–3850, 2020.

[20] Xiaofeng Liao and Zhiming Zhao. Unsupervised approaches for textual semantic annotation, a survey. *ACM Computing Surveys (CSUR)*, 52(4):1–45, 2019.

[21] Vasileios C. Pezoulas, Konstantina D. Kourou, Fanis Kalatzis, Themis P. Exarchos, Aliki Venetsanopoulou, Evi Zampeli, Saviana Gandolfo, Fotini Skopouli, Salvatore De Vita, Athanasios G. Tzioufas, and Dimitrios I. Fotiadis. Medical data quality assessment: On the development of an automated framework for medical data curation. *Computers in Biology and Medicine*, 107:270–283, 4 2019.

[22] Lisa Ehrlinger, Verena Haunschmid, Davide Palazzini, and Christian Lettner. A DaQL to Monitor Data Quality in Machine Learning Applications. pages 227–237. 5 2019.

[23] Srikanth Thudumu, Philip Branch, Jiong Jin, and Jugdutt (Jack) Singh. A comprehensive survey of anomaly detection techniques for high dimensional big data. *Journal of Big Data*, 7(1), 12 2020.

[24] Larry P English. *Improving data warehouse and business information quality: methods for reducing costs and increasing profits*. John Wiley & Sons, Inc., 1999.

[25] Yue Zhao, Zain Nasrullah, and Zheng Li. PyOD: A Python Toolbox for Scalable Outlier Detection. *Journal of Machine Learning Research*, 20(96):1–7, 2019.

[26] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems*, 42(3), 7 2017.

[27] J ¨ Org Sander and Martin Ester. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. Technical report, 1998.

[28] Ajay Sreenivasulu, Alexander Karlsson, and Nikolaos Kourentzes. EVALUATION OF CLUSTER BASED ANOMALY DETECTION. Technical report, 2019.