

7-1-2023

## Empowering Patient Similarity Networks through Innovative Data-Quality-Aware Federated Profiling

Alramzana Nujum Navaz  
*United Arab Emirates University*

Mohamed Adel Serhani  
*University of Sharjah*

Hadeel T. El Kassabi  
*Humber Institute of Technology and Advanced Learning*

Ikbal Taleb  
*Zayed University, ikbal.taleb@zu.ac.ae*

Follow this and additional works at: <https://zuscholars.zu.ac.ae/works>



Part of the [Computer Sciences Commons](#), and the [Medicine and Health Sciences Commons](#)

---

### Recommended Citation

Navaz, Alramzana Nujum; Serhani, Mohamed Adel; El Kassabi, Hadeel T.; and Taleb, Ikbal, "Empowering Patient Similarity Networks through Innovative Data-Quality-Aware Federated Profiling" (2023). *All Works*. 6014.

<https://zuscholars.zu.ac.ae/works/6014>

This Article is brought to you for free and open access by ZU Scholars. It has been accepted for inclusion in All Works by an authorized administrator of ZU Scholars. For more information, please contact [scholars@zu.ac.ae](mailto:scholars@zu.ac.ae).

## Article

# Empowering Patient Similarity Networks through Innovative Data-Quality-Aware Federated Profiling

Alramzana Nujum Navaz <sup>1,\*</sup>, Mohamed Adel Serhani <sup>2</sup>, Hadeel T. El Kassabi <sup>3</sup> and Ikbal Taleb <sup>4</sup>

<sup>1</sup> Department of Computer Science and Software Engineering, College of Information Technology, UAE University, Al Ain P.O. Box 15551, United Arab Emirates

<sup>2</sup> College of Computing and Informatics, Sharjah University, Sharjah P.O. Box 27272, United Arab Emirates; mserhani@sharjah.ac.ae or serhanim@uaeu.ac.ae

<sup>3</sup> Faculty of Applied Sciences & Technology, Humber College Institute of Technology & Advanced Learning, Toronto, ON M9W 5L7, Canada; hadeel.el-kassabi@humber.ca

<sup>4</sup> College of Technological Innovation, Zayed University, Abu Dhabi P.O. Box 144534, United Arab Emirates; ikbal.taleb@zu.ac.ae

\* Correspondence: 201570182@uaeu.ac.ae

**Abstract:** Continuous monitoring of patients involves collecting and analyzing sensory data from a multitude of sources. To overcome communication overhead, ensure data privacy and security, reduce data loss, and maintain efficient resource usage, the processing and analytics are moved close to where the data are located (e.g., the edge). However, data quality (DQ) can be degraded because of imprecise or malfunctioning sensors, dynamic changes in the environment, transmission failures, or delays. Therefore, it is crucial to keep an eye on data quality and spot problems as quickly as possible, so that they do not mislead clinical judgments and lead to the wrong course of action. In this article, a novel approach called federated data quality profiling (FDQP) is proposed to assess the quality of the data at the edge. FDQP is inspired by federated learning (FL) and serves as a condensed document or a guide for node data quality assurance. The FDQP formal model is developed to capture the quality dimensions specified in the data quality profile (DQP). The proposed approach uses federated feature selection to improve classifier precision and rank features based on criteria such as feature value, outlier percentage, and missing data percentage. Extensive experimentation using a fetal dataset split into different edge nodes and a set of scenarios were carefully chosen to evaluate the proposed FDQP model. The results of the experiments demonstrated that the proposed FDQP approach positively improved the DQ, and thus, impacted the accuracy of the federated patient similarity network (FPSN)-based machine learning models. The proposed data-quality-aware federated PSN architecture leveraging FDQP model with data collected from edge nodes can effectively improve the data quality and accuracy of the federated patient similarity network (FPSN)-based machine learning models. Our profiling algorithm used lightweight profile exchange instead of full data processing at the edge, which resulted in optimal data quality achievement, thus improving efficiency. Overall, FDQP is an effective method for assessing data quality in the edge computing environment, and we believe that the proposed approach can be applied to other scenarios beyond patient monitoring.

**Keywords:** federated learning; federated profiling; patient similarity network; federated patient similarity network; data quality profiling; deep learning; edge computing; eHealth



**Citation:** Navaz, A.N.; Serhani, M.A.; El Kassabi, H.T.; Taleb, I. Empowering Patient Similarity Networks through Innovative Data-Quality-Aware Federated Profiling. *Sensors* **2023**, *23*, 6443. <https://doi.org/10.3390/s23146443>

Academic Editors: Flavio Bertini, Rahimeh Rouhi and Enrique Lopez Droguett

Received: 13 June 2023

Revised: 12 July 2023

Accepted: 13 July 2023

Published: 16 July 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

As the internet of things (IoT) became more pervasive, it is evident that big data capabilities are undergoing a revolution, with enhanced domain sensing capabilities. Nevertheless, many IoT-related projects are hampered by real-time connectivity issues and insufficient computing power to handle the ever-increasing volume of information processing. Furthermore, limitations in data transport capabilities further exacerbate these

challenges, necessitating the execution of complex data analysis on heterogeneous computing platforms. These constraints impose a considerable risk of information loss during data processing, particularly when employing aggregations, approximations, and filtrations to overcome resource limitations. This, in turn, has a direct impact on data accuracy, as the outcomes of data processing become susceptible to inaccuracies and uncertainties [1]. In addition to environmental factors, challenges during data creation and collection at the data sources add to the complexities, as IoT generates a massive volume of data that needs to be efficiently collected, stored, processed, and analyzed. Factors such as reduced sensor precision, communication latencies, short battery life, and limited availability of sensor/actuator sets contribute to diminished data accuracy. Additionally, the potential for data breaches and privacy concerns due to the vast amount of sensitive data collected and transmitted further compounds the need for robust security measures to protect the integrity and confidentiality of the collected data. These challenges must be addressed to ensure the successful implementation and operation of IoT projects, particularly in domains where data accuracy, real-time analysis, and efficient data transport are crucial. To optimize IoT and mobile edge computing (MEC) architectures, techniques such as feature selection and data fusion can be leveraged, while also considering strategies for energy conservation and energy harvesting [2]. These approaches aim to improve the quality of low-cost sensor data, enhancing the overall performance and reliability of IoT systems [3]. By overcoming these obstacles, IoT projects can unlock their full potential and effectively contribute to data-driven decision making in various domains.

In the field of healthcare, the influence of data quality (DQ) flaws on physician judgments decreases the likelihood of patients receiving optimal treatment, jeopardizing their health and well-being. While poorly designed DQ attributes result in miscalculations and misinterpretations that can have significant negative effects on healthcare providers and patients, only a few studies document the actual effects of DQ flaws on patient care decisions. This includes reduced validity of critical clinical characteristics and incomplete data, which affects the propensity to prescribe medications or perform invasive procedures [4]. Consequently, it is essential to measure the impact of DQ issues on clinical decisions and emphasize their relative importance. DQ issues must be carefully addressed at the source and effectively managed before they can affect the clinical decision-making process in order to preserve the integrity of the generated data and support the highest standards of clinical decision making.

Gartner [5] estimated that 60% of organizations will utilize machine-learning-enabled DQ technology by 2022 to reduce human operations for data quality improvement, and 50% will use data quality solutions by 2024 to promote digital business. DQ is critical for tracking data value and relevance, and we believe its use in quantifying data will give us a handle on what data are available, what the data's value might be for business decision making, and whether the data should be assessed primarily during the data transformations at the pre-processing and processing stages of the data. The accuracy of a classification model is heavily dependent on the DQ, so measuring DQ [6] is critical for estimating task complexity earlier. DQ attributes should be verified, improved, and regulated throughout their life cycle, as they have a direct impact on the conclusions drawn from data analysis. To capture the quality requirements, characteristics, dimensions, scores, and applications of quality rules, data profiling [7,8] has become a popular approach. DQ assurance and this approach have become so intertwined that they are often referred to as the same. It is a collection of techniques used to facilitate a variety of data management tasks, such as data quality evaluation and metadata management. In healthcare data, on the other hand, ensuring DQ is a time- and resource-intensive process, especially when dealing with large amounts of data. The FL method may be crucial, as it allows for the construction of a common quality model based on multiple sources without sharing data. This method facilitates the need to combine multiple data sources to ensure the quality of data analytics while also protecting the privacy of each individual and reducing the transfer time for data. Our proposal to use a federated data quality profiling model to ensure the privacy and

security of eHealth data, and is motivated by the need to address DQ concerns at every stage of the data's lifecycle, primarily at the edge.

### 1.1. Background

Any work on DQ is incomplete unless the DQ measures and metrics are stated, as they are critical components in measuring data quality. To familiarize the reader with the concept, some PSN background information is also provided.

#### 1.1.1. Data Quality Dimensions and Metrics

For a given situation, some data may be more important than others in terms of achieving the strategic vision, and therefore, when it comes to data quality (DQ), it is important to focus on the most significant data. An entirely new set of metrics incorporating "data weights" has been proposed by the authors in [9]. Choosing a set of dimensions to work with is an important part of the approach, and to measure each dimension, a metric must be selected. To accomplish continuous improvement, TDQM (total data quality methodology) [10] is one of the few approaches that operate in a cyclical or revolutionary fashion. When it comes to metrics, the only thing it uses is "basic percentages", such as the percentage of missing data for the "completeness" dimension. Data quality assessment (DQA) combines subjective and objective evaluations. The root causes of discrepancies can give great insight into the DQ problems and guide data quality improvement efforts. According to the authors of [11], the five requirements of data quality metrics are the existence of minimum and maximum metric values (R1), interval scaling of the metric values (R2), and quality of the configuration parameters, as well as the determination of metric values (R3), sound aggregated metric values (R4), and economic efficiency of the metric values (R5). The term dimension is used to describe elements of data that can be measured, as well as the ways in which data quality can be evaluated and quantified. Data accuracy, completeness, uniqueness, timeliness, and validity are the six primary criteria that determine the quality of data [12]. The following are the accepted definitions of each of these metrics.

**Accuracy:** The degree to which data correctly portray the thing or event under investigation in the "actual world". Validity is a quality dimension linked to accuracy.

**Completeness:** The ratio of data saved to the potential for "completeness". Validity and accuracy are two quality dimensions related to completeness.

**Uniqueness:** No object, regardless of how it is identified, will be recorded more than once. Consistency is the quality dimension connected to uniqueness.

**Timeliness:** The degree to which data correctly represent reality at a given point in time. The quality component associated with timeliness is accuracy.

**Validity:** Data are accepted if they adhere to the syntax of the specification (format, type, and range). The validity-related quality dimensions include accuracy, completeness, consistency, and uniqueness.

It is also important to note that the related literature revealed that most DQ metrics are strongly associated with one another. Data quality characteristics appear in the data collection process as well as the preprocessing stage—this includes both the upstream and downstream stages of the data processing [13]. The upstream influencing factors are determined by the data collection system. The loss of data quality is expressed by missing values in the event of data storage failures or the inability to measure the requested physical values. The completeness indicator considers any missing values. Accessibility, mobility, and recovery all fall under this umbrella term. The data analyst cannot use the signal data if it cannot be accessed or if it cannot be transferred to a database or data mining software. In the event of a failure or loss of data, a lack of recoverability results in a lack of information. When it comes to traceability, the impact is not caused by missing values in the time series, but rather by a lack of details about the dataset itself. The various subdimensions of completeness also cover this influence. When compared to the factors that have an impact on data quality upstream, the factors that have an impact on downstream DQ that appear

during data preprocessing are accuracy, credibility, consistency, and relevance [13]. It is also worth noting that the conversion of signal data to the international system of units (SI), for example, does not have a negative impact on the quality of the data. Compliance is a problem for data quality if there is a lack of information that prevents conversion to a particular standard. Different definitions for data quality metrics have been presented by many research studies. A few of the metrics are listed in the following sections.

#### 1.1.2. Timeliness Metric

The ratio between currency and volatility determines the timeliness and they must be measured using the same units of time. Time tags provide information about the date the data item was acquired. For example, highly volatile data, such as stock quotes or currency conversion tables, have a very short shelf life. Depending on when the information product is delivered to the customer, an information product's timeliness can vary. The data quality metric for timeliness is defined by Ballou et al. in [14] as follows:  $Timeliness = [(1 - currency/volatility), 0]^s$ . The parameter age of the data value represents the elapsed time between the real-world event's occurrence (i.e., the time the data value was created in the real world) and the determination of the data value's timeliness. The maximum amount of time that the values of the considered attribute will remain current is defined as the parameter shelf life or volatility. In other words, a higher value of the metric for timeliness implies a higher value of the parameter volatility and vice versa. The metric's sensitivity to the ratio age of the data value depends on the exponent  $s > 0$ , which must be determined based on expert estimations.

#### 1.1.3. Completeness Metric

Complete data have been described as data with all values recorded. Missing data can typically be indicated by null or another indicator in most applications. The metric for completeness is defined in [15] as:  $Completeness = 1 - (MT/NK)$ , where  $MT$  is the proportion of tuples in relation with null values to the total number of tuples and  $NK$  is the total number of tuples.

#### 1.1.4. Correctness Metric

Heinrichs [11] defined correctness as a metric to evaluate the accuracy of a stored data value:  $Correctness = 1/(d(w, w_m) + 1)$ , where  $w$  is the stored data value,  $w_m$  is the corresponding real-world value, and  $d$  is a domain-specific distance measure.

#### 1.1.5. Patient Similarity Network (PSN)

PSNs are designed to assess whether a patient is likely to benefit from treatment modalities and lifestyle changes of other patients who are likely to be similar to the current patient. The objective of PSNs is to recommend the appropriate therapy and medicine for the patient based on aggregated data extracted from other patients with similar characteristics. A few of the PSN challenges are listed as follows. Clinical narrative data that are diverse and heterogeneous enrich hidden information that is useful in selecting the most comparable patients. Medical occurrences are time-sensitive, and understanding the dynamics of medical terminology and conclusions requires temporal information. When using noisy clinical datasets, interpreting temporal representation is highly challenging, and the result prediction accuracy is low. The dimensionality of health datasets is varied and high. For example, the electronic health record (EHR) stores a wide range of data, such as diagnoses, drugs, laboratory tests, and X-rays, as well as medical events like diseases and treatments. Because the data are a mix of static and dynamic, modeling and processing are difficult.

### 1.2. Motivation

The main motivation of this paper is to address a few of the challenges with respect to data quality. The heterogeneity of eHealth data from diverse data sources may be addressed using the generalized hybrid PSN model proposed in our paper [16]. The model is effective

in solving big data challenges when patient cases contain both structured and unstructured data by employing an autoencoder to enforce dimensionality reduction in the model. The patient similarity network fusion strategy uses PSN distance estimations from static and dynamic data to emphasize patient pair similarity and reduce interference produced by non-similar pairings. However, this PSN fusion strategy was designed with processing at the centralized server, and not considering the quality of data received from the edge nodes or the source.

Through experiments, we assessed the influence of individual edge data quality on FL model accuracy, which motivates us to investigate data quality-aware edge selection and profiling for PSN, integrating it with FL services to address faulty data issues at dispersed client sources. FL training may use a lot of computer resources when there are a lot of training datasets and jobs, and therefore, we propose a model to efficiently improve data quality in remote learning clients. We create a profile based on the data context that can be dynamically sent to all FL clients, and we execute client data selection and augmentation to significantly reduce patient data transmission. When compared to cutting-edge data quality enhancement approaches, our proposed model can significantly improve FL performance for a wide range of learning tasks and FL scenarios. In this paper, we propose the DQA FDQ profiling model, a quality-driven, edge-based federated strategy for sensor-based monitoring setup that is motivated by the following:

- Capture quality at the beginning of the data acquisition process to ensure that DQ is maintained throughout the data lifecycle.
- In the event that quality assessment criteria are dynamically updated in the case of real-time data, it is recommended to introduce a data quality profile, abbreviated as DQP, to support quality assessment at each edge node.
- The federated DQP can provide a more robust and detailed quality evaluation because it will be able to capture the vast majority of quality issues occurring across all nodes. Adopting a strategy to eliminate edges with noisy data and facilitate client selection will reduce the impact of low-quality data on model training. Thus, the federated quality profile will exclude edges with greater quality profile variation.
- Federated profiling will have a low overhead because it will focus only on the quality profile measures and variance and not on the entire datasets stored at the various nodes.

This article aims to address the challenges of data quality in healthcare by proposing a novel federated data quality profiling (FDQP) approach. The proposed FDQP model in federated PSN evaluates the quality of patient data obtained from edge nodes and enhances the accuracy of machine learning models through profiling algorithm and federated feature selection. Experimental results demonstrate the effectiveness of federated profiling in improving data quality and accuracy. The contributions of this paper are described in the following section.

#### Contributions

- Pioneering the federated data quality profiling (FDQP) technique for evaluating patient data quality: This paper introduces a novel approach, referred to as federated data quality profiling (FDQP), which aims to evaluate the quality of patient data obtained from edge nodes. By pioneering this technique, the paper addresses the need for assessing the reliability and accuracy of decentralized healthcare data.
- Development of an FDQP formal model to capture quality dimensions: To encapsulate the various dimensions of data quality profile (DQP), the paper establishes an FDQP formal model. This model serves as a comprehensive framework for representing and analyzing the different aspects of data quality, contributing to the overall understanding and evaluation of the patient data's quality characteristics.
- Utilization of federated feature selection for enhanced precision: By leveraging federated feature selection techniques, this research enhances the precision of classifiers used in analyzing patient data. The paper categorizes features based on metrics

such as feature value, percentage of outliers, and missing data percentage, thereby improving the accuracy and reliability of the classification process.

- Extensive experimental evaluation of the FDQP model: The proposed FDQP model undergoes an extensive series of experiments, utilizing a distributed fetal dataset across diverse edge nodes and varying scenarios. This rigorous evaluation enables the assessment of the effectiveness and performance of the FDQP model in practical healthcare settings.
- Improved data quality and accuracy in FPSN-based machine learning models: The study demonstrates a noticeable enhancement in data quality and accuracy of federated patient similarity network (FPSN)-based machine learning models as a result of adopting the FDQP approach. By incorporating the FDQP model, the paper shows how the proposed methodology positively impacts the performance of FPSN-based models in healthcare data analysis tasks.
- Proposal of a data-quality-conscious federated PSN architecture: This paper presents a novel architecture, termed the data-quality cognizant federated PSN architecture, which integrates the FDQP model to effectively improve the data quality and accuracy of FPSN-based machine learning models. This proposed architecture addresses the challenge of utilizing data collected from edge nodes, while ensuring high-quality and reliable healthcare analytics.
- Application of an efficient profiling algorithm for data quality optimization: The research applies a profiling algorithm that prioritizes efficient lightweight profile exchange over complete data processing at the edge. By adopting this approach, the paper advocates for optimized achievement in data quality, allowing for streamlined data processing and improved efficiency in healthcare data analysis workflows.

The FDQP approach has the potential to be applied in a range of scenarios beyond patient monitoring. Here are a few scenarios that highlight the broader applicability of FDQP:

**Industrial automation:** In industrial settings, where large amounts of sensor data are collected from various machines and equipment, FDQP can be used to assess the quality of data to ensure accurate decision making and optimize production processes.

**Environmental monitoring:** FDQP can play a crucial role in evaluating the quality of environmental sensor data, such as air quality measurements, water quality parameters, and climate data. This can aid in monitoring and addressing environmental issues effectively.

**Smart cities:** With the increasing adoption of IoT technologies in smart city applications, FDQP can be employed to evaluate the quality of data collected from various sensors deployed throughout the city. This can support better urban planning, resource management, and citizen services.

The versatility of the FDQP approach can be extended to diverse IoT applications, enabling data accuracy and integrity across various domains for reliable decision making, improved operational efficiency, and enhanced outcomes.

The paper is organized as follows. In Section 2, we review prior research on DQ and FL. Section 3 describes our FDQ profiling model and proposed algorithm, while Section 4 describes the evaluation of our model using the fetal health monitoring dataset and includes the evaluation methodology and experiments to illustrate the benefits of our approach. Section 5 discusses the experiment findings, underlying principles, and limitations, to present a balanced and realistic view of the proposed approach. Finally, Section 6 concludes the paper and points to some future research directions.

## 2. Related Work

Many challenges must be resolved when managing a federated set of health data, including individual user rights and needs, maintaining user data privacy, establishing rules for creating and changing accounts and roles, providing tools for sharing private data with selected audiences, and allocating shared costs among participants in a collaborative effort, including data sharing, transfer, and integration. Clinical information can be protected

by using FL [17], as data transfer is not required and all necessary information can be stored locally in a patient's healthcare organization or the patient data edges. In the data analytics lifecycle, however, erroneous data discovered at one stage may have originated from the processing performed upstream. It is essential to ensure the integrity of each link in the lifecycle in order to determine the severity of a problem, as the ability to track the propagation of defects through a system can aid in determining its magnitude. Eliminating false positives, or results that do not pertain to a particular problem is another benefit of DQ profiling [7]. Consequently, managing DQ profiling through federation enhances DQ evaluation while maintaining data confidentiality. The following is a discussion of the current state of quality management over distributed and federated sensory data.

Some data quality objectives can be achieved through database integration using DQ attributes, such as querying the source database for inconsistencies and reconciling inconsistencies during merging using federated database approaches [18]. In global applications, the metadata models were used to identify DQ-related query targets and to dynamically combine data from component databases. The number of complex computations that can be accomplished on mobile devices has recently risen, supporting the spread of the "on-demand-edge-computing" paradigm, which is characterized by the concept of FL frameworks. The authors of [19] propose a DQ-centric big data architecture for federated sensor service clouds, in which high-quality data from a large number of separately managed sensors is exchanged or even traded in real time. One of FL's challenges is training clients with poor or severely biased data, which may jeopardize FL's efficiency and efficacy, resulting in a suboptimal global model that requires more rounds of computation to converge. Based on the FL technique, a data representation profiling method called FEDPROF [20] was proposed, which adaptively modifies clients' participation likelihood based on their profile dissimilarity while maintaining data locality.

By utilizing FL based on source data instead of centralized data, mobile edge computing becomes more intelligent. The additional computation may be used in FL to minimize the number of communication rounds necessary to train the model. For example, the in-edge AI framework [21] relies on edge caching and computation offloading to boost computation and communication per edge device as an effective technique to handle the various challenges related to energy consumption, privacy, and modeling fairness at the edge. All stakeholders in the underlying data processing edges can have fair access to high-quality data when using FL methods. When an attack is detected, FairFed [22], a new FL framework, rejects model updates that lead to an attack.

Although there are studies that highlight FL, DQ, and PSN separately, none of these initiatives explored DQ-aware federated PSN to assess the quality of the data itself and its impact on increasing the prediction accuracy of PSN models. In the following section, we highlight the related concepts within the context of FL based on the literature review.

### 2.1. Data Quality Assessment

Many approaches for detecting noisy data have been proposed in the literature, with the most common three being distance-based, density-based, and clustering-based strategies [18,23]. In its most basic form, an outlier is a value that differs from the rest of the data in a collection. Outlier detection assigns each object an outlier score that indicates how much it deviates from the norm. Such algorithms can remove any specified amount of noise, for example, by sorting objects based on their "outlier score" and deleting the items with the highest outlier scores until the appropriate percentage of objects is removed. Data profiling [24] can solve DQ concerns, such as noise and outliers, inconsistent data, duplicate data, and missing values [23,25,26].

Due to the presence of minor class representation in the data, class imbalance problems arise, and when the data are classified, the classifiers are biased toward the dominant class. For instance, in medical diagnosis, the detection of cancerous cells and the misclassification of non-cancerous cells may necessitate additional clinical testing. However, if the misclassification of cancerous cells leads to the prescription of no additional tests, the patient's

cancer may go undetected, which poses a significant health risk. In the medical diagnosis example, however, the classifier might very well mislabel some benign tumors as malignant (cancerous), thereby subjecting a patient without cancer to unnecessary tests. Similarly, there are many instances where the overall classification accuracy of ML algorithms misclassifies minority classes. This leads to increased misclassification costs, time, and risk assessment. Each client has a unique quantity of locally stored training data. Therefore, the reliability of the trained values will vary depending on the client, as there is a possibility that the training sessions will be too brief [27].

In today's information age, where vast amounts of data are being generated at rapid intervals, it can be difficult to sift through this flood of information to find the bits of value. There is a centralization issue, because large tech companies control the majority of the data and resources required to effectively train machine learning models. Traditional machine learning settings explore the subject of relevant data selection under the presumption that all relevant data are available in one location for computation. Due to the fact that the data in FL settings are dispersed among numerous clients and the server is unable to inspect it because of privacy restrictions [28], the conventional methods for choosing pertinent data are not applicable in this situation.

An approach that finds relevant data and stores it in a decentralized way with an innovative approach to network sharding that the authors call the interest group [29]. The data added to the ledger are verified for accuracy by the proof of common interest. In [28], it is suggested to use a method called federated learning with relevant data (FLRD), which allows clients to select data relevant to the server's objective, leading to improved performance of the global learned model. Each client must learn a function that estimates data point relevance without compromising privacy.

Some of the challenges in federated data quality profiling are inspired by the FL environment, which involves data that are distributed and from several sources, which presents its own set of issues, as opposed to a centralized dataset that is simply dispersed around worker nodes.

## 2.2. Quality-Aware Client Selection in FL

A strategy for user recruitment called quality-aware user recruitment [30] has been addressed to handle the problem of optimizing the observed DQ in mobile crowd sourcing (MCS) mode. Through federated learning, this work predicts the quality of sensed data from various users by examining the correlation between data and context information.

The majority of current FL approaches focus on the supervised setting, where each client's data are labeled. FedTriNet [31] is proposed as a federated semi-supervised learning method with a pseudo-labeling method having two learning phases, because it is impossible to completely label client data in real-world applications.

One of the most prized qualities of a data valuation metric is its fairness. Data owners are key to federated learning's success. Each data owner's contribution to the model's performance should be reflected in the FL setting's data valuation metric. Furthermore, it is important to fairly evaluate data owners' quality and contribution to the final model and reward accordingly if you want them to continue contributing. The federated shapley value [32] satisfies many data valuation properties. Two data owners with the same local data may not receive the same evaluation with a federated Shapley value. Completed federated Shapley value [33], based on a low-rank matrix completion formulation, improves the federated Shapley value's fairness. Data source selection is required in practical applications due to the high cost and low availability of communication channels, and FDSS [34], an algorithm, is devised, which can effectively address this issue in both static and dynamic contexts and demonstrate that it can be transformed into a monotone submodular maximization problem.

An adaptive accuracy threshold aggregation strategy based on federated learning [35] is proposed, which can satisfy the practical needs of multi-party data learning without requiring the sharing of sensitive data. The local node obtains information from

the local dataset. After local training, the aggregation node verifies that the conditions for this round's aggregation have been met. In this instance, local model data are retrieved from the local node, and model aggregation is carried out. The algorithm calculates the optimal threshold to reduce the number of communications between local nodes and aggregation nodes. This algorithm can alter the accuracy of each round during model training and determine the optimal threshold. Since the local model data indirectly reflects the node sample information, an attacker can deduce the sample data from the effective model information, reducing the number of communications and the probability of privacy leakage. Hybrid logical security framework (HLSF) [36] proposes strong authentication and data confidentiality mechanisms, which improves security and network capabilities, promotes green IoT, and addresses privacy concerns.

AUCTION [37] is intended to embed the client selection policy into a neural network and then use reinforcement learning to automatically learn client selection rules based on observed client status and feedback incentives quantified by federated learning performance. The policy network is based on an encoder–decoder deep neural network with an attention mechanism that can adjust to dynamic changes in the number of prospective clients. TiFL [38], a tier-based FL system provides an adaptive client selection approach that uses quality as an indirect indicator to infer data heterogeneity information and updates the tiering algorithm on the fly. This minimizes training time while increasing accuracy. They also employed a lightweight profiler to evaluate each client's training time and put them into logical data pools known as tiers depending on measured latency.

### 2.3. Data Heterogeneity on FL

Since data in the FL paradigm cannot be gathered or shared, data heterogeneity inevitably arises, severely impacting the efficacy of federated analytics. The analytics-driven client selection (FedACS) [39] framework proposed a three-pronged approach to dealing with the issue of data heterogeneity. Without revealing any private information, clients first generate insights about their local data; then, the server uses these insights to infer the situation of clients' data using Hoeffding's inequality. Finally, a client pool is formed by selecting individuals from a slightly more diverse set of data. Each individual customer has insufficient data because of costly labeling and the inability to store additional information. Large amounts of unlabeled data are readily available in the public cloud (e.g., social media). To boost DL model efficacy, Ada-FedSemi [40] uses both local and cloud-based labeled and unlabeled data. Due to the high volume and low availability of the clients, a carefully chosen subset of clients is used in FL during each iteration of training. The parameter server compiles these local models into aggregated pseudo-labels for unlabeled data, which is then used to train a better global model. Federated learning has been widely utilized to connect resource-constrained devices for neural network on-the-edge training. When federated learning deploys identical neural network models to heterogeneous edge devices, those with less processing capacity may dramatically delay the synchronized parameter aggregation, resulting in serious computational straggler problems [41]. These computational stragglers are edge devices with little processing capability that are attracting increasing attention from the research community [42]. Although training model improvement can help with stragglers, the optimized models frequently result in divergent structures due to diverse device resource limits, which has a substantial impact on collaborative convergence. FedProx [43], a variant of FedAvg [44], is presented to address system heterogeneity caused by system feature variability on federated training devices. System heterogeneity has a significant influence on model aggregation efficiency and accuracy, causing the optimization to diverge [45].

### 2.4. Implications of Communication in FL

Communication is an important part of FL and approaches to decrease communication, such as local updating and model compression, must be understood through a thorough analysis of the trade-off between accuracy and communication for each strategy. Bulk syn-

chronous and asynchronous techniques are the two most often researched communication strategies in distributed optimization. These strategies are more practical in data center settings because worker nodes are generally dedicated to the task, i.e., they are ready to ‘pull’ their next job from the central node as soon as the results of their previous job are ‘pushed’. In federated networks, on the other hand, each device is typically unfocused on the task at hand, and most devices are inactive at any given time. As a result, it is worthwhile to investigate the impacts of this more realistic device-centric communication architecture, in which each device may select when to ‘wake up’ and connect with the central server in an event-triggered manner [46]. To adapt the FL model to 6G networks [47] with huge, heterogeneous devices and networks, a communication-efficient approach must be developed, considering two major aspects: (a) lowering the total number of communication rounds or (b) reducing the number of gradients in each communication round. Another challenge is that the majority of present FL works use machine learning models with full-precision weights, and virtually all these models contain a high number of redundant parameters that do not need to be transferred to the server, spending an unnecessary amount of communication expenses. In addition, using data from edge devices is not always possible due to the devices’ inability to transmit data quickly enough or due to poor connectivity circumstances [48]. A ternary federated averaging protocol (T-FedAvg) [49] was developed to overcome this issue by reducing the upstream and downstream transmission of FL systems.

### 2.5. Data Model Quality-Aware FL

Another focus of recent research is to raise FL model quality awareness. The model updates contributed by computing nodes training with their local data determine the quality of federated learning. The model update qualities of computing nodes can vary dramatically depending on various factors (e.g., training data size, mislabeled data samples, skewed data distributions). The accumulation of low-quality model updates can degrade the overall model quality. A framework called FAIR [50] integrates three major components to achieve efficient FL learning: learning quality estimation, a reverse auction problem to encourage the participation of high-quality and inexpensive computing nodes, and auto-weighted model aggregation.

Federated edge learning’s privacy-preserving aspect makes it the future of wireless edge network training. In [51], data-quality aspects are used to schedule edges for collaborative training. First, the learning algorithm’s components, dataset diversity, and edge node reputation are defined. The authors proposed a DQ-based scheduling (DQS) algorithm that prioritizes reliable devices with rich and diverse datasets.

### 2.6. FL Data Streaming Concerns

There are several practical difficulties that occur when FL is used with real-life streamed data. Concept drift (underlying data-generation model changes over time), diurnal variations (devices exhibit different behavior at different times of the day or week), and cold start problems (new devices enter the network) are just a few of the issues [46] that must be handled with caution. The Adaptive-FedAVG FL [52] approach works with non-stationary data-generating systems that are impacted by concept drifts by modifying the learning rate to increase the learning phase’s flexibility, allowing it to adapt to concept drift.

Patient privacy and the risks associated with releasing personal health information, such as stigma, continue to be key challenges in healthcare. Because patient data are not exchanged between multiple locations, but the model goes to the data, our recommended DQA-FPSN model is created with privacy in mind. This enables the identification of similar individuals without risking patient privacy, allowing for accurate diagnosis and lifestyle recommendations.

Table 1 provides an overview of notable related work in the paper, highlighting different contexts and challenges addressed in the literature. It presents key features, advantages, and disadvantages of each reference, offering a concise understanding of the discussed work.

**Table 1.** Summary of related work.

Concept	Reference	Key Features	Advantages (+)/Disadvantages (–)
<b>FL for Data Protection</b>	[17]	Uses FL for protecting clinical information, eliminates data transfer, maintains data confidentiality.	+ Preserves data privacy, allows local storage of patient data. – Challenging to train clients with poor or biased data.
<b>DQ Assessment</b>	[7]	Discusses DQ profiling, eliminates false positives, improves DQ evaluation.	+ Enhances DQ evaluation, maintains data confidentiality. – Does not address federated or distributed data specifically.
<b>Federated Sensor Service Clouds</b>	[19]	Proposes DQ-centric architecture for federated sensor service clouds, exchanges high-quality data in real time.	+ Enables real-time data exchange, improves data quality. – Challenges with training clients having poor or biased data.
<b>Quality-aware User Recruitment in MCS</b>	[30]	Introduces quality-aware user recruitment in MCS using FL, predicts the quality of sensed data.	+ Optimizes observed data quality, improves FL performance. – Limited to the MCS context.
<b>FL for In-Edge AI</b>	[21]	Presents an in-edge AI framework using FL, boosts computation and communication per edge device, handles energy consumption, privacy, and fairness challenges.	+ Improves computation and communication efficiency, addresses energy and privacy concerns. – Focuses on edge computing, may have limitations in scalability.
<b>Fairness in FL</b>	[22]	Introduces FairFed, an FL framework that rejects model updates leading to attacks, ensures fairness in FL.	+ Enhances security and fairness in FL, prevents model poisoning attacks. – Does not explicitly address DQA.
<b>Data Heterogeneity in FL</b>	[39]	Proposes FedACS framework to address data heterogeneity in FL, generates insights about local data and forms a diverse client pool.	+ Handles data heterogeneity, improves FL efficacy. – Does not focus on DQA directly.
<b>Communication Strategies in FL</b>	[46]	Discusses communication strategies in FL, investigates the impacts of device-centric communication architecture.	+ Considers realistic communication scenarios, improves efficiency. – Does not specifically address DQ concerns.
<b>Federated Edge Learning</b>	[51]	Uses DQ aspects for scheduling edges in federated edge learning, prioritizes reliable devices with diverse datasets.	+ Enhances edge selection in FL, considers DQ for scheduling. – Limited to federated edge learning context.
<b>Concept Drifts in FL</b>	[52]	Introduces the adaptive-FedAVG FL approach to handle concept drifts in nonstationary data-generating systems.	+ Adapts to concept drifts, improves model flexibility. – Focuses on concept drift, may not cover other data streaming concerns.
<b>Client Selection Policy</b>	[37]	Proposes an auction, client selection policy embedded in a neural network.	+ Reinforcement learning-based client selection. – May require additional computational resources.
<b>Accuracy Threshold Aggregation</b>	[35]	Privacy-preserving, multi-party data learning.	+ Reduces privacy leakage probability. – Requires threshold tuning.

### 2.7. Research Gaps

Table 2 compares the proposed model, DQA FPSN, with state-of-the-art approaches in the field of FL. The listed approaches encompass various aspects, including node/client selection, dimension reduction, data heterogeneity, federated profiling, data quality assessment, FL-based methods, and accuracy Improvement. It is worth noting that the proposed DQA FPSN addresses several research gaps that have not been explored in the existing literature.

Despite the advancements in FL approaches, there is a lack of emphasis on quality-aware client selection, which plays a crucial role in ensuring the reliability and representativeness of the participating clients in the FL process. Another research gap lies in the dimension reduction or feature selection techniques specifically tailored for FL settings. Existing methods often assume homogeneous data, and there is a need to explore dimension reduction approaches that can handle the inherent data heterogeneity across different clients while preserving privacy and data utility. Current approaches focus on aggregating model updates, but there is room for improvement in terms of profiling the participating clients to better understand their characteristics and performance, leading to more informed aggregation strategies, while FL has shown promising results in various applications, ensuring accuracy improvement by the integration of DQ assessment, and DQ dimensions into FL frameworks is another area that requires more attention. There is a need for advanced optimization techniques and personalized learning approaches that can adapt to each client's data distribution and enable more accurate and reliable model updates for heterogeneous data, specific to various domains, such as PSN, in healthcare. Addressing these gaps would significantly enhance the overall effectiveness of FL in real-world scenarios.

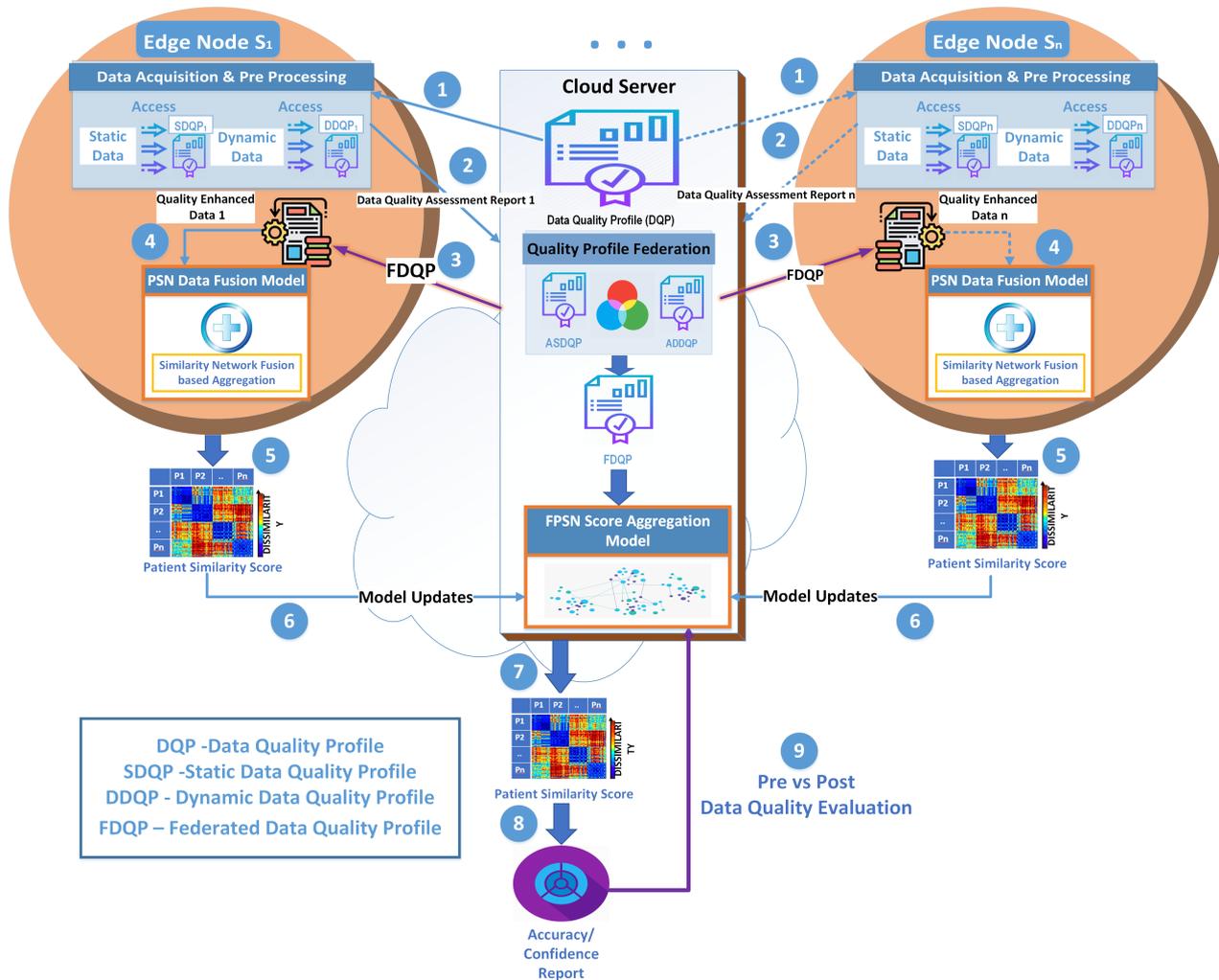
**Table 2.** Comparison of the proposed model against the state-of-the-art approaches.

Reference	Approach	Node/ Client Selection	Dimension Reduction/ Feature Selection	Data Heterogeneity	Federated Profiling/ Aggregated Models	DQ Aware	FL Based	Accuracy Improvement
[30]	Quality-aware user selection	✓	-	-	-	✓	✓	-
[31]	FedTriNet	-	-	-	-	-	✓	-
[32]	Federated Shapley value	-	✓	-	-	-	✓	-
[34]	FDSS	✓	-	-	-	-	✓	-
[37]	AUCTION	✓	-	-	-	✓	✓	-
[50]	FAIR	-	-	-	✓	✓	✓	-
[38]	TiFL	✓	-	✓	-	-	✓	-
[39]	FedACS	✓	-	✓	-	-	✓	-
[40]	Ada-FedSemi	-	-	-	✓	-	✓	✓
[DQA-FDQP]	Proposed FDQP	✓	✓	✓	✓	✓	✓	✓

### 3. Data-Quality-Aware FPSN Model

In this section, we present our proposed data-quality-aware federated PSN (DQA FPSN) model, how to apply FDQP with an illustration, the detailed algorithm, and finally the model formulation. Figure 1 details our proposed DQA PSN model architecture that features federated quality profiling, where the data sources are at the edge. The architecture also features a cloud-based server that facilitates the profiling federation and the federated PSN score aggregation. Pre- and post-data quality evaluation is performed, and the process is repeated until the data quality reaches acceptable tolerance levels. The following are the sequential steps of the model processes.

1. In the initial stage, the centralized cloud server sends the baseline DQP to the edge nodes.
2. Subsequently, at the edge local node, each node verifies and evaluates the data quality acquired from the sensors, updates the DQP, and transmits it back to the server.
3. Finally, the server integrates the DQPs received from the edge nodes to create the federated data quality profile (FDQP), which is transferred to the edge nodes.
4. FDQP is then applied to the local edge data that creates the quality-enhanced data, which will be the basis for the PSN data fusion model at the edge.
5. The resulting patient similarity score is sent back to the cloud server.
6. The FPSN score aggregation model receives the model updates from the edges and the aggregation of the similarity scores takes place at the cloud server.
7. The final patient similarity score is used as a basis to detect the most similar patient.
8. DQ evaluation measures of performance such as accuracy are calculated.
9. The process is further assessed with pre- and post-DQ evaluations.



**Figure 1.** FPSN enhanced by FDQP at the edge.

### 3.1. Data-Quality-Aware FPSN Model Overview

An overview of the proposed model with FPSN and FDQP edge enhancement is shown in Figure 1. The objective of this framework is to enhance the precision and speed of data processing at the edge of the network.

### 3.2. Federated Data Quality Profiling

Figure 2 depicts federated data quality (FDQ) profiling using an example. The baseline quality profile specifies the needed data quality characteristics such as completeness, accuracy, and timeliness, as well as the data quality standards. Based on the baseline quality profile, we identify the dimensions of DQ with severe issues. A missing data-related rule, for example, will infer some actions (e.g., replace missing values with the mean) depending on the kind of data and the degree of tolerance specified in the DQP. This baseline DQP will be forwarded to the source edges, where the local dataset will be reviewed using the profile and a new quality profile will be constructed using some or all of the criteria. If the new profile satisfies the baseline quality profile, it is sent back to the server. It is worth noting that we are considering the edges of collecting and holding identical datasets with similar characteristics. Hence, Figure 2 shows that source edge 1 with attribute ID 1 has a missing value of 70%, and thus, rule 2.2 is used to eliminate the whole column.

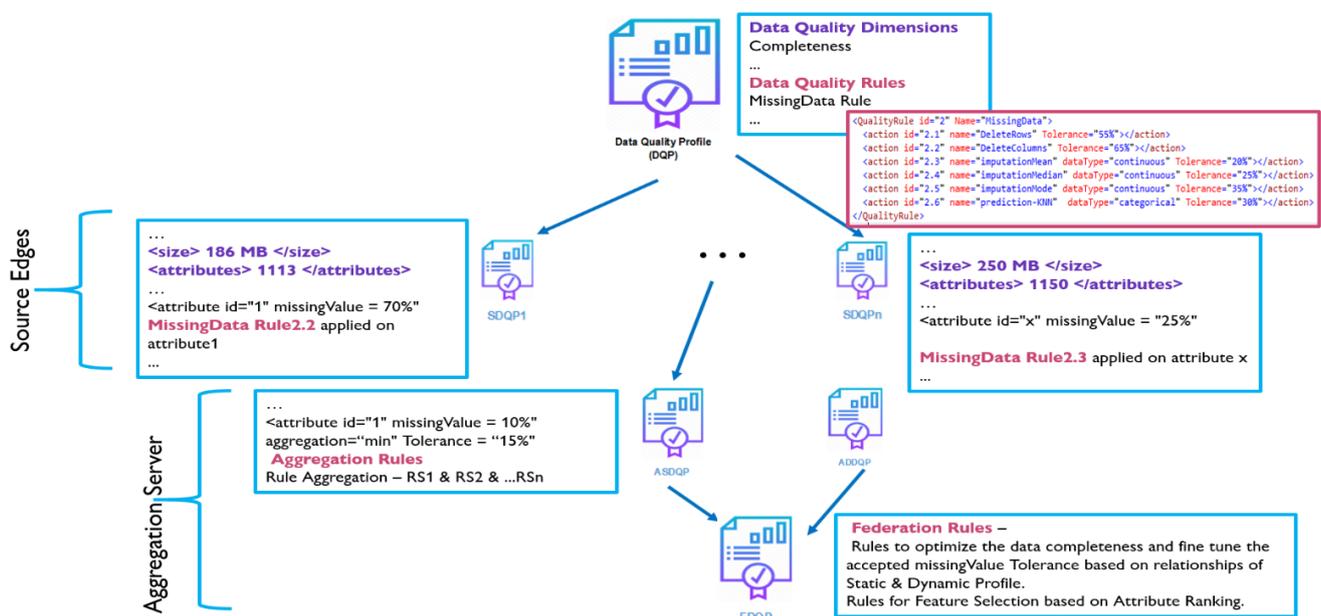


Figure 2. FDQ profiling example.

Similarly, at each edge node, a local profile is relayed to the server node, which aggregates all profiles. According to the context and the rules, the aggregate process will use min, max, total, or average aggregation. The aggregated quality profiles will then be integrated to generate the federated data quality profile, which will include optimal rules based on attribute correlations. Furthermore, the feature selection rules will be defined in the federated data quality profile based on attribute priority and ranking. Thus, the federated data quality profile will have a well-defined purpose, quantifiable metrics, and optimized rules for selection, as well as explicit formulas for combining local/federated data profiles. Finally, the rules in the federated data quality profile are propagated and enforced across all nodes in the federation, ensuring enhanced data quality.

### 3.3. Model Formulation

Let  $A$  represent a collection of data attributes of the dataset expressed by  $A = a_0, \dots, a_j, \dots, a_D$ , where  $D$  is the number of attributes (or the dimensionality of the dataset) and  $a_j$  is an attribute represented by its type, possible values, weight, tolerance, and rules. For each  $a_j$  in  $A$ , the weights of each attribute are represented by a set  $W = w_0, \dots, w_i, \dots, w_p$ , where the sum of all weights should be 1. The weights are defined using a kernel function that provides the priority and importance of each attribute in the set  $A$ , thus enabling feature selection. Each attribute evaluation is mapped to a set of

data quality dimensions  $D = d_0, \dots, d_k, \dots, d_q$ .  $T$  is a collection of minimum acceptable tolerance levels,  $T = t_0, \dots, t_l, \dots, t_r$ , established by data specialists and associated with every quality dimension (e.g., completeness accepted tolerance is 70 percent, which means the attributes need to comply with above 70 percent completeness). There are zero, one, or more applicable rules,  $R = r_0, \dots, r_m, \dots, r_s$ , to the attributes, which are applied when meeting criteria based on acceptable tolerance levels specified for the attributes. Rules can be tailored to handle quality issues and include corrective actions, such as data imputation or outlier elimination.

Our algorithm's main phases are baseline representation profiling, edge representation profiling, and federated profiling. A DQP is, thus, represented by  $(A, W, D, T, R)$ . First, the baseline DQP, DQPv is generated based on the sample dataset available at the server, which is similar to and characterized by independent and identically distributed (IID) datasets available at the edges. Quality requirements or preferences can be set for all attributes (the entire dataset) by default, as well as applied to individual attributes, on demand. The settings are determined by the data quality dimensions selected for profiling and by the requirements of the application. Certain dimensions, such as completeness, can be set for all attributes by specifying the expected ratio or tolerance that must be met. Several other dimensions, such as timeliness, are more focused on specific attributes such as time or date.

The DQPv is subsequently sent to all the edges where it is applied at the edge data sources to form DQPe with the tolerance values, max and min (possible values), and the quality measures (metrics) of the dimensions specified by  $M = m_0, \dots, m_n, \dots, m_t$ .

### 3.3.1. Quality Rules Development

DQPe is, thus, represented by  $(A, W, D, T, R, M)$  and is built at the edges using the baseline profile as a blueprint. The dataset is analyzed, and the profile is updated to include data size, attribute count, row count, and attribute details with the maximum, minimum, and percentage of missing values. Quantitative measures of missing data, unique data, and completeness relative to the dataset are added to the quality profile. To guarantee an increase in edge data quality, certain steps must be taken as outlined in the rules. Here is an example of the XML code that describes how to handle missing data in the dataset.

The quality rules of missing data are specified in the XML, as depicted in Figure 3. With quality rule action 2.1, using mean imputation, we replace missing values of attribute  $X$  with the attribute's mean, which is calculated using the attribute's non-missing values. In a normally distributed variable, the median, mean, and mode will all be close to one another. Thus, using the mean or the median to fill in missing data is equivalent, but for skewed attributes having missing value tolerance under 5%, it is recommended to do imputation with the median [53]. Using the mode to fill in gaps in numerical variables is unusual.

```
<QualityRule id="2" Name="MissingData">
  <action id="2.1" name="imputationMean" dataType="continuous" pattern="numerical" Tolerance="5%"></action>
  <action id="2.2" name="imputationMedian" dataType="continuous" pattern="skewed" Tolerance="5%"></action>
  <action id="2.3" name="ImputationMode" dataType="continuous" pattern="stringOrcategory" Tolerance="5%"></action>
  <action id="2.4" name="prediction-KNN" dataType="categorical" pattern="MNAR" Tolerance="20%"></action>
  <action id="2.5" name="MultipleImputation" dataType="continuous" pattern="numerical" Tolerance="20%"></action>
  <action id="2.6" name="DeleteRows" Tolerance="25%"></action>
  <action id="2.7" name="DeleteColumns" Tolerance="25%"></action>
</QualityRule>
```

**Figure 3.** XML quality rule for missing data in baseline DQP.

When there is a connection between the missing data and the other features of the record, or when there is background knowledge about the likely data values, it is possible to draw conclusions about the missing data and the other features of the record. To fix data issues, supervised machine learning algorithms can be used to predict values and correct missing values. MNAR (missing not at random) means that something is missing in a systematic way rather than by chance. In this case, there is a systematic dissimilarity between the available data and the missing data which can be handled by KNN (K-nearest neighbor). Feedback iterations can aid algorithms in learning and increasing their precision

over time. This quality rule is incorporated in rule 2.4. There is uncertainty in the imputed values, so multiple imputations (MI) were proposed in the literature as a method to fill in the gaps. In MI, the required number of imputations is proportional to the frequency of missing data, i.e., more imputations are needed for a dataset that is missing a lot of information [54]. We have incorporated MI in rule 2.6 when the tolerance is above 20% and the data is numerical. As imputation could introduce some bias into the data, it is recommended to remove the row or columns if the missing tolerance is larger, as indicated in rule action id 2.6 and 2.7. Similarly, other DQ rules are specified in XML and sent back to the server, where it is federated.

### 3.3.2. FDQP Formulation

The FDQP formulation captures and aggregates data quality profiles from multiple sources in federated systems, enabling a comprehensive understanding of data quality at the federated level.

$$[DQP]_{Fed} = [DQ]_{FedProf}(Grp \sum([DQP]_e(A, W, D, T, R, M))) \quad (1)$$

where  $DQP_{Fed}$  represents the data quality profile at the federated level, which is the overall data quality representation obtained through the federated profiling process,  $[DQ]_{FedProf}$  (data-quality aware) is the group aggregation of the DQP received from all the edges; the aggregation is based on the dataset and the data quality dimensions specified in the DQP,  $Grp \sum$  represents a grouping and aggregation operation applied to the individual data quality profiles ( $DQP_e$ ) obtained from the edge data sources, and  $DQP_e$  represents the data quality profile at the edge level, which is obtained by applying the DQPv (baseline data quality profile) to the edge data sources.

Parameters:  $A$  (data attributes),  $W$  (attribute weights),  $D$  (data quality dimensions),  $T$  (tolerance levels),  $R$  (applicable rules), and  $M$  (quality measures).

Figure 3 gives a comprehensive overview of the FDQP-mandated federation rules. Both global and attribute-based summaries of quality profiles are generated. Attributes and the “MissingData” deletion criteria are both standardized at the federation level and then cascaded to the edges. Additionally, there is another rule that uses federation heuristics to fine-tune the tolerance applied to the attributes.

$[DQ]_{FedProf}$  also considers the dimensionality reduction with feature selection where the number of attributes is reduced based on missing data and acceptable tolerance rules. Assume we have a dataset in  $D$ -dimensional space with  $n$  samples. Dimensionality reduction approaches convert a dataset with dimensionality  $D$  into a new dataset with dimensionality  $d$ , while keeping as much of the data’s shape as possible. For each edge, and each attribute  $a_j$ , a missing value  $MVi$  information vector is created, where  $ATol$  is the acceptable tolerance for the attribute.

$$[MV]_i = a_j.isNull \cdot count() / D \quad (2)$$

$$[[MV]_i a]_{i,j}^n = \begin{cases} 1, & \text{if } [MV]_i a_{i,j} < [ATol(a_{i,j})] \\ 0, & \text{Otherwise} \end{cases} \quad (3)$$

Furthermore, the vector is aggregated during  $[DQ]_{FedProf}$ , considering all the edges where rules are applied based on attribute weight and tolerance. For instance, If the attribute weight is above 0.5, the aggregation rule adds a chosen imputation method; however, if the calculated  $MVi$  is 0 (meaning it is above the missing value tolerance) and the attribute weight is insignificant, rules are prescribed to remove the attribute. As a result, these rules override the profile and reduce the dimension to  $d$  at the FedProf aggregation. This federated DQP,  $[DQP]_{Fed}$ , is distributed to the edges, where the rules are applied. All of the data measurements and rules acquired in the preceding profiling’s are considered in the FedProf aggregation.  $[DQP]_{Fed}$  will differ from the  $DQP_e$  in many aspects, including values, size, dimensions, and data metric scores. Federated  $[DQP]_{Fed}$  will be used to

determine the right quality metric functions to assess a data quality dimension  $dk$  for an attribute  $ai$  with a weight  $wj$ . Edges that fail to meet critical data quality metrics will be dropped from further processing. Finally, the  $[DQP]_{Fed}$  is sent to the edge nodes and applied for PSN calculation and evaluation based on different PSN fusion algorithms.

### 3.4. Model Profiling Algorithm

We have detailed the FDQ profiling algorithm in Algorithm 1. This algorithm takes the server and the client processes into account. The number of edges, list of dimensions, list of rules, and version number of the quality profile are the input parameters. The FDQP and quality-enhanced data are the outcomes. Initially, a baseline DQ profile with the quality dimensions and rules list is created and distributed to all edges. Each client creates their local DQP, which is then sent to the profile federation server, where the FDQP is built in accordance with the federation rules. FDQP is further forwarded to clients, where it is applied at the edges depending on DQ tolerances that have been stated in advance. The client is disconnected if the data quality at the edge is not adequate. When applied, the FDQP results in data that is both enriched in quality and useful for further eHealth analytics. For the duration of the data streaming process, iterations are performed, and profile versions change incrementally. Once the profile quality reaches a threshold, the profiling stops and the profile is used in all subsequent iterations. To minimize overhead, only the profile updates are transmitted following the initial run. For continuous real-time profiling, once the edge data are applied with FDQP, the data are marked to be forgotten in the profile, so they will not be used again, and the process is then repeated.

### 3.5. Federated Feature Selection

The FDQP relies heavily on federated feature selection to boost classifier precision. The task of feature ranking [55] entails determining the relative importance of a set of features and then ordering them accordingly. This relative importance is typically determined by a feature selection criterion. Given that the data in our scenario are dispersed across multiple nodes, the challenge is to minimize information loss during federation. The number of available nodes with high-quality data or the varying number of features per node could be the limiting factor, depending on the situation we are tackling. The pseudocode for federated feature selection is provided in Algorithm 2; it ranks features according to criteria including feature value, outlier percentage, and missing data percentage, and these metrics are collectively indicators of data quality.

Given a network with 10 nodes and 100 features per node, we can estimate the difficulty of the ranking combination task because the rankings are not identical.

The federated feature selection is illustrated in Table 3. Column 1 contains the list of features, while columns 2 through 4 contain the aggregated feature selection rank, feature outlier rank, and feature missing rank. The ranking criteria for each column are specified in the header; for example, the feature with the lowest outlier percentage is ranked first (i.e., Rank 1), and the feature with the highest outlier percentage is ranked N, where N is the number of features. The final column, federated feature rank, is calculated as described in Algorithm 2. In the provided features in Table 3 (A, B, C, D, E), D is the best valuable feature (with the least federated rank), while E is the worst one. Priority is given for feature selection if two features have the same federated value, followed by outlier and missing rank. Thus, missing data and outliers have a negative effect on aggregation for making the right decision regarding data selection and feature extraction, as illustrated by the federated feature selection. The federation of feature selection adds value because, in a real-world scenario, each node has only partial information to rank the features as it does not have the entire dataset, making it impossible to accurately compute the importance given to each feature.

**Algorithm 1** Federated Data Quality Profiling Algorithm

---

**Input:**

$S_n$ , ▷ No of Edges  
 $AList$ , ▷ List of Attributes  
 $DList$ , ▷ List of Dimensions  
 $RList$ , ▷ List of Rules  
 $Q_{Tol}$ , ▷ Acceptable Quality Tolerance  
 $v = 0$ , ▷ Version of the DQ Profile(DQP)

**Output:**

$DQP_{Fed}$  ▷ Federated DQ Profile  
 $PScore_{Fed}$  ▷ Federated Patient Similarity Score

**//Baseline Profiling**

1:  $DQP_v \leftarrow initializeProfile(DList, AList, RList)$  ▷ Data Quality Profile - Generate baseline profile with quality dimensions and rules.

**//Edge Profiling**

2:  $WLP_e \leftarrow EdgeWorkloadProfileCreate(Config_e, RealTime_e)$  ▷ Create Workload Profile based on config file and edge resource real time parameters.

3:  $DQP_e \leftarrow ClientProfileCreate(DQP_v) + WLP_e$  ▷ each Edge i

**//Federated Profiling**

4:  $DQP_{Fed} \leftarrow DQ_{FedProf}(\sum_{e=1}^n DQP_e)$  ▷ Federated DQ Profile Aggregation.

**//Edge Processing**

5:  $DQEnrichedData_e \leftarrow ClientProfileUpdate(DQP_{Fed})$  ▷ each Edge i  
6:  $PScore_{edge} \leftarrow PSNFusionModelpssing(DQEnrichedData)$  ▷ Quality Enriched Data is passed to the Edge PSN Model to determine the patient similarity score.

**//Federated PSN and Centralised Processing**

7:  $PScore_{Fed} \leftarrow FederatedPSNModelpssing(\sum_{e=1}^n PScore_{edge})$   
8: **while** ( $Accuracy(PScore_{Fed}) < Q_{Tol}$ ) ▷ Quality Score Evaluation  
9:  $v \leftarrow v + 1$  ▷ Increments Profile Version  
10:  $DQP_v \leftarrow DQP_{Fed}$  ▷ Baseline Profile is updated with the Federated Profile.  
11: **Repeat** ▷ The process is repeated until the target accuracy is obtained.

**//Client process: running on the clients**

12: **procedure** CLIENTPROFILECREATE( $DQP_v$ )  
13:      $DQP = GenerateEdgeProfile(DQP_v)$  ▷ DQ Profile  
14:     **return**( $DQP$ )  
15: **end procedure**

16: **procedure** CLIENTPROFILEUPDATE( $DQP_{Fed}$ )  
17:      $DQ_{tolerance} \leftarrow extractDQLimits(DQP_{Fed})$   
18:     **for**  $DList_i \leftarrow 1, DList_n$  **do** ▷ each quality dimension  
19:         **if**  $DList_i \leq DQ_{tolerance}[DList_i]$  **then**  $disconnectClient()$   
20:             **return** 0  
21:         **end if**  
22:     **end for**  
23:      $LFDQP \leftarrow UpdateLocalClientProfile(DQP_{Fed})$  ▷ Local Federated DQ Profile  
24:      $ClientData_{DQEnriched} \leftarrow ApplyClientDataProfiling(LFDQP)$   
25:     **return**  $ClientData_{DQEnriched}$   
26: **end procedure**

---

**Algorithm 2** Federated Feature Selection Algorithm

**Input:**

$S_n$ , ▷ Participating Edge Source Nodes  
 $DQP_{Features}$ , ▷ Node Features extracted from DQP  
 $Feature_{Tol}$  ▷ Tolerance of number of selected features

**Output:**

$Features_{Fed}$  ▷ Federated Selected Features

**//Aggregating Features based on FeatureValue, Outlier and Missing Data from n nodes**

```

1: procedure FEDERATEDFEATURESELECTION( $S_n, DQP_{Features}, Feature_{Tol}$ )
2:    $Features_{Agg} \leftarrow \sum_{e=1}^n DQP_{Features}(FeatureValue, Outlier, Missing)$ 
3:    $FeatureValue_{Rank} \leftarrow \text{sort DESC } Features_{Agg}(FeatureValue)$ 
4:    $OutlierData_{Rank} \leftarrow \text{sort } Features_{Agg}(Outlier)$ 
5:    $MissingData_{Rank} \leftarrow \text{sort } Features_{Agg}(Missing)$ 
6:    $[Feature]_{Rank} \leftarrow FeatureValue_{Rank} + OutlierData_{Rank} + MissingData_{Rank}$ 
7:    $Features_{Fed} \leftarrow \text{sort } [Feature]_{Rank}$  limit by  $Feature_{Tol}$ 
8:   return( $Features_{Fed}$ )
9: end procedure
    
```

**Table 3.** Illustration of federated feature selection.

Federated Feature Selection				
Ranking Criteria	Best Rank: 1 (Most Feature Value) Worst Rank: N	Best Rank: 1 (Least Outlier %) Worst Rank: N	Best Rank: 1 (Least Missing %) Worst Rank: N	Best Rank: 1 (Most Valuable) Worst Rank: N
Feature (N)	Aggregated Feature Selection Rank	Aggregated Feature Outlier Rank	Aggregated Feature Missing Rank	Federated Feature Rank
A	3	1	8	12
B	5	4	7	16
C	9	8	2	19
D	1	2	5	8
E	7	23	13	43
...	...	...	...	...

3.6. Computational Complexity of FDQP

The overall computational complexity of the FDQP approach includes the computational complexity of the FDQP algorithm and the federated feature selection algorithm.

3.6.1. Federated Data Quality Profiling Algorithm (FDQP)

**Baseline profiling:** The complexity of creating the baseline DQ profile ( $DQP_v$ ) is  $O(D * W * R)$ , where  $D$  represents the number of attributes,  $W$  represents the number of data quality dimensions, and  $R$  represents the number of rules.

**Edge profiling:** Creating the edge workload profile ( $WLP_e$ ) at each edge node can be assumed to have a complexity of  $O(1)$ , as it depends on the specific configuration and real-time parameters.

**Federated profiling:** Aggregating the DQPs from multiple edge nodes involves combining the profiles, which can be completed in linear time, resulting in a complexity of  $O(E * D * W * R)$ , where  $E$  is the number of edge nodes.

**Edge Processing:** Updating the local client profile ( $DQEnrichedData_e$ ) based on the federated profile ( $DQP_{Fed}$ ) can be assumed to have a complexity of  $O(1)$ , as it depends on the specific update operation.

### 3.6.2. Federated Feature Selection Algorithm

**Aggregating features** based on feature values, outliers, and missing data from  $E$  node edges can be performed in linear time, resulting in a complexity of  $O(E)$ .

**Sorting the aggregated features** based on ranks can be performed in  $O(N \log N)$  time complexity, where  $N$  is the number of features.

**Selecting the top  $K$  features** based on the tolerance ( $Feature_{Tol}$ ) can be done in  $O(N)$  time complexity.

Thus, the overall complexity of the federated feature selection algorithm is  $O(N \log N + N)$ , which is  $= O(N \log N)$ .

**Overall Complexity = Baseline Profiling + Edge Profiling + Federated Profiling + Edge Processing + Federated Feature Selection**

Therefore, the overall computational complexity can be expressed as:

$$O(D * W * R + E * D * W * R + N \log N) = O(E * D * W * R + N \log N)$$

It is linear with the number of edges, and has log-linear complexity with the number of features. It is important to note that the above analysis focuses on the computational complexity of the algorithm itself. Other factors, such as data transfer, network latency, and edge node resources, can also impact the overall performance in edge computing environments. Thorough investigation and benchmarking considering the specific constraints and requirements of the edge environment are necessary to assess the computational demands of the FDQP approach accurately.

### 3.7. DQ XML Profile Illustration

All of our DQPs are written in XML format, which are lightweight and easily readable by humans and machines alike. This makes them incredibly versatile, allowing for the data to be used in various application contexts and shared between any OS platforms at different nodes. A representative XML profile at one of the selected nodes after LDQP processing is shown in Figure 4. The profiling process takes care of the following, which can provide valuable insight into the structure of the data and make it easier to manage DQ.

1. Attribute-wise feature updating that includes maximum value, minimum value, mode, uniqueness, skew, outlier detected, uniqueness, etc. The XML profile structure includes these elements, allowing for a more thorough comprehension of the data.
2. Missing value identification and reporting in XML (attribute-wise, overall). It also allows for quick debugging, as well as making it easier to identify and deal with data anomalies.
3. How missing values were removed is detailed, using either a specified threshold or specific criteria.
4. Missing data imputation rules and the DQ measurements
5. Rank the attributes for feature selection according to their relative importance, and the DQ metrics are evaluated.

Figure 5 depicts the federated XML attributes snippets as a subset of the full FDQP attributes in the FDQP model. We can see that according to the FDQP formulation in Section 3.3, only the resulting decisions are highlighted in FDQP. The global threshold value for all the nodes has been established for the delete rows as seen in XML. The selected features are listed according to the algorithm presented in federated feature selection (Algorithm 2), and the data imputation rules for each attribute common to all nodes are presented. After applying FDQP to the data along all of its edges and enforcing constraints in accordance with FDQP's rules, we remeasure the data metrics to assess how well the profiling worked. We can guarantee that as the data model evolves, so will the associated attribute-level constraints, allowing for a broader range of query and data manipulation capabilities.

```

<root_element>
  <DataSource name="Source1">
    <Dataset name="Node1">
      <DataProfile>
        <AttributeCount>21</AttributeCount>
        <Attributes>
          <column name="baseline value" type="float64" minvalue="106.0" maxvalue="160.0" mean="133.25" mode="133.0" uniqueness="48"
            skew="0.03" std="9.84" missing_data_count="216" missing_percentage="0.1%" outliers_count="0"/>
          <column name="accelerations" type="float64" minvalue="0.0" maxvalue="0.019" mean="0.0" mode="0.0" uniqueness="20"
            skew="1.21" std="0.0" missing_data_count="201" missing_percentage="0.09%" outliers_count="12"/>
          <column name="fetal_movement" type="float64" minvalue="0.0" maxvalue="0.481" mean="0.01" mode="0.0" uniqueness="96"
            skew="7.95" std="0.05" missing_data_count="213" missing_percentage="0.1%" outliers_count="267"/>
          ...
        </Attributes>
        <MissingValueRows>
          <MissingRows>
            <missingdata rows="779" missing_val_total_count="4314"/>
          </MissingRows>
        </MissingValueRows>
      </DataProfile>
      <DQList>
        <MissingRowsRemoved>
          <DQRows threshold="40%" rows_removed="131" accuracy="94.54%"/>
        </MissingRowsRemoved>
        <MissingDataImputation>
          <DQD name="accuracy" value="92.59%"/>
          <DQD name="completeness" value="89.0%"/>
          <DQD name="consistency" value="77.39%"/>
        </MissingDataImputation>
      </DQList>
      <QualityProfile>
        <SelectedFeaturesCount>19</SelectedFeaturesCount>
        <SelectedFeaturesAccuracy>92.59%</SelectedFeaturesAccuracy>
        <SelectedFeatures>
          <feature name="baseline value" rank="0.1302"/>
          <feature name="accelerations" rank="0.1252"/>
          ...
          <feature name="histogram_tendency" rank="0.0301"/>
        </SelectedFeatures>
        <ImputationRules>
          <ImputationRule name="baseline value" rule="(int(df['baseline value'].mean()))"/>
          <ImputationRule name="accelerations" rule="(round(df['accelerations'].mean(),3))"/>
          ...
        </ImputationRules>
      </QualityProfile>
    </Dataset>
  </DataSource>
</root_element>

```

Figure 4. Node 1 consolidated LDQP-snippet.

```

<FDQP>
  <FederationRules>
    <DeleteRows scope="all">
      <threshold value="10"/>
    </DeleteRows>
  </FederationRules>
  <QualityProfile>
    <SelectedFeatures>
      <feature name="abnormal_short_term_variability"/>
      <feature name="histogram_min"/>
      ...
      <feature name="light_decelerations"/>
    </SelectedFeatures>
    <ImputationRules>
      <ImputationRule name="light_decelerations" rule="(round(df['light_decelerations'].mean(),3))"/>
      <ImputationRule name="severe_decelerations" rule="(df['severe_decelerations'].mode().iloc[0])"/>
      ...
      <ImputationRule name="histogram_tendency" rule="(df['histogram_tendency'].value_counts().index[0])"/>
    </ImputationRules>
  </QualityProfile>
</FDQP>

```

Figure 5. FDQP XML snippet.

#### 4. Experimental Evaluation

This section describes the experiments conducted to evaluate our proposed FDQ profiling model. The primary goals are to evaluate the data quality and the accuracy of the training model both before and after the FDQ profiling model is applied to the data at the edges. In the subsections that follow, we will describe the various aspects of the experiments conducted, such as the experimental setup, the dataset employed, the experimental design, and the various scenarios tested. We will conclude by discussing the results, which depict an improvement in accuracy, and the reasoning behind it. Initial assumption: We assume that the data collected at the nodes are homogenous and identically distributed (i.i.d).

#### 4.1. Dataset

The dataset we chose for our experiments contains 2126 instances and 23 attributes derived from cardiocograms, which are continuous measurements of the fetal heart rate using an ultrasound transducer placed on the mother's abdomen and categorized by expert obstetricians. The parameters used for data analysis are instantaneous fetal heart rate (FHR) and simultaneously communicated uterine contraction signals. The classification results were based on the fetal state labels (N = normal; S = suspect; P = pathologic) [56].

For the selection stage and reduction of attributes, in particular the numerical input data and a categorical (class) target variable, there are two well-known feature selection techniques, mainly ANOVA-f statistics and mutual information statistics. The results of this test can be used for feature selection by removing from the dataset those features that are independent of the target variable.

#### 4.2. Experiment Setup

All of our experiments for this study were conducted in IPython, which is enhanced interactive Python software V.3.11.1 for the experimental configuration that includes multiple machine learning libraries for deep learning and federated learning. The entire Fetal Health dataset is randomly divided into five datasets, each corresponding to one of the five edge nodes. The dataset has a few quality issues, so we synthesized errors and noise at the dataset's edges to reflect a true clinical distribution and to illustrate how FDQP will improve it.

#### 4.3. Scenarios—Proposed FDQP Evaluation

To evaluate our proposed FDQ profiling model, we implemented eight scenarios. The first scenario evaluates the data quality profiling with the main focus on accuracy considering baseline accuracy, after missing data imputation, applying local DQP (LDQP), which includes feature selection and rules application. The second scenario illustrates the node selection criteria defined in FDQP, which takes into consideration the completeness and consistency of DQ parameters. In the third scenario, the federated feature selection and ranking are evaluated. The fourth scenario compares the accuracy of the FDQP to that of the LDQP and the baseline. In the fifth scenario, the accuracy, completeness, and consistency of the data quality metrics are examined before and after FDQP. The sixth scenario analyzes the accuracy of various classifiers using FQQP to determine the most accurate classifier. The seventh scenario indicates the number of features selected and the associated training time at each node. Finally, the eighth scenario depicts the PSN accuracy before and after FDQP.

##### 4.3.1. Scenario 1

Create a data quality profile (DQP) based on the XML file containing the dataset and send it to the edge nodes. The quality characteristics of the profile are updated by generating LDQP and sent to the server. As shown in Figure 6, LDQP is applied at the edge node and the experiment's accuracy is evaluated node-by-node with baseline and after each of the processes (missing data imputation and LDQP).

Node 2 showed the highest increase in accuracy, with a boost from 70.71% to 91.92% after LDQP, and Node 4 exhibited a notable improvement in accuracy, increasing from 65.13% to 89% after LDQP. Node 3 demonstrated a considerable increase in accuracy after rows were removed, with the accuracy rising from 53.66% to 83.53%. This significant enhancement underscores the impact of data preprocessing steps, such as removing rows with missing or incomplete data, in improving the overall accuracy within the FDQP framework.

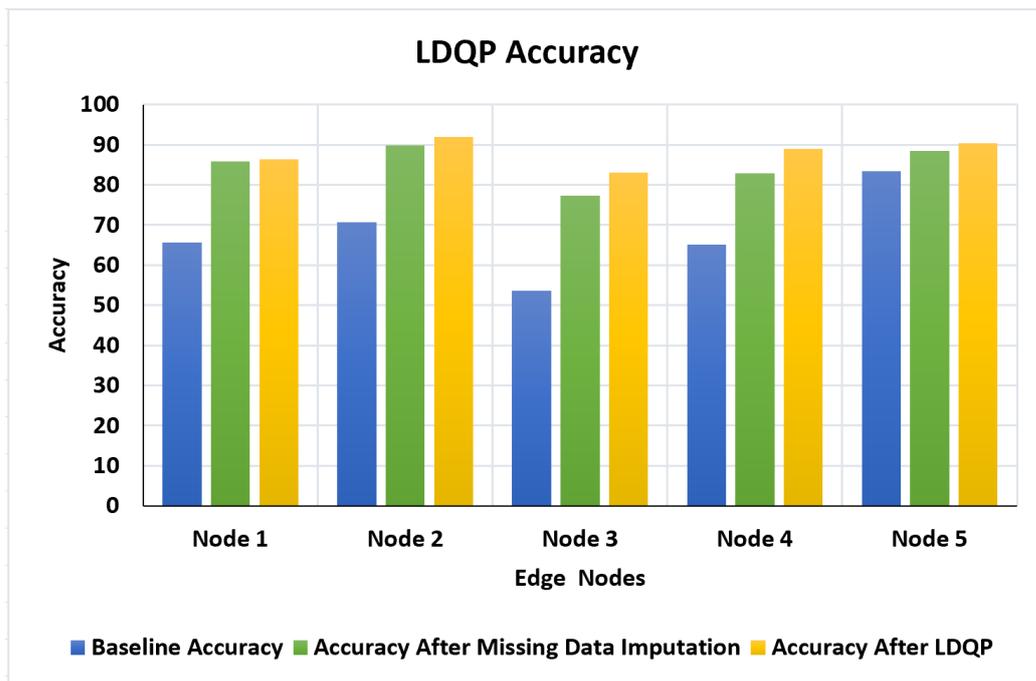


Figure 6. Local data quality profile (LDQP) accuracy evaluation.

4.3.2. Scenario 2

Edge node selection occurs before FDQP is applied at the edge node. In other words, LDQP will be used to select nodes on the server. If the DQ metrics values (accuracy, completeness or consistency) are less than the tolerance, the node is eliminated.

One of the first and most important steps in our proposed FDQ profiling is assessing the edge-level distribution of classes and establishing criteria for node selection. Figure 7 illustrates the problem with the consistency and completeness of the Node-3 dataset. The FDQ profiling criteria determine what levels of information must be present in a node’s representation. Accordingly, Node-3 is eliminated from further processing after failing to comply with the DQ requirements.

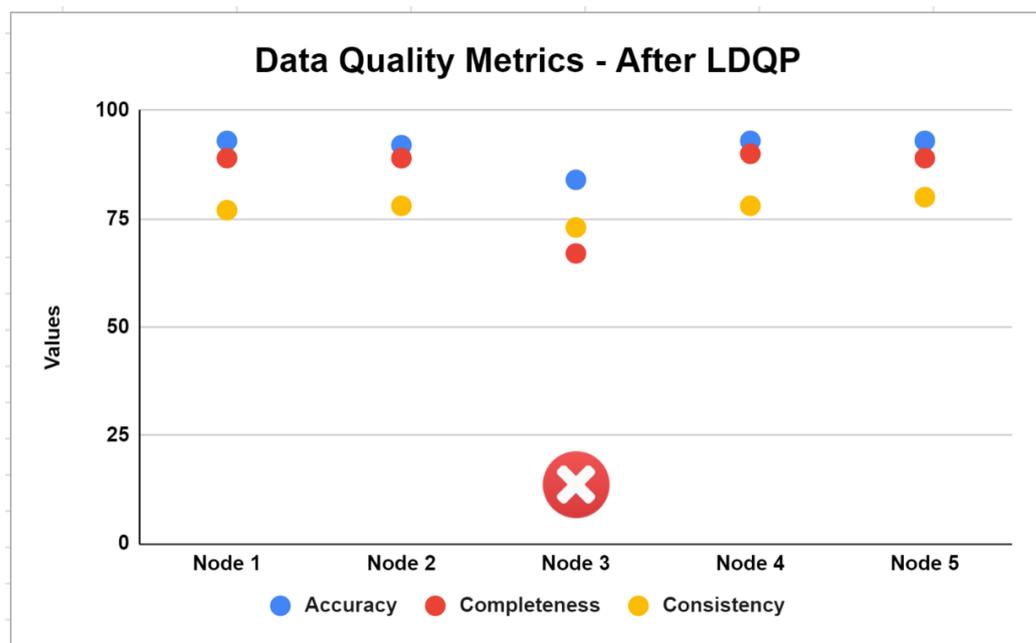
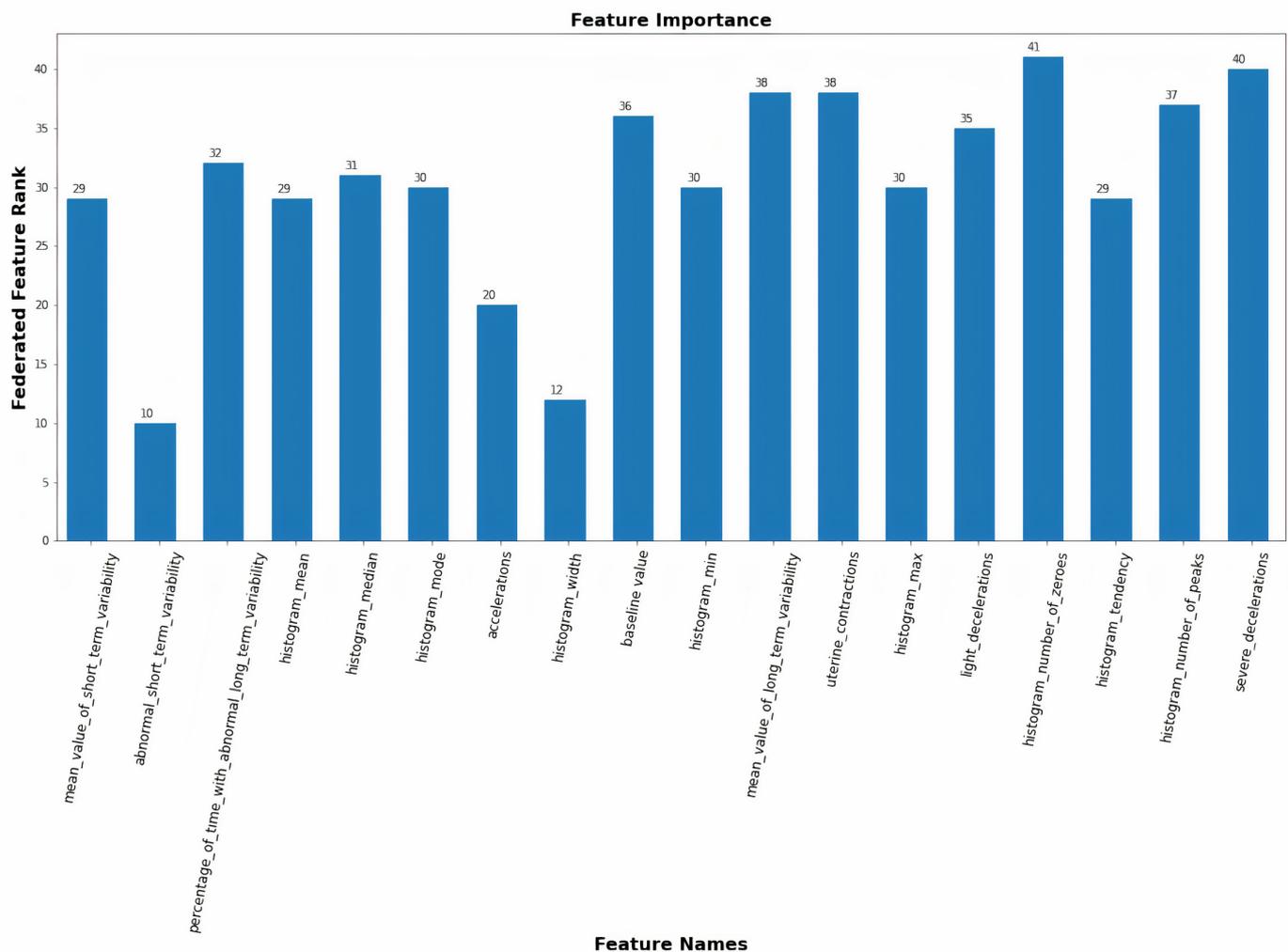


Figure 7. Node selection based on LDQP data quality metrics.

#### 4.3.3. Scenario 3

The server federates the profile data from the edges. Features are eliminated (feature selection) according to the feature selection algorithm, and the retained features are documented in the feature driven FDQP. The resulting FDQP includes data imputation rules and is transmitted to the edge nodes. In order to assess the PSN's accuracy and precision, it is necessary to apply FDQP to the edge node. Figure 8 depicts the federated feature rank for each of the attributes in our experimental dataset.



**Figure 8.** Federated feature selection.

#### 4.3.4. Scenario 4

Evaluation of FDQ profiles is one of the most crucial aspects of our experimentation. Node 3 has already been deleted, and the accuracy after FDQP profiling for the remaining nodes is compared to accuracy after LDQP and baseline accuracy, as seen in Figure 9. We can see that accuracy has improved significantly around 10% with LDQP and to a maximum of 5% with FDQP.

#### 4.3.5. Scenario 5

Before and after applying FDQP, we analyzed the metrics for data quality. Figure 10 shows that the coefficient of variation was reduced following FDQP, suggesting enhanced consistency and that the completeness factor was increased to 100%, indicating that certain data quality issues had been addressed by the process.

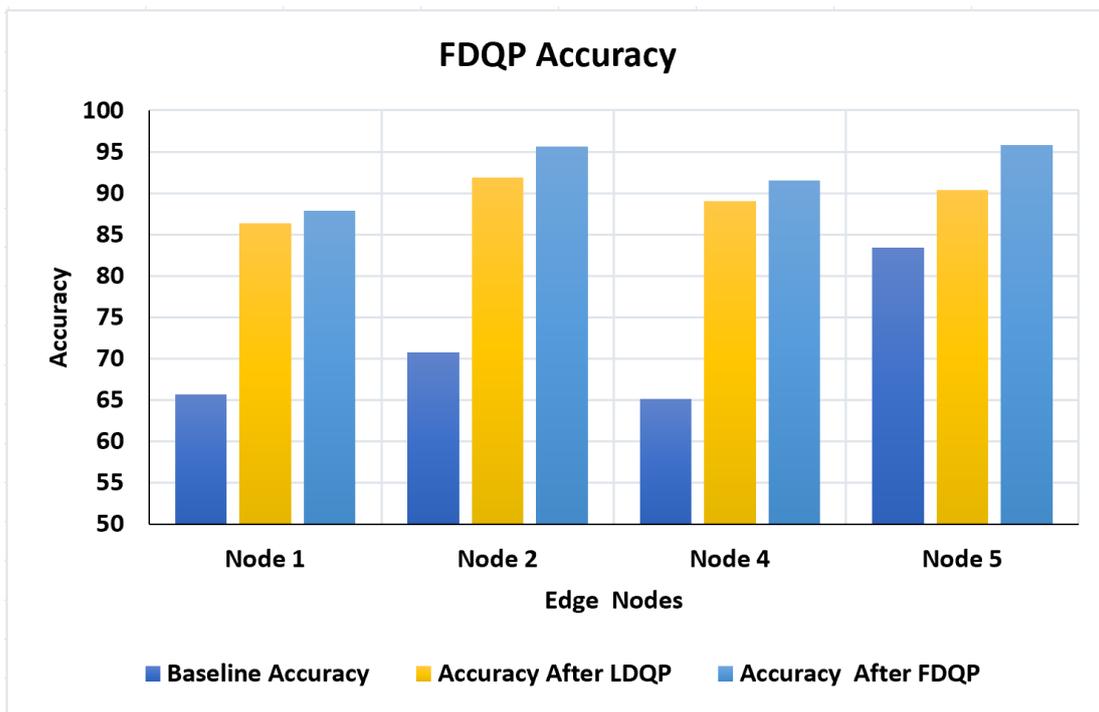


Figure 9. Accuracy (baseline, after LDQP, and after FDQP).

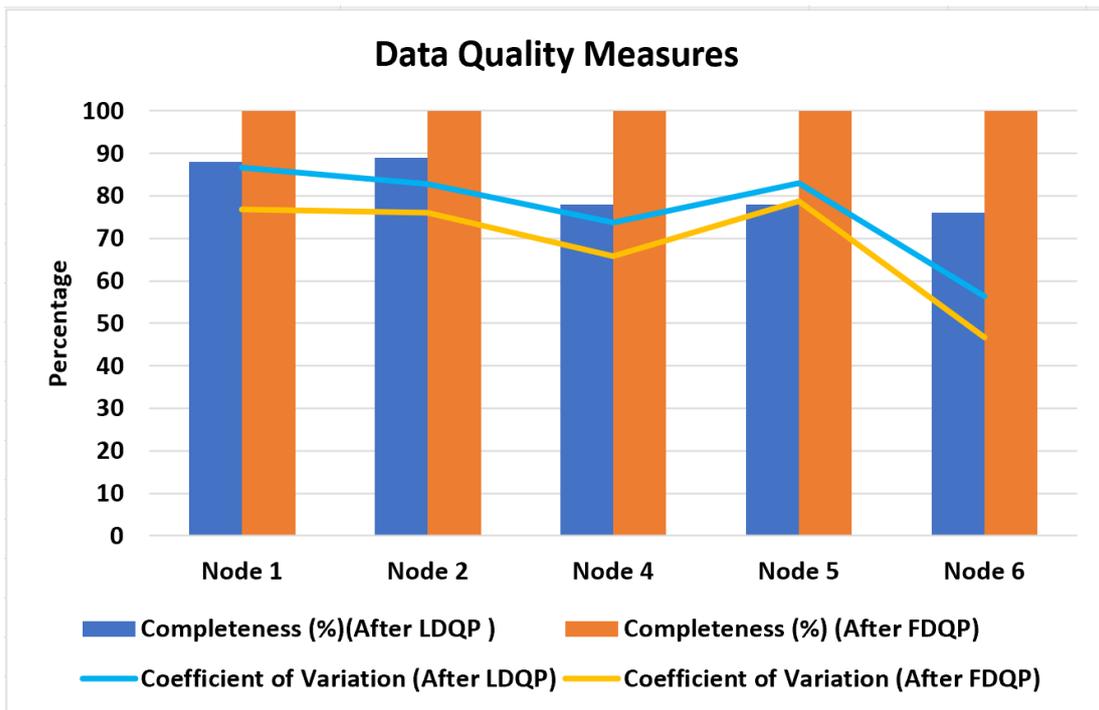


Figure 10. Data quality metrics assessment after LDQP and FDQP.

#### 4.3.6. Scenario 6

In this scenario, an evaluation is performed using various ML models to determine which classifier is the most accurate in forecasting fetal health at each node, and the results are displayed in Figure 11. Both the random forest classifier and the decision tree classifier have been shown to have the best performance in all of the edge nodes. Edge 2 had the greatest accuracy gain, with random forest classifier attaining 95

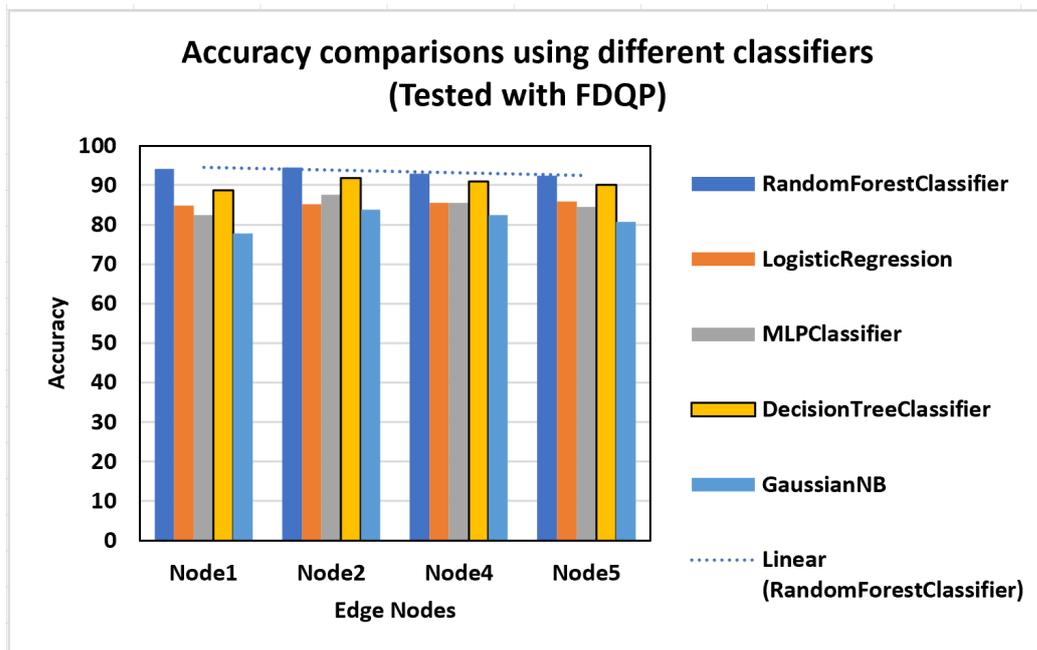


Figure 11. Accuracy comparisons with different classifiers.

4.3.7. Scenario 7

The selection of features is one of the primary characteristics of our suggested FDQP, and its evaluation can be found in Figure 12. We can see that the number of features has been reduced in each of the nodes, which has led to a shorter amount of training time when compared to the initial amount of training time.

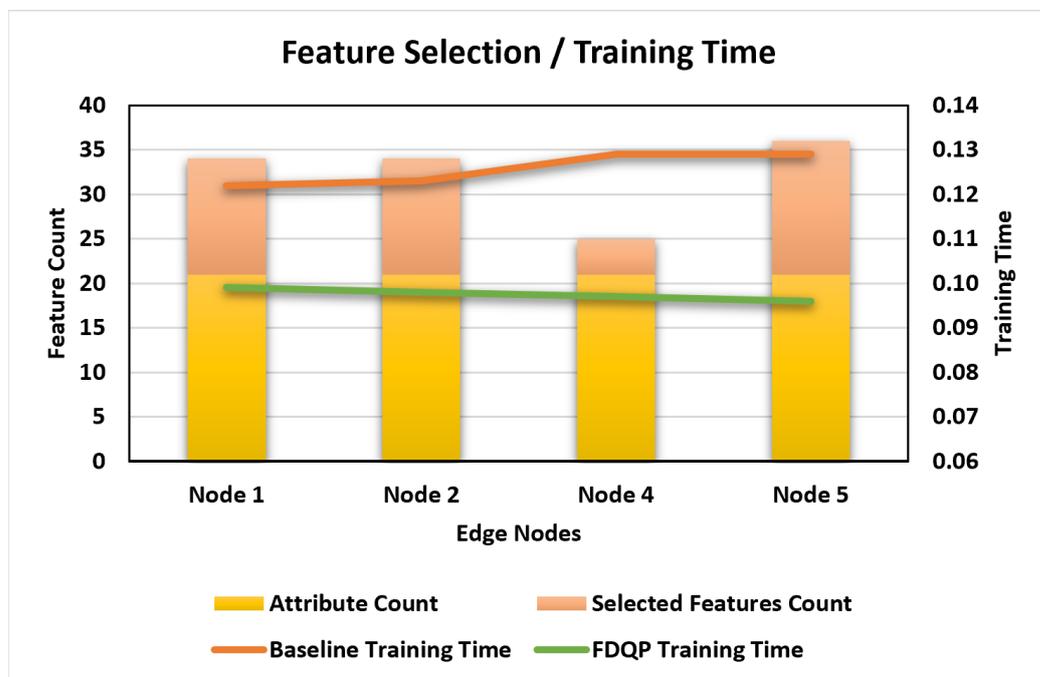


Figure 12. FDQP feature selection vs. training time.

4.3.8. Scenario 8

Patient similarity evaluation is performed after FDQP is reviewed, and we observed in Figure 13 that FDQ profiling has unquestionably increased the accuracy, with an average increase of 7% and a maximum gain of 9% accuracy.

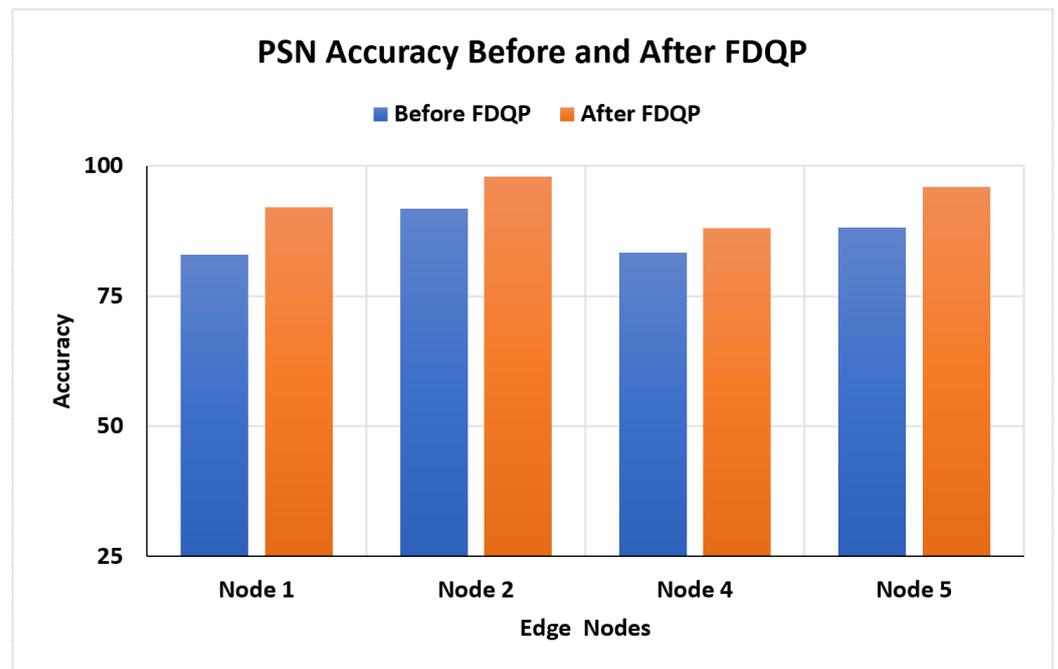


Figure 13. PSN accuracy before and after federated data quality profiling.

## 5. Discussion

The experimental evaluation of our proposed FDQ profiling model provided valuable insights into its effectiveness and impact. The key findings from the experiments are as follows:

- **Improved accuracy:** FDQ profiling resulted in an average accuracy increase of 10% with LDQP and a maximum improvement of 15%, attributed to data quality enhancement through missing data imputation, feature selection, and rule-based processing.
- **Enhanced data quality metrics:** FDQ profiling significantly improved consistency and completeness, reducing the coefficient of variation and achieving 100% completeness.
- **Effective node selection:** FDQ profiling's node selection criteria successfully identified high-quality nodes, ensuring accuracy improvement by considering completeness and consistency.
- **Classifier evaluation:** Random Forest Classifier and Decision Tree Classifier consistently achieved the highest accuracy in fetal health forecasting, suggesting the potential for selecting the best classifier based on dataset and node characteristics.
- **Patient similarity accuracy:** FDQ profiling improved patient similarity accuracy by 7% on average and up to 9%, which can have significant implications for various applications, such as personalized medicine and recommendation systems.
- **Profile aggregation and optimized feature selection:** Profile aggregation with federated feature selection of attributes from various nodes can improve the efficiency of discriminative features and restrain interference from relatively ineffective features. They are calculated by feature aggregation and then optimized via the proposed rules for elimination, combining the idea of survival of the fittest. The feature with low weight is eliminated in the experiments, leading to improved accuracy and reduced training time.
- **Ensemble-like results:** We can see that the results obtained by profiling the data at multiple nodes are, in some cases, more stable with the overall ranking concerning management and resource utilization. The concept of distributing features across nodes and then federating the profile's results into a final one is analogous to that of ensemble learning or a mixture of experts, and produces more reliable results than a single expert.

In summary, the experimental evaluation showcased the effectiveness of FDQ profiling in enhancing data quality and improving the accuracy of machine learning models. The results not only validate the approach, but also highlight its potential for broader applications across various domains beyond healthcare. By prioritizing data quality components and selecting feature nodes based on relevant metrics, FDQ profiling emerges as a valuable tool for optimizing data quality and advancing machine learning model classification.

The FDQP approach proposed in this paper is founded on several underlying principles that make it a promising technique for evaluating the quality of patient data collected from edge nodes in federated environments. Understanding these principles is essential to value the relevance and potential of FDQP across various domains.

**Federated learning and data quality profile (DQP):** The foundation of FDQP lies in the federated learning paradigm, which enables collaborative model training across distributed edge nodes without centralizing raw data. By preserving data privacy at the edge, FDQP addresses the challenges of data silos and privacy concerns in healthcare and other domains. The DQP encapsulates data quality dimensions, defining a framework to assess the quality of data attributes, such as completeness, accuracy, and consistency. FDQP encapsulates DQPs from different edge nodes into a unified formal model. This formalization enables seamless comparisons, aggregation, and analysis of data quality metrics, facilitating better decision making during model aggregation in federated learning.

**Federated feature selection:** FDQP employs federated feature selection, which combines local feature selections at each edge node and global feature ranking. This technique enhances the precision of classifiers by selecting relevant features while mitigating the impact of noisy or irrelevant attributes. The use of outlier percentage and missing data percentage as criteria in feature selection makes FDQP robust to variations in data quality across different edge nodes.

**Lightweight profile exchange:** FDQP introduces a lightweight profile exchange mechanism based on XML that shares summary statistics of data quality attributes among edge nodes. This exchange avoids the transmission of raw data, optimizing data quality achievement, and improving the overall efficiency of the federated learning process. This approach is particularly relevant in resource-constrained edge computing environments.

**Enhancing data quality and accuracy:** The primary motivation behind FDQP is to enhance data quality and, consequently, the accuracy of federated learning models. By identifying and quantifying data quality issues at the edge nodes, FDQP enables targeted data cleaning, outlier detection, and imputation strategies. This results in improved data quality, reducing the impact of noisy or biased data on the global model's performance.

**Scalability and efficiency:** FDQP addresses the scalability and efficiency challenges of traditional centralized data quality assessment methods. By conducting quality evaluations locally at the edge nodes and sharing aggregated quality metrics, FDQP preserves privacy and reduces the risk of data breaches.

### *Limitations*

While our current work focuses on addressing the challenges of federated data quality profiling in edge computing environments, there are still a few challenges that warrant further investigation and future research.

**Generalizability:** The methodology's validation on a specific fetal dataset raises concerns about its generalizability to other types of patient data and different domains. Further research, including different types of data and real-world scenarios, would provide a more comprehensive evaluation of FDQP's effectiveness.

**Trustworthiness of edge nodes:** The assumption of reliable and trustworthy edge nodes might not always align with real-world scenarios. FDQP should be examined under situations where edge nodes might be unreliable or malicious to ensure the approach's robustness and security.

**Limited quality dimensions:** The current implementation of FDQP focuses on specific data quality dimensions such as accuracy, completeness, and consistency. Extending the

model to incorporate additional quality dimensions could provide a more comprehensive evaluation.

**Privacy and security:** While FDQP incorporates security measures through federated learning, it might not address extremely stringent privacy and security requirements, particularly in high-sensitivity data scenarios. Future research should explore additional privacy-enhancing techniques to accommodate diverse security needs.

By recognizing these remaining challenges and conducting further research to address them, the FDQP approach can be refined to enhance its applicability, robustness, and credibility. This will facilitate its adoption in various domains, leading to improved data quality, more accurate machine-learning models, and better decision-making processes.

## 6. Conclusions and Future Work

To ensure reliable and meaningful insights from any data analytics process, quality is a determining factor of utmost importance. The end-to-end process of data integrity is required to achieve DQ, which is attained through a meticulous process of data cleansing and data governance. Poor data, on the other hand, will have an adverse effect on ML classification results because it will lead to poor analytics. In conclusion, we have demonstrated that FDQ profiling can handle various quality issues at multiple edges while keeping data localized, thereby enabling data privacy and reducing data resource consumption and transportation costs. Erroneous data reduce the performance of learning algorithms, but this is the first time federated profiling has been used to mitigate the consequences of improving data quality and enhancing ML classification at multiple edges. In our future research, we plan to extend the application of FDQ profiling beyond healthcare and explore its synergies with other domains, such as finance, e-commerce, manufacturing, and more, with the goal of optimizing data quality and decision-making processes in diverse industries. By embracing this interdisciplinary approach, we aim to contribute to the advancement of data quality management practices and drive innovation in various fields. In addition, we will explore potential advancements on edge node data compression and PSN similarity as a powerful tool for data imputation, which has the potential to significantly enhance data quality.

**Author Contributions:** A.N.N. conceived the main conceptual ideas related to data-quality-aware FPSN (DQAFPSN), architecture, literature, and overall implementation and execution of experimentation. H.T.E.K. contributed to the formal modeling of FDQP, the literature review, and the analysis of the results. M.A.S. contributed to the architecture of the DQAFPSN model, the identification of scenarios, and he ensured that the study was carried out with utmost care and attention to detail, while overseeing the overall direction and planning. I.T. was involved in the general evaluation of the proposed model. All authors have read and agreed to the published version of the manuscript.

**Funding:** This work is supported by the Zayed health science center under fund # 12R005.

**Institutional Review Board Statement:** Not applicable

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The dataset is available at UCI Machine Learning Repository, <https://archive.ics.uci.edu/ml/datasets/Cardiotocography> (accessed on 9 June 2023).

**Conflicts of Interest:** The authors declare no conflict of interest.

## Abbreviations

The following abbreviations are used in this manuscript:

DQ	Data quality
FL	Federated learning
MI	Multiple imputations
ML	Machine learning
SI	International system of units
DQA	Data quality assessment

DQP	Data quality profile
EHR	Electronic health record
FDQ	Federated data quality
FHR	Fetal heart rate
IID	Independent and identically distributed
IoT	Internet of things
MCS	Mobile crowd sourcing
PSN	Patient similarity network
XML	Extensible markup language
FDQP	Federated data quality profile
FPSN	Federated patient similarity network
LDQP	Local data quality profile
TDQM	Total data quality methodology
DQA-FPSN	Data-quality-aware federated PSN

## References

- Klein, A.; Lehner, W. Quality and Performance Optimization of Sensor Data Stream Processing. *Int. J. Adv. Netw. Serv.* **2010**, *3*, 249–262.
- Tian, M.W.; Yan, S.R.; Guo, W.; Mohammadzadeh, A.; Ghaderpour, E. A New Task Scheduling Approach for Energy Conservation in Internet of Things. *Energies* **2023**, *16*, 2394. [[CrossRef](#)]
- Okafor, N.U.; Alghorani, Y.; Delaney, D.T. Improving Data Quality of Low-cost IoT Sensors in Environmental Monitoring Networks Using Data Fusion and Machine Learning Approach. *ICT Express* **2020**, *6*, 220–228. [[CrossRef](#)]
- Kramer, O.; Even, A.; Matot, I.; Steinberg, Y.; Bitan, Y. The impact of data quality defects on clinical decision-making in the intensive care unit. *Comput. Methods Progr. Biomed.* **2021**, *209*, 106359. [[CrossRef](#)] [[PubMed](#)]
- Chien, M.; Jain, A. *Gartner Report 2021—Magic Quadrant for Data Quality Solutions*; Technical Report; Gartner, Inc.: Stamford, CT, USA, 2021.
- Bello, M.; Nápoles, G.; Vanhoof, K.; Bello, R. Data quality measures based on granular computing for multi-label classification. *Inf. Sci.* **2021**, *560*, 51–67. [[CrossRef](#)]
- Loshin, D. Data Profiling. *The Practitioner's Guide to Data Quality Improvement*; Elsevier: Amsterdam, The Netherlands, 2011; pp. 241–259. [[CrossRef](#)]
- Taleb, I.; Serhani, M.A.; Bouhaddioui, C.; Dssouli, R. *Big Data Quality Framework: A Holistic Approach to Continuous Quality Management*; Springer International Publishing: Berlin/Heidelberg, Germany, 2021; Volume 8. [[CrossRef](#)]
- Vaziri, R.; Mohsenzadeh, M.; Habibi, J. Measuring data quality with weighted metrics. *Total Qual. Manag. Bus. Excell.* **2019**, *30*, 708–720. [[CrossRef](#)]
- Cichy, C.; Rass, S. An Overview of Data Quality Frameworks. *IEEE Access* **2019**, *7*, 24634–24648. [[CrossRef](#)]
- Heinrich, B.; Hristova, D.; Klier, M.; Schiller, A.; Szubartowicz, M. Requirements for data quality metrics. *J. Data Inf. Qual.* **2018**, *9*, 1–32. [[CrossRef](#)]
- Patterson, C. The Six Primary Dimensions for Data Quality Assessment: Defining Data Quality Dimensions. 2017. Available online: <https://silotips.com/download/the-six-primary-dimensions-for-data-quality-assessment> (accessed on 20 March 2023).
- Kirchen, I.; Schutz, D.; Folmer, J.; Vogel-Heuser, B. Metrics for the evaluation of data quality of signal data in industrial processes. In Proceedings of the 2017 IEEE 15th International Conference on Industrial Informatics, INDIN 2017, Emden, Germany, 24–26 July 2017; pp. 819–826. [[CrossRef](#)]
- Ballou, D.; Wang, R.; Pazer, H.; Tayi, G.K. Modeling information manufacturing systems to determine information product quality. *Manag. Sci.* **1998**, *44*, 462–484. [[CrossRef](#)]
- Blake, R.; Mangiameli, P. The effects and interactions of data quality and problem complexity on classification. *J. Data Inf. Qual.* **2011**, *2*, 1–28. [[CrossRef](#)]
- Navaz, A.N.; El-kassabi, H.T.; Serhani, M.A.; Oulhaj, A.; Khalil, K. A Novel Patient Similarity Network (PSN) Framework Based on Multi-Model Deep Learning for Precision Medicine. *J. Pers. Med.* **2022**, *12*, 768. [[CrossRef](#)]
- Lee, G.H.; Shin, S.Y. Federated Learning on Clinical Benchmark Data: Performance Assessment. *J. Med. Internet Res.* **2020**, *22*, e20891. [[CrossRef](#)]
- Decker, H. Answers That Have Quality. In *Computational Science and Its Applications, Proceedings of the ICCSA 2013, Ho Chi Minh City, Vietnam, 24–27 June 2013*; Murgante, B., Misra, S., Carlini, M., Torre, C.M., Nguyen, H.Q., Taniar, D., Apduhan, B.O., Gervasi, O., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 543–558. [[CrossRef](#)]
- Ramaswamy, L.; Lawson, V.; Gogineni, S.V. Towards a quality-centric big data architecture for federated sensor services. In Proceedings of the 2013 IEEE International Congress on Big Data, BigData 2013, Santa Clara, CA, USA, 27 June–2 July 2013; pp. 86–93. [[CrossRef](#)]
- Wu, W.; He, L.; Lin, W.; Mao, R. FedProf: Efficient Federated Learning with Data Representation Profiling. *arXiv* **2021**, arXiv:2102.01733.

21. Wang, X.; Han, Y.; Wang, C.; Zhao, Q.; Chen, X.; Chen, M. In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning. *IEEE Netw.* **2019**, *33*, 156–165. [[CrossRef](#)]
22. Habib Ur Rehman, M.; Mukhtar Dirir, A.; Salah, K.; Svetinovic, D. Fairfed: Cross-device fair federated learning. In Proceedings of the Applied Imagery Pattern Recognition Workshop, Washington, DC, USA, 13–15 October 2020; Institute of Electrical and Electronics Engineers Inc.: Piscataway, NJ, USA, 2020; Volume 2020. [[CrossRef](#)]
23. Christy, A.; Gandhi, M.G.; Vaithyasubramanian, S. Cluster based outlier detection algorithm for healthcare data. *Procedia Comput. Sci.* **2015**, *50*, 209–215. [[CrossRef](#)]
24. Jang, W.J.; Lee, S.T.; Kim, J.B.; Gim, G.Y. A study on data profiling: Focusing on attribute value quality index. *Appl. Sci.* **2019**, *9*, 5054. [[CrossRef](#)]
25. Özsu, M.T.; Valduriez, P. Data Replication. In *Principles of Distributed Database Systems*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 247–280. [[CrossRef](#)]
26. Ghaderpour, E.; Mazzanti, P.; Mugnozsa, G.S.; Bozzano, F. Coherency and phase delay analyses between land cover and climate across Italy via the least-squares wavelet software. *Int. J. Appl. Earth Obs. Geoinf.* **2023**, *118*, 103241. [[CrossRef](#)]
27. Fantacci, R.; Picano, B. Federated learning framework for mobile edge computing networks. *CAAI Trans. Intell. Technol.* **2020**, *5*, 15–21. [[CrossRef](#)]
28. Nagalapatti, L.; Mittal, R.S.; Narayanam, R. Is Your Data Relevant?: Dynamic Selection of Relevant Data for Federated Learning. In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2022; Volume 36, pp. 7859–7867. [[CrossRef](#)]
29. Doku, R.; Rawat, D.B.; Liu, C. Towards federated learning approach to determine data relevance in big data. In Proceedings of the 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science, IRI, Los Angeles, CA, USA, 30 July–1 August 2019; pp. 184–192. [[CrossRef](#)]
30. Zhang, W.; Li, Z.; Chen, X. Quality-aware user recruitment based on federated learning in mobile crowd sensing. *Tsinghua Sci. Technol.* **2021**, *26*, 869–877. [[CrossRef](#)]
31. Che, L.; Long, Z.; Wang, J.; Wang, Y.; Xiao, H.; Ma, F. FedTriNet: A Pseudo Labeling Method with Three Players for Federated Semi-supervised Learning. In Proceedings of the 2021 IEEE International Conference on Big Data, Big Data 2021, Orlando, FL, USA, 15–18 December 2021; pp. 715–724. [[CrossRef](#)]
32. Wang, G. Interpret Federated Learning with Shapley Values. *arXiv* **2019**, arXiv:1905.04519.
33. Fan, Z.; Fang, H.; Zhou, Z.; Pei, J.; Friedlander, M.P.; Liu, C.; Zhang, Y. Improving Fairness for Data Valuation in Horizontal Federated Learning. In Proceedings of the International Conference on Data Engineering, Kuala Lumpur, Malaysia, 9–12 May 2022; pp. 2440–2453. [[CrossRef](#)]
34. Zhang, R.; Wang, Y.; Zhou, Z.; Ren, Z.; Tong, Y.; Xu, K. *Data Source Selection in Federated Learning: A Submodular Optimization Approach*; Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Springer: Berlin/Heidelberg, Germany, 2022; pp. 606–614. [[CrossRef](#)]
35. Geng, D.; He, H.; Lan, X.; Liu, C. An Adaptive Accuracy Threshold Aggregation Strategy Based on Federated Learning. In Proceedings of the 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering, ICBAIE, Nanchang, China, 26–28 March 2021; pp. 28–31. [[CrossRef](#)]
36. Batra, I.; Verma, S.; Malik, A.; Kavita; Ghosh, U.; Rodrigues, J.J.; Nguyen, G.N.; Hosen, A.S.; Mariappan, V. Hybrid logical security framework for privacy preservation in the green internet of things. *Sustainability* **2020**, *12*, 5542. [[CrossRef](#)]
37. Deng, Y.; Lyu, F.; Ren, J.; Wu, H.; Zhou, Y.; Zhang, Y.; Shen, X. AUCTION: Automated and Quality-Aware Client Selection Framework for Efficient Federated Learning. *IEEE Trans. Parallel Distrib. Syst.* **2022**, *33*, 1996–2009. [[CrossRef](#)]
38. Chai, Z.; Ali, A.; Zawad, S.; Truex, S.; Anwar, A.; Baracaldo, N.; Zhou, Y.; Ludwig, H.; Yan, F.; Cheng, Y. TiFL: A Tier-based Federated Learning System. In Proceedings of the HPDC 2020—29th International Symposium on High-Performance Parallel and Distributed Computing, Stockholm, Sweden, 23–26 June 2020; pp. 125–136. [[CrossRef](#)]
39. Wang, Z.; Zhu, Y.; Wang, D.; Han, Z. FedACS: Federated Skewness Analytics in Heterogeneous Decentralized Data Environments. In Proceedings of the 2021 IEEE/ACM 29th International Symposium on Quality of Service, IWQOS, Tokyo, Japan, 25–28 June 2021. [[CrossRef](#)]
40. Wang, L.; Xu, Y.; Xu, H.; Liu, J.; Wang, Z.; Huang, L. Enhancing Federated Learning with In-Cloud Unlabeled Data. In Proceedings of the International Conference on Data Engineering, Kuala Lumpur, Malaysia, 9–12 May 2022; pp. 136–149. [[CrossRef](#)]
41. Xu, Z.; Yu, F.; Xiong, J.; Chen, X. ELFISH: Resource-Aware Federated Learning on Heterogeneous Edge Devices. In Proceedings of the Design Automation Conference, San Francisco, CA, USA, 5–9 December 2021; pp. 997–1002. [[CrossRef](#)]
42. Chen, Y.; Ning, Y.; Slawski, M.; Rangwala, H. Asynchronous Online Federated Learning for Edge Devices with Non-IID Data. In Proceedings of the 2020 IEEE International Conference on Big Data, Big Data 2020, Atlanta, GA, USA, 10–13 December 2020; pp. 15–24. [[CrossRef](#)]
43. Li, T.; Sahu, A.K.; Zaheer, M.; Sanjabi, M.; Talwalkar, A.; Smith, V. Federated Optimization in Heterogeneous Networks. *arXiv* **2018**, arXiv:1812.06127.
44. Loog, M.; Duin, R.P.; Haeb-Umbach, R.; Chen, Y.; Ning, Y.; Slawski, M.; Rangwala, H.; Brendan McMahan, H.; Moore, E.; Ramage, D.; et al. Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS, Ft. Lauderdale, FL, USA, 20–22 April 2017; pp. 15–24. [[CrossRef](#)]

45. Abreha, H.G.; Hayajneh, M.; Serhani, M.A. Federated Learning in Edge Computing: A Systematic Survey. *Sensors* **2022**, *22*, 450. [[CrossRef](#)]
46. Li, T.; Sahu, A.K.; Talwalkar, A.; Smith, V. Federated Learning: Challenges, Methods, and Future Directions. *IEEE Signal Process. Mag.* **2020**, *37*, 50–60. [[CrossRef](#)]
47. Liu, Y.; Yuan, X.; Xiong, Z.; Kang, J.; Wang, X.; Niyato, D. Federated learning for 6G communications: Challenges, methods, and future directions. *China Commun.* **2020**, *17*, 105–118. [[CrossRef](#)]
48. Xiong, H.; Pandey, G.; Steinbach, M.; Kumar, V. Enhancing data analysis with noise removal. *IEEE Trans. Knowl. Data Eng.* **2006**, *18*, 304–319. [[CrossRef](#)]
49. Xu, J.; Du, W.; Jin, Y.; He, W.; Cheng, R. Ternary Compression for Communication-Efficient Federated Learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, *33*, 1162–1176. [[CrossRef](#)]
50. Deng, Y.; Lyu, F.; Ren, J.; Chen, Y.C.; Yang, P.; Zhou, Y.; Zhang, Y. Improving Federated Learning With Quality-Aware User Incentive and Auto-Weighted Model Aggregation. *IEEE Trans. Parallel Distrib. Syst.* **2022**, *33*, 4515–4529. [[CrossRef](#)]
51. Taik, A.; Moudoud, H.; Cherkaoui, S. Data-quality based scheduling for federated edge learning. In Proceedings of the Conference on Local Computer Networks, LCN, Edmonton, AB, Canada, 4–7 October 2021; pp. 17–23. [[CrossRef](#)]
52. Canonaco, G.; Bergamasco, A.; Mongelluzzo, A.; Roveri, M. Adaptive Federated Learning in Presence of Concept Drift. In Proceedings of the International Joint Conference on Neural Networks, Shenzhen, China, 18–22 July 2021; pp. 1–7. [[CrossRef](#)]
53. Amballa, A. Feature Engineering Part-1 Mean/ Median Imputation | Analytics Vidhya | Medium. 2018. Available online: <https://medium.com/analytics-vidhya/feature-engineering-part-1-mean-median-imputation-761043b95379> (accessed on 25 June 2023).
54. Dong, Y.; Peng, C.Y.J. Principled missing data methods for researchers. *SpringerPlus* **2013**, *2*, 222. [[CrossRef](#)]
55. Bolón-Canedo, V.; Sechidis, K.; Sánchez-Marño, N.; Alonso-Betanzos, A.; Brown, G. Insights into distributed feature ranking. *Inf. Sci.* **2019**, *496*, 378–398. [[CrossRef](#)]
56. Ayres-de campos, D.; Bernardes, J.; Garrido, A.; Marques-desá, J.; Pereira-Leite, L. SisPorto 2.0: A Program for Automated Analysis of Cardiotocograms. *J. Matern.-Fetal Neonatal Med.* **2000**, *9*, 311–318. [[CrossRef](#)]

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.