IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 July 2023, accepted 12 September 2023, date of publication 20 September 2023, date of current version 27 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3317354

RESEARCH ARTICLE

Quality Anomaly Detection Using Predictive Techniques: An Extensive Big Data Quality Framework for Reliable Data Analysis

ELOUATAOUI WIDAD, ELMENDILI SAIDA, AND YOUSSEF GAHI, (Senior Member, IEEE)
Laboratory of Engineering Sciences, National School of Applied Sciences, Ibn Tofail University, Kenitra 14000, Morocco
Corresponding author: Elouataoui Widad (widad.elouataoui@uit.ac.ma)

ABSTRACT The increasing reliance on Big Data analytics has highlighted the critical role of data quality in ensuring accurate and reliable results. Consequently, organizations aiming to leverage the power of Big Data recognize the crucial role of data quality as an integral component. One notable type of data quality anomaly observed in big datasets is the presence of outlier values. Detecting and addressing these outliers have become a subject of interest across diverse domains, leading to the development of numerous anomaly detection approaches. Although anomaly detection has witnessed a proliferation of practices in recent years, a significant gap remains in addressing anomalies related to the other aspects of data quality. Indeed, while most approaches focus on identifying anomalies that deviate from the expected patterns, they do not consider irregularities in data quality, such as missing, incorrect, or inconsistent data. Moreover, most of approaches are domain-correlated and lack the capability to detect anomalies in a generic manner. Thus, we aim through this paper to address this gap in the field and provide a holistic and effective solution for Big Data quality anomaly detection. To achieve this, we suggest a novel approach that allows a comprehensive detection of Big Data quality anomalies related to six quality dimensions: Accuracy, Consistency, Completeness, Conformity, Uniqueness, and Readability. Moreover, the framework allows for sophisticated detection of generic data quality anomalies through the implementation of an intelligent anomaly detection model without any correlation to a specific field. Furthermore, we introduce and measure a new metric called "Quality Anomaly Score," which refers to the degree of anomalousness of the quality anomalies of each quality dimension and the entire dataset. Through the implementation and evaluation of our framework, the suggested framework has achieved an accuracy score of up to 99.91% and an F1-score of 98.07%.

INDEX TERMS Anomaly detection, big data, big data quality, data quality dimensions, quality anomaly score.

I. INTRODUCTION

With the increasing digitization of almost all aspects of business and society, large volumes of data are being generated and captured, providing valuable insights into customer behavior, market trends, and operational efficiencies. As a result, Big Data has become a critical component in driving innovation, optimizing decision-making processes, and improving organizational performance. However, Big Data quality remains a significant challenge, with many data sources containing errors, inconsistencies, and inaccuracies that can significantly impact the accuracy of their insights [1]. These challenges are further worsened by the significant Volume, high Velocity, and Variety of Big Data (known as Big Data 3 Vs), making it challenging to identify and correct data anomalies [2]. Thus, addressing quality anomalies is an essential component of any Big Data strategy and should be prioritized by organizations seeking to maximize the value of their data resources. Although quality anomalies are generally addressed using basic and conventional techniques, identifying data quality anomalies, particularly in the context of Big Data, is more challenging and requires

The associate editor coordinating the review of this manuscript and approving it for publication was Chong Leong Gan.

sophisticated and intelligent methods for their detection. One of the quality anomalies that can significantly impact the accuracy and reliability of Big Data analysis is outlier values. Outlier values refer to data points that deviate from the normal distribution of data values [3]. These values can occur due to various factors, including measurement errors, data entry errors, or rare occurrences. Many Big Data anomaly detection approaches have been suggested in the literature to address issues related to outlier values [4], [5], [6]. These approaches have contributed significantly to advancing anomaly detection in various fields, including finance, healthcare, transportation, and security. While outlier values are a critical aspect of data quality, they represent only one of many other aspects that can impact Big Data analysis. Organizations that focus only on addressing outlier values may overlook other essential quality anomalies that can affect the accuracy and reliability of the analytical results. For instance, missing data, incorrect data formats, and duplicate data can introduce biases and errors into analytical results, leading to wrong business decisions. Thus, despite significant progress in anomaly detection, the existing approaches have not fully addressed the issue of quality anomalies. While most methods aim to identify deviations from expected patterns, they do not consider the potential anomalies related to data quality. Such anomalies can be particularly challenging to detect as they may not exhibit characteristic patterns associated with abnormal behavior. Therefore, there is a pressing need for more sophisticated and practical approaches to identify and address data quality anomalies, especially for Big Data. To address the concerns raised regarding the importance of managing all aspects of data quality anomalies, we propose a comprehensive approach for Big Data Quality Anomaly Detection, with the following main contributions:

1) Defining a new end-to-end Data Quality Anomaly Detection Framework that allows the identification of potential generic data quality anomalies for Big Data using predictive techniques as a proactive approach to ensure data accuracy and reliability.
2) Detecting the quality anomalies related to six quality dimensions: Accuracy, Consistency, Completeness, Conformity, Uniqueness, and Readability.
3) Introducing and computing a new metric called ''Quality Anomaly Score.'' This metric refers to the degree of anomalousness and low-quality of the detected anomalies for each quality dimension and the entire dataset.

The rest of this paper is organized as follows: In the next section, we discuss the importance of ensuring high data quality for Big Data and the impact of quality anomalies on the accuracy of data analysis. Section III reviews existing research on anomaly detection and data quality for Big Data. In the fourth section, we present the different steps of our proposed Big Data Quality Anomaly Detection Framework. Section V presents the implementation of our proposed framework for two big datasets and the obtained results. Finally, conclusions are made, and future work directions are discussed.

II. IMPROVING DATA QUALITY THROUGH ANOMALY DETECTION

A. THE IMPORTANCE OF DATA QUALITY

Big Data has shown excellent capabilities for supporting organizations in different fields to improve their business and manage their operations. Nevertheless, the advantages of Big Data can only be fully realized if data quality is improved. Big Data usually contain quality anomalies such as inconsistent, missing, or inaccurate values, which may impact the accuracy and reliability of the analytical results. Thus, data quality is critical for any organization, as the accuracy and completeness of data can directly impact business decisions and outcomes. In fact, good data quality ensures that the decisions made based on the data are accurate, reliable, and well-informed. Poor quality data can lead to incorrect conclusions, significantly affecting the organization [7]. In addition, clean, accurate data can reduce the need for manual data cleaning and processing, saving time and resources. Moreover, poor-quality data can be costly for organizations [8], leading to additional data cleaning, correction, and rework costs. Organizations can avoid these costs by ensuring data quality and using their resources better.

The data quality problem for Big Data is particularly challenging because of the sheer volume and velocity of the generated data. Big Data is often characterized by the ''3Vs'': volume, speed, and variety [9], [10]. This means that large amounts of data are generated at a high velocity, coming in various formats and from multiple sources. In addition to the particular characteristics of Big Data, data quality issues can arise for various reasons, including errors in data collection, processing, storage, inconsistencies, and discrepancies in data from different sources. As a result, ensuring the quality of Big Data can be difficult. Inaccurate, incomplete, or inconsistent data can lead to incorrect conclusions and decisions, which can have significant negative consequences. Moreover, Big Data is often used to train machine learning models, and the quality of the data used to train these models can significantly impact their accuracy and reliability [11]. Therefore, ensuring the quality of Big Data is essential to derive meaningful insights and make informed decisions.

B. ANOMALY DETECTION FOR DATA QUALITY

To improve data quality, quality anomalies should be identified and addressed. A data quality anomaly can be defined as an irregularity or abnormality in a dataset caused by errors, inconsistencies, inaccuracies, or incompleteness of the data. It can also refer to unexpected or unusual patterns in data that do not fit the expected norms or distributions, considering data quality anomalies as deviations or irregularities from what is expected or normal. Data quality anomalies can be detected using anomaly detection techniques.

Anomaly detection, also known as outlier detection, identifies rare or unusual events, patterns, or observations that deviate from the expected or standard behavior of a system or a dataset [12]. Anomalies may indicate fraudulent activities, errors in data collection, equipment malfunctions, security breaches, or other unexpected phenomena that require attention or investigation. Anomaly detection can be performed in various domains, such as finance, healthcare, manufacturing, cybersecurity, and transportation, where abnormal behavior or events can have significant consequences. However, addressing data quality anomalies using anomaly detection techniques requires representing the quality anomalies in a detectable format that highlights the quality deviation and makes the quality anomaly noticeable as an abnormal point. Thus, this paper introduces a Big Data quality anomaly detection framework that detects different data quality anomalies related to each quality metric. Moreover, the framework allows us to measure the anomaly score of each anomaly, the quality anomaly score for each metric, and a global quality anomaly score.

III. RELATED WORKS

Anomaly detection has been the subject of extensive research in recent years, with a wide range of approaches proposed to address this challenging problem. Numerous studies and articles have been published on the subject, especially for high-dimensional datasets, indicating the importance and significance of Big Data in various domains [13]. Moreover, this topic continues to gain more interest as organizations recognize the value of identifying unusual patterns or events in their data, which could significantly impact their business operations [14]. Table 1 summarizes the most recent articles that have addressed anomaly detection for different domains sorted by the year of publication. In the process of conducting the literature review, whether a paper is about Big Data is based on the focus, methodologies employed, and alignment with the main characteristics and requirements of Big Data. While the primary objective is to address anomaly detection and data quality for Big Data, we included some papers not about Big Data based on their exceptional performance in particular scenarios and their exploration of anomaly detection for data quality, even in a limited scope and within a domain-specific context.

All these approaches have significantly contributed to the state of the art of anomaly detection for high-dimensional datasets. However, most anomaly detection approaches are domain-correlated. Actually, anomalies can look very different depending on their field. As a result, the success of an anomaly detection approach often depends on its ability to capture the unique properties of the area being analyzed. Consequently, many researchers have focused on developing domain-specific methods that can accurately detect anomalies within a specific context but do not consider anomalies related to data quality. Based on the above exploration of state of the art, reviewing the existing studies related to Big Data quality and Big Data anomaly detection led us to make the following points:

1) Due to the multivariate nature of anomalies, most approaches are domain-correlated and allow the detection of anomalies related to a specific application domain.
2) Current anomaly detection approaches address "domain anomalies" related to a specific field and do not address anomalies in data quality. Thus, to the best of our knowledge, the use of anomaly detection techniques to comprehensively address data quality, including its various dimensions, has not been explored in the academic literature.
3) Most of anomaly detection approaches focus only on detecting outlier data that deviate from expected patterns and disregard the other aspects of data quality.
4) Conventional data cleaning tools are often employed to handle data quality anomalies. However, these tools may not suffice in uncovering latent anomalies that necessitate the application of sophisticated and intelligent techniques for detection.

To overcome the raised points, we propose a Big Data Quality Anomaly Detection Framework with the following main contributions:

1) Defining a new end-to-end Data Quality Anomaly Detection Framework that allows an intelligent detection of generic data quality anomalies for Big Data without any correlation to a specific field based on a predictive model.
2) Identifying a broad range of generic data quality anomalies related to six data-quality dimensions: Accuracy, Consistency, Completeness, Conformity, Uniqueness, and Readability.
3) Introducing and computing a new metric called "Quality Anomaly Score." This metric refers to the degree of incompatibility and the low-quality of the detected anomalies of each quality dimension and the entire dataset.

IV. BIG DATA QUALITY ANOMALY DETECTION FRAMEWORK

This section presents a Data Quality Anomaly Detection Framework that allows for identifying anomalies through an extended isolation forest model. Anomaly detection is mainly based on data anomalies being generally unusual and unexpected compared to the rest of the data. However, anomaly detection in a data quality context is quite challenging because quality anomalies are usually hidden, not obvious, and can be related to in-depth features that may not be apparent in the data. Such anomalies cannot be detected using conventional cleaning tools and require intelligent features based on machine learning that can automatically learn and adapt to complex anomaly patterns and behaviors. Therefore, we propose an advanced predictive anomaly detection model that allows the identification of potential data quality anomalies as a proactive approach to ensure high Big Data quality. In addition, the proposed model can be improved over time by being exposed to more data, making it particularly effective for detecting emerging data quality anomalies or detecting patterns that may evolve over time. Thus, the idea

TABLE 1. Related works.

Ref	Year	About Big Data?	Main findings
[4]	2023	Yes	The authors investigated how Big Data technologies impact the banking sector. They examined credit card inconsistencies and how the toolkit can be used to detect anomalies in a specific Wireless Application Protocol (WAP) instrument.
[5]	2022	Yes	The authors have investigated anomalies in the banking sector associated with Big Data, particularly anomalies in credit cards, and proposed a method for their detection and removal.
[6]	2022	Yes	The authors have presented new algorithms for detecting Big Data anomalies based on online learning and on distributed learning. The approach taken in the thesis is different from previous literature, as it employs the class-imbalance learning approach to tackle point anomalies.
[15]	2022	Yes	The authors have proposed a data quality tracking system based on a sequential combination of textual analysis methods to capture data anomalies. In addition, the authors have proposed using lightweight time series-based anomaly detection systems that can operate efficiently on growing dimensionality data, thus alleviating the burden of analyzing additional data without the need for extensive retraining on past data.
[16]	2022	Yes	The authors have proposed a Big Data quality assessment framework that allows for assessing 12 quality metrics related to Big Data. The authors have also defined a weighted Big Data quality score that applies data weights at multiple levels and defines and measures five quality aspects to provide a holistic view of data quality.
[17]	2022	Yes	The authors highlighted the importance of data quality in machine learning systems and the impact of data on their performance. They also raised the need for a data-centric approach in machine learning, where datasets are carefully designed and data quality is assessed and improved.
[18]	2021	Yes	To detect Network anomalies, the authors have proposed a method for classifying anomalies in Big Data streams that involves a two-step approach. Firstly, a dense stream algorithm generates class labels for the data stream. These predicted class labels are then utilized for training and classifying recurrent gated unit (GRU)-based recurrent neural networks, which can effectively organize the anomalies. The authors evaluated their algorithm using a network traffic dataset obtained from Cisco.
[19]	2021	Yes	To tackle the problem of inaccurate data in multi-source heterogeneous data settings, the authors proposed a data-cleaning approach known as hierarchical reduction and classification cleaning. This strategy involves constructing an augmented tree Bayesian Tan network using machine learning classification algorithms and attribute weight, with the Tan network's probability value used to classify accurate and imprecise data. The experimental analysis indicates that the proposed hierarchical reduction algorithm can extract the required data, particularly in complex multi-source heterogeneous data environments.
[20]	2021	No	The authors have proposed an unsupervised method for detecting anomalies by comparing models, using consensus learning, and combining heuristic rules with iterative hyperparameter tuning to improve data quality. A case study was performed to evaluate the effectiveness of this approach.
[21]	2021	Yes	The authors proposed a Big Data Quality Management Framework that focuses on enhancing pre-processing activities and data control through the use of a Big Data Quality Profile concept. The framework includes components for exploratory profiling, quality rule generation, and quality estimation, aiming to improve the efficiency and accuracy of data quality assessment before and after the pre-processing phase.
[22]	2020	No	The authors presented innovative methods to enhance the precision of Data Quality applications in High Energy Physics experiments. Specifically, they demonstrated the effectiveness of Machine Learning-based anomaly detection techniques in identifying unexpected detector malfunctions, using the CMS experiment at the Large Hadron Collider as a case study.
[23]	2019	Yes	The authors have presented a new algorithm, Stray (Search and Trace Anomaly), which aims to overcome the limitations of the HDoutliers algorithm. The Stray algorithm can detect anomalies in the original dataset and other data structures through feature engineering. The algorithm assigns an anomalous score to each data instance, indicating the extent of its outliers and providing a label. This paper presents experimental evidence to demonstrate the efficacy of the Stray algorithm.
[24]	2011	Yes	The authors have presented a novel approach to improve the computation time of the LOF (Local Outlier Factor) algorithm. In this methodology, the authors adapted the kd-tree indexing and approximated the nearest neighbor algorithm to reduce the computation time during LOF calculation. The authors also conducted a theoretical analysis on reaching the most immediate neighbor search. They demonstrated through experiments with accurate data that the proposed algorithm significantly reduces computation time while maintaining acceptable approximation errors.

behind this framework consists of formatting the dataset in a manner that uncovers the hidden anomalies related to each metric and highlights their deviation from the statistical norm of the dataset. In this framework, the dataset is transformed into several patterns, each highlighting and emphasizing the anomalies related to a quality metric, making them distinguishable in terms of values and, therefore, easily detectable. This data quality framework consists of four main phases: the first phase is the preprocessing phase, where the raw data is cleaned and prepared to be effectively used by the anomaly detection model. Then, the data is transformed and formatted into multiple patterns, each presenting the quality anomalies related to a specific quality metric as abnormal values, allowing their detection. The resulting prints are then attributed to the anomaly detection model that, based on an advanced isolation forest model, detects the anomalies related to each quality metric. Finally, the model computes an anomaly quality score for each abnormality detected and

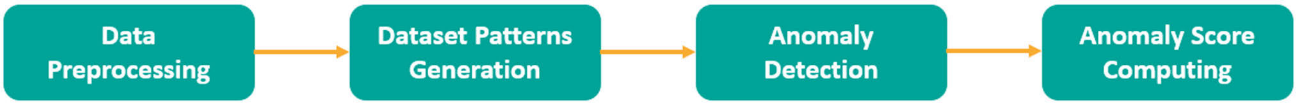

FIGURE 1. The big data quality anomaly detection framework.

computes the resulting anomaly score for each metric as well as the global anomaly score of the dataset. Figure 1 shows the different phases of the Big Data Quality Anomaly Detection Framework.

A. DATA PREPROCESSING

Data preprocessing is a crucial step in preparing data for anomaly detection. In the Big Data context, data preprocessing becomes more critical due to the high volume, velocity, and variety of data that must be processed. Moreover, big datasets are usually unstructured and poorly formatted, leading to a biased model and inaccurate results. Preprocessing is a crucial step in anomaly detection, as it helps to ensure that the data is accurate, consistent, and ready for analysis. Below, we outline the necessary data preprocessing tasks to enable precise anomaly detection in big datasets:

1) FEATURE SELECTION

Feature selection can be defined as the methodological process aimed at identifying and retaining the most significant features or variables from a dataset for analytical purposes, while discarding those that lack informative value [25]. The significance of feature selection becomes particularly pronounced in the context of anomaly detection, where it plays a crucial role in isolating the relevant features associated with anomalies, thereby enhancing the precision of predictive models. Moreover, this process enables faster and more accurate anomaly detection by reducing the amount of noisy data. Feature selection is performed first because it helps eliminate irrelevant and redundant features, thus making the subsequent preprocessing tasks more effective and efficient.

2) NORMALIZATION AND SCALING

As anomaly detection models are mainly based on the statistical distribution of the data values, they are highly affected by variations in the scale or range of the data. In the Big Data context, data variations issues are more common due to various data sources. The same feature may be represented differently depending on the scale each data source adopts [25]. Normalization and scaling are techniques used to address these issues by transforming the data into a consistent range of values, making detecting data points outside the expected range more accurate.

3) UPPERCASING/LOWERCASING

In the context of anomaly detection models, it is essential to consider case sensitivity to ensure a consistent and normalized data representation. Thus, we propose converting all text to either uppercase or lowercase format for a reliable and effective quality anomaly detection.

4) DIMENSIONALITY REDUCTION

Dimensionality reduction is a technique used to reduce the number of data features by combining them without losing information [26]. Dimensionality reduction is significant in Big Data where there is a large number of features making anomaly detection more difficult because of the size and the complexity of data. Thus, we suggest using dimensionality reduction to improve the anomaly detection process's scalability and performance by reducing the data's dimensionality.

5) STOP WORDS AND SYMBOL REMOVAL

Stop words are commonly used words in a language that do not carry much meaning or significance, such as ''the'', ''and'', ''of'', etc. The removal of stop words and punctuation marks is a crucial preprocessing step in anomaly detection, as it refines the data, focusing on meaningful patterns while reducing irrelevant elements from the text data. Notably, a recent survey on big data anomaly detection [27] emphasized the impact of input data noise on the efficacy of anomaly detection methods in the realm of Big Data. Furthermore, as part of this process, white spaces and specific placeholders like ''N/A,'' ''NA,'' ''NULL,'' and ''NaN'' values are also cleared to be recognized and treated as null values.

We consider these transformations critical in preparing the data for anomaly detection. Nonetheless, additional text-cleaning measures may be necessary based on the dataset's specifics. In addition, specific data cleaning tasks will be performed in the next step, where data are transformed into new patterns depending on the addressed metric.

B. DATASET PATTERNS GENERATION

Once the dataset is preprocessed and cleaned, the next step is generating new patterns by transforming the original dataset. Each dataset pattern is designed to highlight the deviations related to a specific data quality aspect and detect the quality anomalies about each metric. The following explains how the associated dataset pattern is generated for each data quality metric.

1) ACCURACY

Data Accuracy refers to the reliability and the correctness of the information contained within data. It measures the closeness of the collected data to the actual or expected value [28]. Examples of inaccurate values can be an age that exceeds 120 or a birth date in a future date, as these values are unusual and outside of the expected scale for these attributes. These values are called outliers and are inaccurate because they deviate significantly from a dataset's normal range of values. The presence of outliers in a dataset can lead to incorrect statistical analyses and can skew results. Outliers can occur due to errors in data collection, measurement errors, or wrong human entries. Detecting accuracy anomalies does not require a dataset transformation, as accuracy is related to data values and information contained within the data. Therefore, for this metric, the dataset should be kept in its original form, so the dataset will be assessed in terms of its values that can be accurate or non-accurate. Thus, directly applying the anomaly detection model to the dataset allows the detection of inaccurate outlier values.

2) CONFORMITY

This metric refers to data respecting the expected rules and constraints regarding data type [16]. Thus, a conform dataset means that all data values in a particular column or field have the same data type and format. Instances of nonconforming values may include a nominal value of "40k" instead of the expected numerical representation of "40000" in a given "Salary" column or the presence of percentages in a column designed to hold decimal values. For instance, in a "Profit Margin" column, some values may be "0.75", while others may be "75%". Thus, conformity anomalies refer to values that do not comply with the expected data type. Conformity anomalies occur especially in Big Data contexts where multiple data sources are gathered. Each uses a different data type and format for the same column. These anomalies should be addressed to avoid errors and inconsistencies in the data analysis and processing. As conformity is related to data type, conformity patterns should represent data values in terms of their class. There are three main types of data: numeric, string, and date. Thus, we generate the conformity pattern by converting each data value into three binary digits: The first digit refers to whether the entire data value is numeric, the second digit refers to whether the entire data value is alphabetic, and the third digit refers to whether the data value is a date. Therefore, a numeric value will be represented as "100", a string value as "010", a date value as "001", and an alphanumeric value as "000", as the data value is neither entirely numerical nor entirely alphabetical. Data values are represented in a binary format because the anomaly detection model easily understands and processes them. Thus, nonconform records with different data types will be represented differently and detected as outliers by the anomaly detection model. This representation can be extended with other data types depending on the used dataset.

3) COMPLETENESS

The collected raw data are usually messy and incomplete in Big Data environments. Data completeness ensures no missing values and refers to how data are sufficiently complete and fulfill the required information [29]. Completeness anomalies refer to missing values and can occur for many reasons, such as human non-response, incomplete data collection that may miss collecting some information, or data unavailability. The impact of missing data on analysis can be significant, especially if there is a considerable number of missing values, which can bias data analysis and lead to compromised results. Based on the definition of completeness, the completeness pattern should highlight the extent to which the dataset is complete. Thus, the completeness pattern is generated by converting each data value to one binary digit that indicates whether the data value is missing. Therefore, a data value will be represented as '1' if missing and otherwise as '0'. Thus, a missing data value will be expressed differently than the other existing values and be detected as an outlier by the anomaly detection model. As mentioned in the preprocessing phase, white spaces, 'N/A', 'NA', 'NULL', 'NaN' values are cleaned and therefore considered missing.

4) UNIQUENESS

Data Uniqueness refers to the fact that an actual entity should not be recorded more than once in a dataset [30]. A uniqueness anomaly is a redundancy of data entries referring to the same real-world entity. In large-scale datasets, redundancy is commonly observed as data is typically collected from various sources, resulting in multiple records of the same information in different formats. A dataset with redundant entries is misleading and inappropriate for accurate data analysis. Thus, the corresponding dataset pattern should highlight the degree of similarity between records to detect redundant entries as anomalies. For this, we used Record Linkage as a technique that consists of grouping records that likely refer to the same entity based on their values, comparing them, and then computing a similarity score for each pair of records. Thus, with a set of similarity scores of potentially similar records constituting our dataset pattern, the most similar records with a very high similarity score will be detected as anomalies by the anomaly detection model.

5) CONSISTENCY

Inconsistency refers to a lack of coherence or agreement between two or more records or information referring to the same entity. Thus, a consistency anomaly refers to records referring to the same world entity but with conflicting information [31]. An example of a consistency anomaly can be two records with conflicting information about the same real-world entity. For instance, two records have the same name and ID number but different addresses. Inconsistencies can occur for various reasons, such as data entry errors or when multiple data sources are used for data integration. To detect data consistency anomalies, potential redundant records should first be detected and then compared to identify consistency anomalies. Therefore, detecting consistency anomalies requires using the output of the Uniqueness anomaly detection as a dataset pattern to compare the similarity degree of the data values of the potential redundant records.

6) READABILITY

Data quality encompasses not only the format of the data but also its semantics. Data can include misspelled words and non-readable values, mainly when the database is populated with a large volume of human-generated data. As misspelled words can make it difficult for users to search, filter, or analyze the data accurately. The ability to extract information from data is called data readability [16]. Thus, data readability patterns should highlight how much data is readable and represent meaningful insights. The data readability pattern is generated by converting each data value to one binary digit that refers to whether the data value is readable using semantic libraries. Therefore, a data value will be represented as '1' if it is misspelled and as '0' otherwise. Thus, a non-readable data value will be expressed differently than the other existing values and detected as an anomaly by the anomaly detection model.

C. ANOMALY DETECTION

Once the dataset patterns are generated for each quality metric, the generated patterns are attributed to an anomaly detection model to identify the quality anomalies associated with each metric. This framework uses an extended isolation forest version as an anomaly detection model. Isolation Forest is an unsupervised learning algorithm that isolates anomalies from regular data points by creating isolation trees. Isolation Forest is a decision tree-based algorithm that produces a set of binary trees that recursively generates partitions by randomly selecting a feature and a split value within the range of that feature. The partitioning process will continue until it separates all the data points from the rest of the samples. This process generates shorter paths in the decision trees for anomalies since outliers will require fewer partitions on average to get isolated compared to standard instances, making them more distinguishable from the rest of the data. Using this partitioning approach, Isolation Forest can quickly identify outliers without building a normality model, as most outlier models do. This makes it more suitable, especially in a Big Data context when the expected behavior of data values is not well-defined. Another main point about Isolation Forest is its scalability, as it can handle data with many features, making it appropriate for Big Data applications with high-dimensional data. Moreover, Isolation Forest is not sensitive to the curse of dimensionality as some other Forest-based methods, so it can effectively identify outliers even in the presence of noisy or irrelevant features in the data, as in Big Data. Furthermore, Isolation Forest has a short computation time because it has a linear time complexity O(n), which makes it one of the best to deal with high-volume data sets. This framework uses an Extended Isolation Forest (EIF) version for outlier detection. The main difference between the algorithms is how the split branching is selected. In the original algorithm, the split branch can only be parallel to the axes making some regions with few or single observations split many times, which results in improper anomaly scores for some of the statements. In the extended version of Isolation Forest, the dataset is sliced using hyperplanes with random slopes. Indeed, instead of selecting an unexpected feature and value, it establishes a random gradient for the branching cut and a random intercept. This approach allows the model to capture more complex dependencies among components and handle high-dimensional data more effectively as it uses hyperplanes to isolate anomalies instead of straight lines. As mentioned earlier, the isolation forest is an unsupervised model as it does not require labeled data for training. The generation of the decision trees (The isolation Forest) represents the training phase of the model, so during scoring, a data point is traversed through all the trees which were trained earlier. A data point is then classified as an anomaly or non-anomaly based on the depth of all the trees required to reach the data point. In fact, anomaly data points mostly have a shorter tree path than the usual data points, as they are easily isolable. This framework applies the isolation model to each pattern to detect the anomalies associated with each quality metric. The model is applied separately to each column, allowing it to capture the unique properties of each feature and detect outliers more accurately. This is because the algorithm can focus on the most relevant features for each column and identify anomalies based on their unique values in that particular feature. Once all the dataset patterns are imputed to the model, the resulting detected anomalies are gathered into a single dataset, where all quality anomalies associated with each data value are identified.

D. QUALITY ANOMALY SCORE COMPUTING

The isolation forest consists of continuously splitting a subset of the dataset until isolating each data point. Then, the depth of the data point in each tree is determined. The depth of a data point is the number of splits required to isolate the point in the tree. A quality anomaly score for the data point is then determined as the average depth across all trees in the forest. The intuition behind this is that anomalous data points require fewer splits to be isolated and have a smaller average depth. Thus, the quality anomaly score measures how likely the data point is to be a quality outlier. Data points with higher anomaly scores are more likely to be of poor quality, while those with lower scores are more likely to be regular data points. For quality metrics where the dataset patterns are binary, namely: Completeness, Readability, and Conformity, the quality anomaly score for the detected anomalies is set to 100% as these quality metrics are categorical, so the associated quality anomaly score can either be 0% or 100%. Thus, each data point should have at the end the associated detected anomalies for each quality metric as well as their respective quality anomaly scores. For Instance, applying the quality anomaly detection to the data value "05/2030" of the Birth Date column, will lead to an outcome similar to: "Accuracy (48%), Conformity (75%)" implying that the data value is inaccurate by 48% and nonconform by 75%. Once the

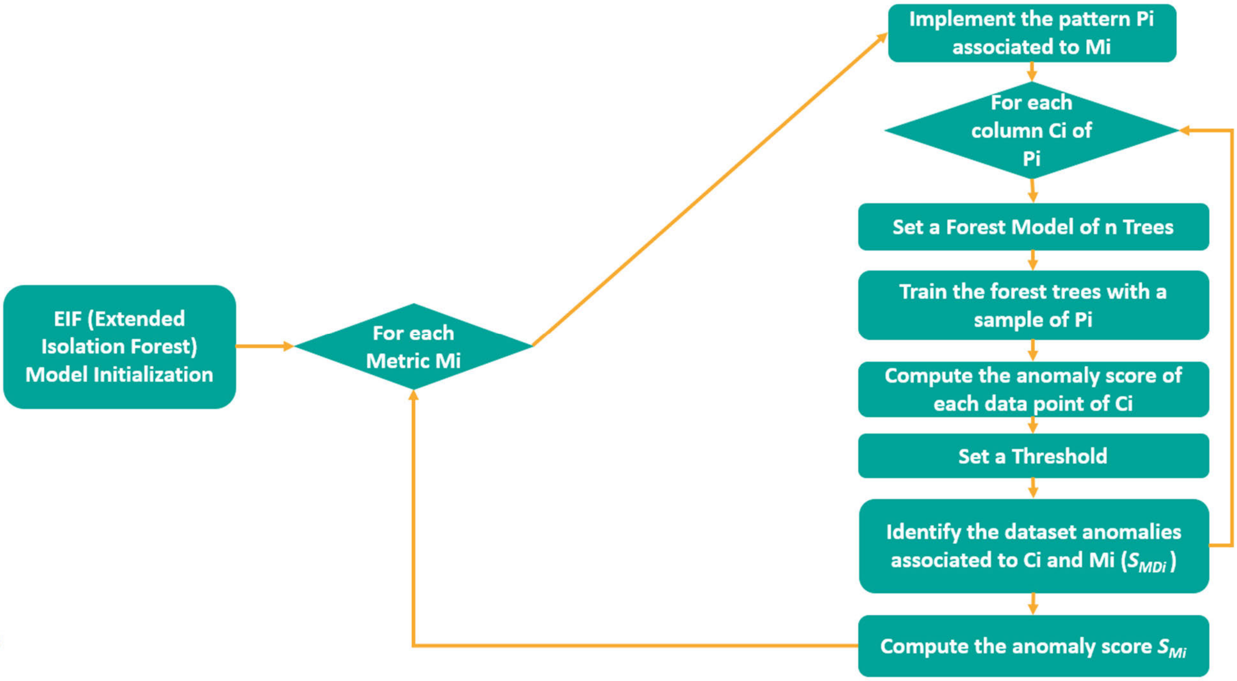
EIF (Extended Isolation Forest) Model Initialization
For each Metric Mi
Implement the pattern Pi associated to Mi
For each column Ci of Pi
Set a Forest Model of n Trees
Train the forest trees with a sample of Pi
Compute the anomaly score of each data point of Ci
Set a Threshold
Identify the dataset anomalies associated to Ci and Mi (S_{MDi})
Compute the anomaly score S_{Mi}

FIGURE 2. The big data quality anomaly detection framework pipeline.

quality anomalies are measured for all data values, a global anomaly score associated with each quality metric can be then deducted by averaging the quality anomaly scores of all the detected anomalies for that quality metric. Thus, each quality metric should have a resulting quality score that refers to the anomaly level in the entire dataset for that quality metric. Formally, the global anomaly score associated with a metric M is given by:

$$S_M = \frac{\sum_i S_{MDi}}{\sum_i D_i} \tag{1}$$

where S_{MDi} refers to the anomaly score of the detected D_i anomaly related to the Metric M. The quality anomaly score of the entire dataset refers to the degree of the overall anomalousness and the poor quality of the dataset. Moreover, a global quality anomaly score associated with the whole dataset can also be deducted as an average of the anomality scores related to each quality metric. For more precision, the global quality anomaly score of the dataset can be computed using a weighted average. As the quality metrics can be of varying relevancy, the most significant metrics can be attributed a higher weight to have a higher impact on the measured quality anomaly score. Formally, the global anomaly score is given by:

$$S_A = \frac{\sum_i w_i \times S_{Mi}}{\sum_i S_{Mi}} \tag{2}$$

where w_i refers to the weight attributed to each metric, and S_{Mi} refers to the average anomaly score related to the metric M. The quality anomaly score S_A of the dataset can provide helpful information about the overall level of an anomaly in the dataset. A high-quality anomaly score of the dataset implies that the dataset contains a large number of anomalous data points and therefore is of poor quality. Conversely, a low-quality anomaly score of the dataset means that the dataset includes relatively few anomalous data points and so, is of a good quality. Thus, the global quality anomaly score can be used as an indicator of the quality level of the dataset, and so to compare the anomalousness of different datasets. Moreover, it can be used as a quality diagnostic approach that helps estimate the effort required to clean the dataset based on the detected anomalies and their respective quality anomaly scores. Figure 2 shows a detailed pipeline for detecting quality anomalies in the proposed framework. In the next section, we present the implementation of the framework introduced in this section. Also, we present the tools and the datasets used for the implementation. Finally, a discussion is conducted about the obtained results as well as the possible evolutions.

V. IMPLEMENTATION

A. USE CASE AND DATASETS DESCRIPTION

In the present section, we describe the execution of the framework proposed in the preceding section. We have applied the suggested framework to two datasets:

1) DATASET 1

The first dataset is a Taxi and Limousine Commission (TLC) Trip Record dataset [32]. It is a big dataset that contains information on every taxi and for-hire vehicle trip in New York City. The dataset is collected and updated every month by the TLC, which regulates and licenses the city's taxi and for-hire vehicle industries. In this research, we addressed the green taxi trips data covering the year 2022. The collected dataset is a big dataset with over 1 million records that provides information on green taxi trips, including pick-up and drop-off locations, trip distances and durations, fares, payment types, and more. The dataset also provides information about payment data, such as payment type, fare amount, and type amount. However, the dataset contains many types of errors and inconsistencies that require careful cleaning and preprocessing, such as inaccurate and missing values, duplicate records, and non-conforming record types. The taxi trip data were captured by taxi meters, which were subject to numerous sources of errors, for instance, hardware and software malfunction, wireless signal issues, and human interference.

2) DATASET 2

The second dataset is a synthetic dataset of over 2 million records generated using a Python script [33]. The dataset simulates personal information and contains columns such as Name, Address, Gender, Age, Salary, etc. Anomalies were then introduced by changing the data distribution with outlier values to address accuracy metric, modifying data values, and introducing missing or incorrect data to address the completeness, conformity, and readability metrics. Additionally, 10,000 random rows were slightly modified and added as duplications to address Uniqueness and Consistency metrics. The reason behind using synthetic dataset is the non-availability of prelabelled dataset with known quality anomalies. A dataset with foreknown anomalies is required to evaluate the suggested framework's accuracy. However, labeling can be tedious and time-consuming, especially for a large dataset with over 1 million records. Therefore, we built a synthetic dataset with foreknown anomalies to assess the performance of our suggested framework. Another reason for using a synthetic dataset is to assess our framework's performance and limitations by introducing challenging and confusing data anomalies. Nevertheless, to assess our framework for the real-world dataset, a representative sample of 100 000 records in the first dataset were manually labeled as anomalies or non-anomalies to measure the Precision and the framework's accuracy for the first dataset. Table 2 presents the characteristics of the two datasets used for our experiments. The following section presents the tools and techniques used in our implementation.

TABLE 2. Datasets description.

Dataset	Number of Records	Number of columns
Dataset 1	1 370 000	20
Dataset 2	2M	9

B. ADOPTED ARCHITECTURE AND LIBRARIES

1) ARCHITECTURE

The framework was implemented in the Cloudera Data Platform [34], a platform for Big Data processing. Cloudera Data Platform includes various tools and technologies for processing and analyzing large-scale datasets in a scalable and secure way. This Big Data platform was used for this implementation as it has improved performance through faster data processing and is compatible with various Big Data tools such as Hadoop, Spark, HBase, Kafka, etc.

Data storage was managed in a distributed mode using Hadoop Distributed File System (HDFS), which provides scalable Big Data storage. HDFS is designed to scale horizontally, so it can easily accommodate growing data volumes by adding more nodes to the cluster. Another advantage of HDFS is its fault tolerance, as it is designed to replicate data across multiple nodes in the cluster to ensure that data is not lost in case of node failures. Apache Spark was used for data processing, especially Pyspark, as a Python API of Spark. Spark is a robust Big Data processing framework designed to be faster than traditional MapReduce processing in Hadoop. In this implementation, Spark was used with the Hadoop YARN cluster for scalable and parallel data processing, allowing us to take advantage of Hadoop's scalability and fault tolerance. In the adopted architecture, Zeppelin was used as a development platform monitored by Ambari to ensure an efficient operation orchestration. Figure 3 shows the global architecture used in the implementation based on Cloudera Data Platform [34].

2) LIBRARIES

Data preprocessing involved using string functions and various Python libraries for data preprocessing, including NLTK (Natural Language Toolkit), DateTime, and RE (Regular Expressions). Also, dataset patterns were generated using Pyspark functions to convert the original data into the appropriate form. The anomalies related to the Uniqueness metric are detected using Record Linkage. Thus, data were indexed using Sorted Neighborhood to compute the similarity score of potential records. In this approach, each record is represented as a vector of features, and a weighted similarity score between records is computed using Jaccard similarity and cosine similarity. According to a recent investigation [35], various indexing methods for scalable linkage were compared, and the results demonstrated that sorted neighborhood is the most suitable technique for indexing Big Data. The Isolation Forest model was implemented using the Sparkling Water library [36]. Sparkling Water is an open-source library that integrates Apache Spark and H2O.ai [37], an open-source machine learning platform that allows combining the fast, scalable machine learning algorithms of H2O with the capabilities of Spark. Thus, an isolation forest model of 100 trees with a sample size of 265 was initialized. The sample size refers to the number of randomly sampled observations used to train the isolation tree. As mentioned, the model was credited with the appropriate dataset pattern for each quality metric. Then, the iterative application of the isolation model to each dataset pattern column resulted in the detection of quality anomalies associated with that specific metric. Firstly, a subset of the dataset was analyzed to

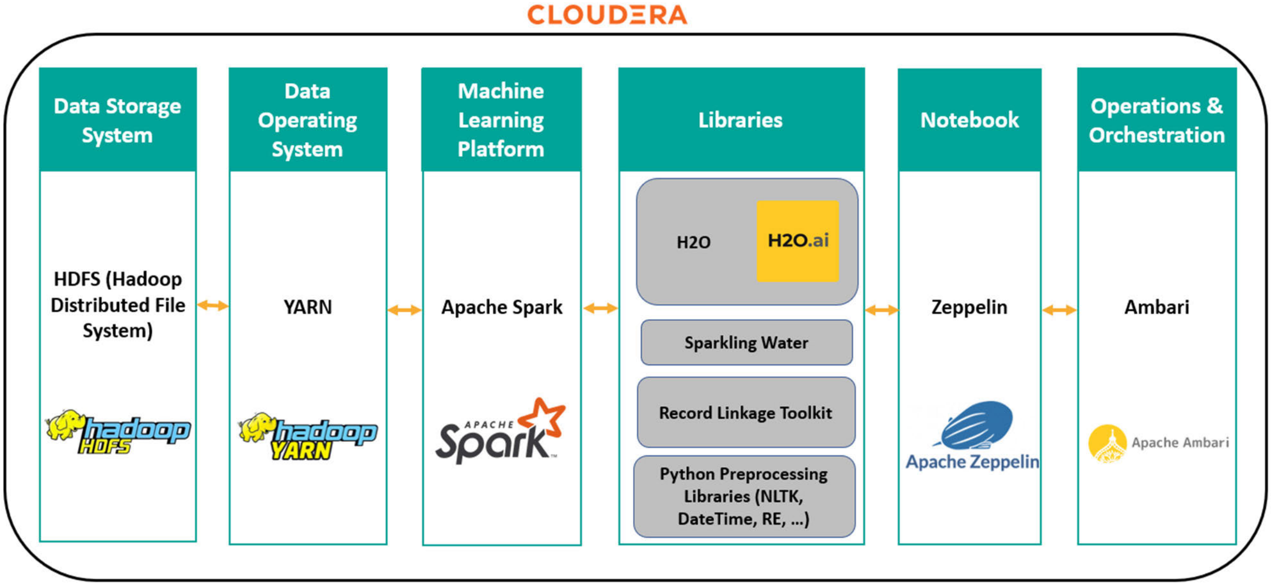

FIGURE 3. Implementation architecture.

determine the proportion of outliers present in each dataset. Then, the corresponding quantile of the score was estimated and used as a threshold for predicting quality anomalies. Then, the anomalies were detected, and the related quality anomaly score was measured. Finally, each metric's corresponding quality anomaly score and global quality score were calculated.

C. RESULTS

1) ACCURACY

Both datasets were first preprocessed by converting the data values to the appropriate formats not to mislead the anomaly detection model, removing white spaces and stop words, and extracting new features based on the existing ones to detect the hidden anomalies. Then, the dataset patterns corresponding to each metric were generated for both datasets. Finally, the anomaly detection model was applied to each dataset pattern to detect the anomalies related to the corresponding metric. The model was used for each metric, and then the obtained results were labeled manually. For the first dataset, we have biased the dataset with additional anomalies by deleting some values and converting some data values to a non-conform or unreadable format. The framework's performance was evaluated using the metrics of the confusion matrix:

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

$$F1 - score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{5}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

TP, FP, TN, and FN are True Positive, False Positive, True Negative, and False Negative, respectively. We can understand how well the suggested framework identifies anomalies by evaluating these metrics. For example, a high precision score indicates that the algorithm accurately detects anomalies and generates few false positives. On the other hand, a high recall score suggests that the algorithm is seeing a high proportion of actual anomalies in the dataset but may also generate many false positives. The F1-score and the Accuracy metrics measure the overall performance and the model's effectiveness in detecting anomalies. The results for each quality metric were computed by averaging the resulting metric of all columns. The measurement of the readability metric was considered less relevant for the first dataset, given that most of the data entries do not consist of textual content and thus lacked potential readability anomalies. Consequently, no measurement of the readability metric was conducted for the first dataset. Tables 3 and 4 show the obtained results for each quality metric for the first and second datasets.

As mentioned earlier, the anomaly score was also computed as part of the suggested framework. First, the anomaly score for each metric and each column was measured based on the obtained anomaly score of the detected anomalies. Then, the global anomaly score related to each metric was calculated by averaging anomaly scores for each column. Finally, the global anomaly score for each dataset has been deducted as the average of the anomaly score of all metrics. In this implementation, all metrics were considered equally when measuring the global anomaly score, so no weights were applied to the global anomaly score. Table 5 shows the obtained anomaly scores for both datasets for each metric as well as the global anomaly score.

TABLE 3. Dataset 1 confusion matrix.

Metrics	Accuracy	Conformity	Completeness	Uniqueness	Consistency	Readability	Average Metrics
Precision	96.00%	93.46%	100%	87.5%	88.89%	_	94.30%
Recall	87.5%	95.71%	100%	77.68%	100%	_	93.48%
F1-score	90.77%	94.26%	100%	82.35%	94.12%	_	93.58%
Accuracy	99.98%	99.97%	100%	99.99%	97.84%	_	99.63%

TABLE 4. Dataset 2 confusion matrix.

Metrics	Accuracy	Conformity	Completeness	Uniqueness	Consistency	Readability	Average Metrics
Precision	99.71%	99.51%	100%	99.46%	97.50%	99.07%	99.20%
Recall	98.52%	98.93%	100%	92.70%	100%	91.85%	97.00%
F1-score	99.20%	99.22%	100%	95.96%	98.73%	95.32%	98.07%
Accuracy	99.99%	99.99%	100%	99.99%	99.54%	99.98%	99.91%

TABLE 5. Anomaly scores.

Metrics	Accuracy	Conformity	Uniqueness	Consistency	Completeness	Readability	Global Score
Dataset 1	79.21%	100%	74.78%	89.53%	100 %	--	88.70%
Dataset 2	82.33%	100%	76.70%	92.01%	100 %	100%	91.84%

TABLE 6. Execution time.

Datasets	Number of Records	Execution Time of the Anomaly Detection	Execution Time of the Whole Framework
Dataset 1	1 370 000	1.6 h	2.4 h
Dataset 2	2M	2.76 h	5.83 h

2) SCALABILITY

Since the framework is designed to handle Big Data, its scalability must also be assessed. For this, we computed the processing time of the model to detect anomalies as well as the execution time of the whole framework's pipeline, including the preprocessing, the datasets pattern generations as well as anomaly detection, and anomaly score computing. Table 6 shows the processing time for both datasets. The framework demonstrated favorable processing time outcomes with a linear O(n) complexity. As mentioned earlier, the framework was implemented using appropriate tools for Big Data in terms of architecture, such as HDFS, Spark, and distributed processing with Hadoop. Also, scalable libraries and models were used in the implementation, such as the isolation forest of Sparkling water for anomaly detection and the sorted neighborhood for indexing.

3) DISCUSSION

Based on the obtained results, the framework has achieved a good score in terms of accuracy and scalability. The framework was able to detect most of the quality anomalies. Only extreme cases of anomalies were not detected. The not detected anomalies were somewhat related to semantics. For the first dataset, categorical encoded features were excluded from detecting anomalies related to the accuracy metric as there are no outlier values, so the model was only applied to continuous elements. In the first dataset, the model detected outlier values such as distance trips exceeding 50 miles or below 0, duration exceeding 2 hours, or fare over 300 $. For the completeness metric, the model detected all the missing values. For Uniqueness, among the nine duplicated records that were located based on the pickup datetime, locations, and other features, seven were detected as anomalies by the model. Then, each pair of records were compared to detect consistency anomalies. The model was able to see the differences between the pairs. However, data values with the same information but different formats were also detected as inconsistent since the framework is based on a comparison of data values to identify consistency anomalies. Hence, this observation highlights an area for potential improvement in the framework's methodology. In the second dataset, the framework has achieved a good quality score for the accuracy metric. The framework can see inaccurate data values for numeric and date values, but not for text data. In fact, the framework was not able to detect the erroneous text data values for the text values that have a non-appropriate text format have been seen as nonconform data values; however, inaccurate text values with a correct format were not detected as anomalies except those which were in duplicated records that were detected as inconsistent. The framework was able to catch most of the non-conform data values due to the generated dataset pattern. To evaluate the model's performance regarding the uniqueness metric, we have infiltered the dataset with non-identical duplicated records and some challenging duplicates where most columns were partially or completely modified to assess the framework limitations, requiring intelligent detection techniques. The framework has achieved a good score for detecting duplicate records as anomalies. Only some extreme cases of duplicate records were not accurately detected, where all column values are completely inconsistent but still refer to the same real-world entity. As for the first dataset, the model can see all missing values and has achieved an excellent readability metric score. In this paper, our proposed framework was not compared to existing methodologies since no anomaly detection framework currently addresses data quality anomalies. Indeed, given that our framework focuses on data quality anomalies, it would be inappropriate to compare its performance to existing anomaly detection frameworks with different scopes. Nevertheless, the proposed framework outperforms the current anomaly detection frameworks in its scope as it not only detects outlier values but also effectively tackles a broad range of generic data quality anomalies pertaining to six data quality metrics. Moreover, the framework has demonstrated promising results in terms of accuracy based on the measured confusion matrix metrics for both datasets. Furthermore, the framework was designed to fit Big Data features and demonstrated favorable processing time outcomes for scalability with a linear O(n) complexity.

VI. CONCLUSION AND FUTURE WORKS

Big Data can significantly enhance organizational operations and business performance across various fields. However, its advantages can only be fully realized if the data quality is improved. While many anomaly detection frameworks were suggested in the literature to address anomalies related to outlier values for different fields, they do not consider anomalies related to data quality, such as missing, incorrect, or inconsistent data. Such anomalies can lead to incorrect conclusions and hinder the ability of organizations to draw meaningful insights from their data. To address this raised issue, we suggest a comprehensive approach for Big Data Quality Anomaly Detection that allows the detection of generic data quality anomalies for Big Data without any correlation to a specific field. The suggested system enables the detection of quality anomalies related to 6 quality dimensions: Accuracy, Consistency, Completeness, Conformity, Uniqueness, and Readability. Moreover, we have defined and measured a new metric called ‘‘Quality Anomaly Score,’’ which refers to the degree of the anomalousness of the quality anomalies. This metric was calculated for the detected quality anomalies, the quality dimensions, and the whole dataset. This framework was implemented using two datasets and has shown acceptable results in terms of accuracy and scalability, with an accuracy score of up to 99.91% and an F1-score of 98.07%. In future work, we aim to extend the suggested framework by addressing the detected anomalies and automatically correcting them to the appropriate data value rather than simply removing them, which will improve the dataset's quality. Automatically correcting anomalies will improve the dataset's overall quality and guarantee reliable and comprehensive information for subsequent analyses and decision-making. Moreover, it will save valuable time and effort that would otherwise be required for manual inspection and correction of anomalies.

REFERENCES

[1] I. E. Alaoui, Y. Gahi, and R. Messoussi, ‘‘Big data quality metrics for sentiment analysis approaches,’’ in *Proc. Int. Conf. Big Data Eng.* New York, NY, USA: Association for Computing Machinery, Jun. 2019, pp. 36–43, doi: 10.1145/3341620.3341629.

[2] A. Z. Faroukhi, I. E. Alaoui, Y. Gahi, and A. Amine, ‘‘An adaptable big data value chain framework for end-to-end big data monetization,’’ *Big Data Cogn. Comput.*, vol. 4, no. 4, p. 34, Nov. 2020, doi: 10.3390/bdcc4040034.

[3] N. Seralina and A. Akzhalova, ‘‘Anomaly detection framework,’’ in *Innovations in Bio-Inspired Computing and Applications* (Lecture Notes in Networks and Systems), A. Abraham, A. Bajaj, N. Gandhi, A. M. Madureira, and C. Kahraman, Eds. Cham, Switzerland: Springer, 2023, pp. 75–85, doi: 10.1007/978-3-031-27499-2_7.

[4] V. V. Keskar, J. Yadav, and A. Kumar, ‘‘Enhancing data quality by detecting and repairing inconsistencies in big data,’’ in *Proc. 2nd Int. Conf. Mech. Energy Technol.*, in Smart Innovation, Systems and Technologies, S. Yadav, A. Haleem, P. K. Arora, and H. Kumar, Eds. Singapore: Springer, 2023, pp. 185–197, doi: 10.1007/978-981-19-0108-9_20.

[5] V. Keskar, J. Yadav, and A. Kumar, ‘‘Perspective of anomaly detection in big data for data quality improvement,’’ *Mater. Today, Proc.*, vol. 51, pp. 532–537, Jan. 2022, doi: 10.1016/j.matpr.2021.05.597.

[6] C. Maurya, ‘‘Anomaly detection in big data,’’ Ph.D. thesis, Indian Inst. Technol., Roorkee, India, Mar. 2022.

[7] I. E. Alaoui and Y. Gahi, ‘‘The impact of big data quality on sentiment analysis approaches,’’ *Proc. Comput. Sci.*, vol. 160, pp. 803–810, Jan. 2019, doi: 10.1016/j.procs.2019.11.007.

[8] W. Elouataoui, I. E. Alaoui, and Y. Gahi, ‘‘Metadata quality in the era of big data and unstructured content,’’ in *Advances in Information, Communication and Cybersecurity* (Lecture Notes in Networks and Systems), vol. 357. Cham, Switzerland: Springer, 2022, pp. 145–157, doi: 10.1007/978-3-030-91738-8_11

[9] I. E. Alaoui, Y. Gahi, and R. Messoussi, ‘‘Full consideration of big data characteristics in sentiment analysis context,’’ in *Proc. IEEE 4th Int. Conf. Cloud Comput. Big Data Anal. (ICCCBDA)*, Apr. 2019, pp. 126–130, doi: 10.1109/ICCCBDA.2019.8725728.

[10] W. Elouataoui, I. E. Alaoui, and Y. Gahi, ‘‘Data quality in the era of big data: A global review,’’ in *Big Data Intelligence for Smart Applications* (Studies in Computational Intelligence), vol. 994, Y. Baddi, Y. Gahi, Y. Maleh, M. Alazab, and L. Tawalbeh, Eds. Cham, Switzerland: Springer, 2022, pp. 1–25, doi: 10.1007/978-3-030-87954-9_1.

[11] Y. Gahi and I. E. Alaoui, ‘‘Machine learning and deep learning models for big data issues,’’ in *Machine Intelligence and Big Data Analytics for Cybersecurity Applications* (Studies in Computational Intelligence), Y. Maleh, M. Shojafar, M. Alazab, and Y. Baddi, Eds. Cham, Switzerland: Springer, 2021, pp. 29–49, doi: 10.1007/978-3-030-57024-8_2.

[12] M. He, M. Petering, P. LaCasse, W. Otieno, and F. Maturana, ‘‘Learning with supervised data for anomaly detection in smart manufacturing,’’ *Int. J. Comput. Integr. Manuf.*, vol. 36, no. 9, pp. 1331–1344, Feb. 2023, doi: 10.1080/0951192X.2023.2177747.

[13] M. Shahin, S. A. Peious, R. Sharma, M. Kaushik, S. B. Yahia, S. A. Shah, and D. Draheim, ‘‘Big data analytics in association rule mining: A systematic literature review,’’ in *Proc. 3rd Int. Conf. Big Data Eng. Technol. (BDET)*. New York, NY, USA: Association for Computing Machinery, Oct. 2021, pp. 40–49, doi: 10.1145/3474944.3474951.

[14] S. A. Shah, D. Z. Seker, S. Hameed, and D. Draheim, ‘‘The rising role of big data analytics and IoT in disaster management: Recent advances, taxonomy and prospects,’’ *IEEE Access*, vol. 7, pp. 54595–54614, 2019, doi: 10.1109/ACCESS.2019.2913340.

[15] P. Mahajan, ‘‘Textual data quality at scale for high dimensionality data,’’ in *Proc. Int. Conf. Data Sci., Agents Artif. Intell. (ICDSAAI)*, Dec. 2022, pp. 1–4, doi: 10.1109/ICDSAAI55433.2022.10028903.

[16] W. Elouataoui, I. El Alaoui, S. El Mendili, and Y. Gahi, ‘‘An advanced big data quality framework based on weighted metrics,’’ *Big Data Cogn. Comput.*, vol. 6, no. 4, p. 153, Dec. 2022, doi: 10.3390/bdcc6040153.

[17] H. Chen, J. Ding, W. Lu, and J. Bhuyan, ‘‘Data quality for big data and machine learning,’’ *Frontiers Big Data*, 2022. [Online]. Available: https://www.frontiersin.org/researchtopics/36569/evaluating-data-quality-in-research

[18] R. Surapaneni, S. Nimmagadda, and K. Pragathi, "Unsupervised classification approach for anomaly detection in big data streams," in *Next Generation of Internet of Things*. 2021, pp. 71–79, doi: 10.1007/978-981-16-0666-3_8.

[19] Z. Ying, Y. Huang, K. Chen, and T. Yu, "Big data cleaning model of multi-source heterogeneous power grid based on machine learning classification algorithm," *J. Phys., Conf. Ser.*, vol. 2087, no. 1, Nov. 2021, Art. no. 012095, doi: 10.1088/1742-6596/2087/1/012095.

[20] L. Poon, S. Farshidi, N. Li, and Z. Zhao, "Unsupervised anomaly detection in data quality control," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2021, pp. 2327–2336, doi: 10.1109/BigData52589.2021.9671672.

[21] I. Taleb, M. A. Serhani, C. Bouhaddioui, and R. Dssouli, "Big data quality framework: A holistic approach to continuous quality management," *J. Big Data*, vol. 8, no. 1, p. 76, May 2021, doi: 10.1186/s40537-021-00468-0.

[22] A. A. Pol, G. Cerminara, C. Germain, and M. Pierini, "Data quality monitoring anomaly detection," in *Artificial Intelligence for High Energy Physics*. Singapore: World Scientific, 2020, pp. 115–149, doi: 10.1142/9789811234033_0005.

[23] P. D. Talagala, R. J. Hyndman, and K. Smith-Miles, "Anomaly detection in high-dimensional data," *J. Comput. Graph. Statist.*, vol. 30, no. 2, pp. 360–374, Apr. 2021, doi: 10.1080/10618600.2020.1807997.

[24] S. Kim, N. W. Cho, B. Kang, and S.-H. Kang, "Fast outlier detection for very large log data," *Expert Syst. Appl.*, vol. 38, no. 8, pp. 9587–9596, Aug. 2011, doi: 10.1016/j.eswa.2011.01.162.

[25] S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, "Big data preprocessing: Methods and prospects," *Big Data Anal.*, vol. 1, no. 1, p. 9, Nov. 2016, doi: 10.1186/s41044-016-0014-0.

[26] A. A. Kumar and S. Chandrasekhar, "Text data pre-processing and dimensionality reduction techniques for document clustering," *Int. J. Eng. Res. Technol.*, vol. 1, no. 5, pp. 1–6, Aug. 2012, doi: 10.17577/IJERTV1IS5278.

[27] R. A. A. Habeeb, F. Nasaruddin, A. Gani, I. A. Targio Hashem, E. Ahmed, and M. Imran, "Real-time big data processing for anomaly detection: A survey," *Int. J. Inf. Manage.*, vol. 45, pp. 289–307, Apr. 2019, doi: 10.1016/j.ijinfomgt.2018.08.006.

[28] G. Mylavarapu, J. P. Thomas, and K. A. Viswanathan, "An automated big data accuracy assessment tool," in *Proc. IEEE 4th Int. Conf. Big Data Anal. (ICBDA)*, Mar. 2019, pp. 193–197, doi: 10.1109/ICBDA.2019.8713218.

[29] W. Elouataoui, I. E. Alaoui, and Y. Gahi, "Metadata quality dimensions for big data use cases," in *Proc. 2nd Int. Conf. Big Data, Modelling Mach. Learn. (BML)*. Setúbal, Portugal: SciTePress, 2022, pp. 488–495, doi: 10.5220/0010737400003101.

[30] W. Elouataoui, I. E. Alaoui, S. E. Mendili, and Y. Gahi, "An end-to-end big data deduplication framework based on online continuous learning," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 9, 2022, doi: 10.14569/IJACSA.2022.0130933.

[31] G. Mylavarapu, K. A. Viswanathan, and J. P. Thomas, "Assessing context-aware data consistency," in *Proc. IEEE/ACS 16th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Nov. 2019, pp. 1–6, doi: 10.1109/AICCSA47632.2019.9035250.

[32] *TLC Trip Record Data—TLC*. Accessed: Mar. 30, 2023. [Online]. Available: https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page

[33] *Synthetic Big Dataset for Anomaly Detection*. Accessed: Mar. 29, 2023. [Online]. Available: https://www.kaggle.com/datasets/elouataouiwidad/synthetic-bigdataset-anomalydetection

[34] Cloudera. (2023). *Cloudera Data Platform (CDP)*. Accessed: Jul. 30, 2023. [Online]. Available: https://www.cloudera.com/products/cloudera-data-platform.html

[35] S. Yeddula and K. Lakshmaiah, "Investigation of techniques for efficient & accurate indexing for scalable record linkage & deduplication," *Int. J. Comput. Commun. Technol.*, vol. 6, no. 1, pp. 24–30, Sep. 2020, doi: 10.47893/IJCCT.2015.1275.

[36] S. Hariri, M. C. Kind, and R. J. Brunner, "Extended isolation forest," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 4, pp. 1479–1489, Apr. 2021, doi: 10.1109/TKDE.2019.2947676.

[37] *H2O.ai | The Fastest, Most Accurate AI Cloud Platform*. Accessed: Mar. 30, 2023. [Online]. Available: https://h2o.ai/

ELOUATAOUI WIDAD received the Engineering degree in software engineering from the National School of Applied Sciences (ENSA), Ibn Tofail University, Kenitra, Morocco, in 2020, where she is currently pursuing the Ph.D. degree. She is an international firm as a Customer Relationship Management (CRM) Consultant. Her current research interests include big data management and big data quality using machine learning and artificial intelligence.

ELMENDILI SAIDA received the State Engineering Diploma degree in computer engineering from Cadi Ayyad University, Morocco, and the Ph.D. degree in computer science from Ibn Tofail University, Morocco. She is currently an Assistant Professor with the Institute of Sport Professions, Ibn Tofail University. Her current research interests include artificial intelligence, big data analytics, machine learning, and smart city.

YOUSSEF GAHI (Senior Member, IEEE) is currently an Associate Professor of computer science with Ibn Tofail University, Morocco. He is a passionate Data Transformation Specialist with extensive big data and data management expertise. He has more than 18 years of experience in academia, research, corporate business, and the IT industry. He has been dedicated to exploring the limitless possibilities that data offers in transforming businesses and shaping the future. His current research interests include big data management, big data quality, big data security, recommendation systems, and cloud computing models for data privacy.