processes

MDPI

Article

A Healthcare Quality Assessment Model Based on Outlier Detection Algorithm

Nawaf Alharbe [1,*], Mohamed Ali Rakrouki [1,2,3] and Abeer Aljohani [1]

[1] Applied College, Taibah University, Medina 42353, Saudi Arabia; mrakrouki@taibahu.edu.sa (M.A.R.); aahjohani@taibahu.edu.sa (A.A.)
[2] Ecole Supérieure des Sciences Economiques et Commerciales de Tunis, University of Tunis, Tunis 1089, Tunisia
[3] Business Analytics and DEcision Making Lab (BADEM) at Tunis Business School, University of Tunis, Tunis 2059, Tunisia
* Correspondence: nrharbe@taibahu.edu.sa

Abstract: With the extremely rapid growth of data in various industries, big data is gradually recognized and valued by people. Medical big data, which can best reflect the significance of big data value, has also received attention from various parties. In Saudi Arabia, healthcare quality assessment is mostly based on human experience and basic statistical methods. In this paper, we proposed a healthcare quality assessment model based on medical big data in a region of Saudi Arabia, which integrated traditional evaluation methods and machine learning based techniques. Healthcare data has been accurate and effective after noise processing, and the outliers could reflect certain medical quality information. An improved *k*-nearest neighbors (KNN) algorithm has been proposed and its time complexity have been reduced to be more suitable for big data processing. An outlier indicator has been established based on statistical methods and the improved KNN algorithm. Experimental results showed that the proposed approach has good potential for detecting hospitals with financial fraud and poor-quality medical care.

Keywords: big data; machine learning; health informatics; KNN algorithm; statistics

Citation: Alharbe, N.; Rakrouki, M.A.; Aljohani, A. A Healthcare Quality Assessment Model Based on Outlier Detection Algorithm. *Processes* **2022**, *10*, 1199. https://doi.org/10.3390/pr10061199

Academic Editor: Chien-Chih Wang

Received: 27 April 2022
Accepted: 13 June 2022
Published: 16 June 2022

1. Introduction

Big data is data collection [1] with a large capacity, many types, fast access and high application value. Globally, big data is developing rapidly, and data has become a fundamental resource for all industries and even countries. It has become a computer technology and service industry that collects, stores, and data mines data from a wide range of sources, in various formats, and in large quantities, and then obtains new knowledge, new values, and new capabilities.

Healthcare is one of the most important pillars of any country, and medical data, as a carrier of information in this field, deserves to be dug deeper. Many countries are very active in promoting the development of medical information technology and medical big data [2–4], which gives the medical industry sufficient financial and human resources to analyze the data.

In Saudi Arabia, medical analysis is still basically based on statistical data analysis methods, and the concept of probability-based sampling statistics is to infer the overall state and behavior with a small number of random samples. The general steps of analysis are to follow the design of the questionnaire, data collection, cleaning, statistical analysis and finally the formation of a report. This process generally takes a long time, the data collection costs are high, and the analysis results are limited by the questionnaire. In short, it is the three major problems of sampling statistics, namely "slow, less, expensive". At the same time, since the reform and opening up, the healthcare field has also been gradually

transformed into a market-oriented economy, which has caused many medical institutions to focus only on economic efficiency and not on quality management, and its own monopoly property and non-transparent nature make quality control difficult. Furthermore, some hospitals have fatal management defects and safety hazards and various illegal behaviors are emerging, such as insurance fraud and excessive medical treatment. These factors emphasize the importance of healthcare quality assessment. The National Health Information Center, which is the basis of the experiment of this paper, is one of the most important data sources in healthcare big data in Saudi Arabia. Based on these data, this paper proposes a healthcare quality assessment model considering traditional statistical methods and machine learning algorithms. This evaluation system is applied to the hospitals of Saudi Arabia in order to evaluate the quality of healthcare for some types of diseases, especially for the detection of hospitals with financial fraud and medical vulnerability.

The remainder of this paper is organized as follows. In Section 2, some important related works and projects to healthcare quality assessment are presented. The description of the problem under consideration is presented in Section 3. Sections 4 and 5 are devoted the presentation of the outlier detection approaches based on statistics and KNN, respectively. An improved outlier detection algorithm is proposed to our problem in Section 6. The experimental results are presented in Section 7. Finally, Section 8 summarizes this research work.

2. Related Works and Projects

Big data in healthcare has emerged as an emerging technology in recent years. Currently, there are many healthcare quality assessment systems and projects proposed by relevant international organizations [5]. The United States has invested heavily in research applications for big data-related industries that include healthcare big data many years ago. The U.S. has also established a variety of healthcare quality assessment systems based on its own data. One of the most influential is the system proposed by the Agency for Healthcare Research and Quality [6]. Other evaluation systems have been proposed by the U.S. Baldrige National Quality Program (BNQP), the U.S. Maryland Hospital Evaluation System, the U.S. Joint Commission on the Evaluation of Health Organizations [7]. Of course, in addition to these professional organizations and hospitals, many third-party companies have also proposed quality assessment methods for hospitals, such as the U.S. News & World Report's hospital evaluation method, Truven Health's 100 Top Hospitals evaluation method, and so on. In addition to the U.S. research applications for healthcare quality assessment systems, the United Kingdom Department of Health's Hospital Quality Assessment Framework [8], Norway, Japan, and Taiwan's hospital quality assessment systems have all achieved good application results.

Healthcare quality management has specific requirements, and the best ones are those with standardized diseases, such as medical insurance diseases. In the medical healthcare data in this paper, these diseases are generally classified by the internationally accepted ICD-10 or ICD-9 [9], and the evaluation indicators are relatively uniform and standardized, including hospitalization days, average hospitalization costs, cure rate, etc.

From the above survey, we can see that healthcare quality evaluation is complicated and diverse, and there is not a completely universal healthcare quality assessment system [10]. It is an accepted method to propose a targeted healthcare quality assessment model or system for different data. In this paper, we propose a healthcare quality assessment model based on the outlier detection algorithm. In the literature, this type of approach has been proposed for different problems. Knorr et al. [11] proposed various distance-based outlier algorithms for k-dimensional datasets in real-world applications. The experimental results showed that the proposed algorithms provided the best results for $k \leq 4$. Petrovskiy [12] suggested an improved outlier detection algorithm based on fuzzy theory and kernel functions. The performance of the proposed algorithm has been tested on an intrusion detection system. Christy et al. [13] studied the problem of outlier reduction and proposed two different approaches: cluster-based and distance-based. The experimen-

tal results revealed that the cluster-based outlier detection approach provided better results. For detecting fraud in the Medicaid dental domain, van Capelleveen et al. [14] proposed an unsupervised outlier technique in order to detect fraudulent patterns at the post-payment stage. A comparative evaluation of outlier detection algorithms has been provided by Domingues et al. [15]. Jyothi et al. [16] proposed a statistical and distance-based outlier detection approach in healthcare claims and experimented the proposed approach on large-scale real-life data. Some state-of-the-art approaches have been benchmarked on real-world datasets from various domains. Based on big data features, Shao et al. [17] proposed an improved rapid density peak outlier detection algorithm. The improved proposed method successfully detected outliers, according to the experimental results.

3. Healthcare Quality Assessment Model

In order to establish a healthcare quality assessment model, we assign reasonable weights to two indicators, known as the outlier index (OI) and the excellent and good cases rate index ($ECRI$). For each hospital, the values for both indicators are sorted in descending order and we perform a quantile segmentation. The values are split into four equal-sized segments. Each 25% interval corresponds to a class, from A to D $(1, 2, 3, 4)$. The better the healthcare quality, the smaller the value of OI, and the larger the value of $ECRI$. Thus, the model score M calculation is carried out as shown in Equation (1):

$$M = a \times OI + b \times ECRI \tag{1}$$

where $a = \{1, 2, 3, 4\}$, $b = \{4, 3, 2, 1\}$ are the weights of the two indicators OI and $ECRI$, respectively.

Regarding the selection of a and b values, it should be that in the case excellent rate index, this paper analyzes three major categories of case analysis (Model), medical defect (Defect), and medical outcome (Trend), and contains a lot of attribute dimensions, which is a comprehensive evaluation method. So, the weight of $ECRI$ is larger, and the outlier index is mainly to analyze the medical defect in one of the three dimensions of the excellent and good rate of cases, so $a = 3 \times b$ is set by default. $a = 0.25$, $b = 0.75$.

A total of the result sets shown in Table 1 can be obtained. According to the calculation results, the model finally divides hospitals into first-level classification and second-level classification. According to different needs, hospitals with corresponding medical quality levels can be displayed.

Definition 1 (**Outlier index—OI**).

$$Outlier\ index = \frac{A}{B} \tag{2}$$

where A is the outlier percentage of a single hospital calculated based on the global statistical outlier algorithm, and B is the outlier percentage of a single hospital calculated based on the improved KNN outlier algorithm.

For example, in a hospital, A is 16%, and B is 7%, so the outlier index of this hospital is 2.3.

Definition 2 (**Excellent and good cases rate index—ECRI**). *The ECRI in each hospital can be calculated on the basis of the quality of each medical case as follows.*

$$p = \frac{S + G}{T} \times 100 \tag{3}$$

where p is the ECRI in the hospital, S and G are the number of excellent and good cases in the hospital, respectively, and T is the total number of cases.

Table 1. Classification table of model calculation results.

ECRI	OI	M	Primary Classification	Secondary Classification
4	4	4.00	A	a
	3	3.75		b
	2	3.50		c
	1	3.25	B	a
3	4	3.25		
	3	3.00		b
	2	2.75		c
	1	2.50	C	a
2	4	2.50		
	3	2.25		b
	2	2.00		c
	1	1.75	D	a
1	4	1.75		
	3	1.50		b
	2	1.25		c
	1	1.00		d

4. Outlier Detection Algorithm Based on Statistics

We found that various statistical methods are widely used in the medical field, and a complete departure from traditional statistical methods is unrealistic. Therefore, an outlier detection algorithm based on statistics is not ideal, but it is suitable for most of the numerical data in this paper. Therefore, this section briefly describes and applies traditional statistical methods.

Statistics-based outlier detection algorithms are usually divided into two parts:

1. The training phase based on an unsupervised method, which tries to build a statistical model that can contain the vast majority of data points, and the other is a semi-supervised method, which usually only estimates the probability density of outliers. This depends partly on the availability of class labels, and the other is a supervised approach, estimating the probability density of non-outliers or outliers.
2. The detection stage, which detects and judges whether the data points are outliers according to the model.

Statistics-based outlier detection algorithms are generally based on three models: the Gaussian model, Histogram-based and Regression model. Histogram is a parameter-free outlier detection method. The following mainly describes the algorithm based on the Gaussian model used in this paper.

This method is the most widely used statistical outlier detection method. By default, the detected data satisfy the Gaussian (Normal) distribution $N(\mu, \sigma^2)$, where μ is the mean of the data and σ is the standard deviation. These two parameters can be obtained by using the maximum likelihood estimation method. The outlier degree is based on the distance from the data value point to the average value. When the outlier degree is larger than the set threshold, it will be considered as an outlier. It can be seen that this is a probability-based outlier detection method, and small probabilities are considered outliers. For the distance between the most critical data point and the mean, the commonly used methods are the mean-variance test method and the boxplot method. The two are briefly introduced below, which are also the methods used in the statistical-based outlier detection in this paper.

4.1. Boxplot Method

The boxplot method is a method commonly used in the medical field. Generally speaking, it contains five statistics. These five statistics are the minimum value (min), the lower quartile (Q1), the median, the upper quartile (Q3), the maximum value (max), and on top of these five bases there is the Inter Quartile Range (IQR), which is the difference (IRQ $=$ Q3 $-$ R1) between the upper quartile and lower quartile. The corresponding steps are:

1. Take Q1, the lower quartile, as the lower end of the rectangular box, and the upper end of the rectangular box corresponds to Q3, the upper quartile, and draw the median line between Q1 and Q3.
2. Draw two vertical lines, parallel to the median line, at positions Q3 $+$ 1.5QR and Q1 $-$ 1.5QR, generally referred to as outlier demarcation lines.
3. Two lines are drawn at Q3 $-$ 3QR and Q1 $+$ 3QR positions. The data points outside these two boundary lines are considered extreme outliers, while points between these two lines and the inner limit are considered mild outliers.

4.2. Mean-Variance Test Method

The mean-variance test method is widely used in various fields, in particular in quality inspection. It simply treats points that are three standard deviations away from the average of the entire sample as outliers. In this way, the area of $\mu \pm 3\sigma$ has about 99.7% of the data points. Therefore, the remaining very small part of the data is regarded as outliers, which is simpler and straightforward.

However, the mean variance test method and the boxplot method are mostly the same. The Q1 $-$ 1.5QR and Q3 $+$ 1.5QR intervals in the boxplot method contain 99.7% of the data points in the dataset. There is no difference between the two in terms of results and probability distributions.

This paper's data is based on medical information from a Saudi Arabian region. Although the amount of data is relatively large and the data dimensions are relatively large, this paper uses methods to reduce dimensionality. The dimensionality reduction principle is to put the data points of the same hospital or the same hospital together and sort them according to the time dimension, try to ensure the consistency of the data points in the time dimension and the space dimension, and then use the divide-and-conquer method to carry out the medical data in this article many times. The statistics-based outlier detection algorithm performs global-based statistical outlier detection for dimension 1, dimension 2, and dimension 3, respectively. The outliers of each dimension can be combined with each other to obtain each dimension of single or multiple dimensions. The results of the proportion of outliers in the hospital are shown in Equation (4).

$$OP_1 = \frac{N_{\text{outlier}}}{N_{\text{all}}} \times 100 \tag{4}$$

where N_{outlier} is the number of outliers in each hospital, and N_{all} is the number of all data points for the hospital. In this paper, the outlier detection method based on the Gaussian model is adopted, specifically, the mean-variance detection method. As shown in the Figure 1, based on the actual situation, we select twice the standard deviation as the relevant parameter of the mean-contrast detection in this paper, and select twice the standard deviation. So, the total proportion of outliers is about 5%, which is more in line with the actual situation.

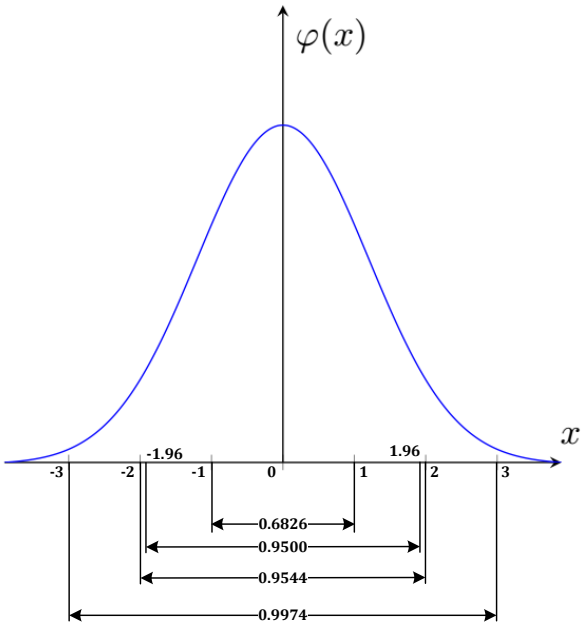

Figure 1. Normal Distribution Probability Plot.

5. Outlier Detection Algorithm Based on *k*-Nearest Neighbors (KNN)

This kind of algorithm is a classic outlier detection algorithm, which has also attracted much attention in the field of outlier detection in recent years. Whether it is an outlier is mainly evaluated by comparing it to the outlier degree of nearby neighbors.

Definition 3 (Distance-based outliers). *According to Knorr et al. [11], the definition of outliers is that if the distance between at least p part of the objects in the data point set T and the object O is greater than MinD, then the object O is called an outlier of* $DB(p, MinD)$.

The idea of the outlier detection algorithm based on KNN is defined as the outlier degree of the data p to be detected. It is the distance from the point p to its kth nearest neighbor, which is denoted here as $D^k(p)$*, where D is the dataset. First calculate the value* $D^k(p)$ *of each point in the dataset D. Then, do a quick-sort and select the top n points as the set of outliers. The algorithm itself does not need to set the parameter values of p and MinD, so the artificial influence is relatively small.*

However, this algorithm has an obvious drawback that it ignores the case of $k-1$ objects between the *k*-nearest neighbors of the data p to be detected. As shown in Figure 2, when $k = 7$ then D^k of points m and n are the same, but the outlier degrees of m and n are definitely different.

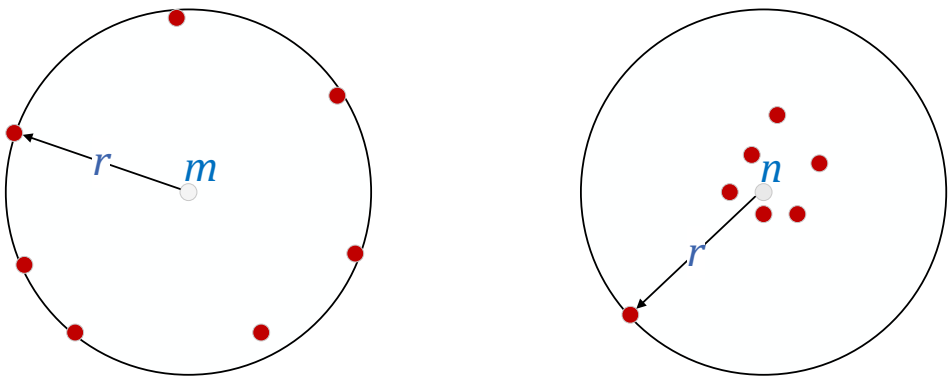

Figure 2. m and n have the same value, but different outlier scores.

Considering the closeness or sparseness of neighbors, we use the average distance as a measure of the outlier. In the following definition, the dataset is $D = \{d_1, d_2, d_3, \ldots, d_i, \ldots, d_N\}$, where N is the size of the dataset D, $d_i \in D$ is a data point in dataset D, counted as $(d_{i1}, d_{i2}, \ldots, d_{ij}, \ldots, d_{im})$. One of the data points represents a type, $A = A_1, A_2, A_3, \ldots, A_m$ represent the M dataset of A, M refers to the dimension of the dataset D, the attributes of the dataset are represented by A_j, distance(d_i, d_j) is the distance function between the data d_i and d_j.

Definition 4 (r-neighborhood). *Let r a positive integer, the data $p(p \in D)$'s r-neighborhood $N(p, r)$ is defined as follows:*

$$N(p, r) = \{q \in D | \boldsymbol{distance}(q, p) \leq r\} \tag{5}$$

Definition 5 (*k*-nearest neighbor distance). *Let k a positive integer, the k-nearest neighbor distance of the object k-distance(p) defined as* distance(p, o)*(where $o \in D, o \neq p$), satisfy:*

- $|N(p, \text{distance}(p, o))| \geq k$
- $|N(p, r)| \leq k - 1; \forall r <$ distance(p, o)

Definition 6 (*k*-nearest neighbor). *Let k a positive integer, the k-nearest neighbors of a data point p are formed by objects with distance $\leq k -$* distance(p) *to the data point p, and expressed as Equation (6).*

$$N_k(p) = \{q \in D, q \neq p | \text{distance}(p, q) \leq k - \text{distance}(q)\} \tag{6}$$

Definition 7 (Neighbor average distance). *Let D a dataset and k the number of neighbors, if $p \in D$, then the k-nearest neighbor distance of data point p can be calculated as follows.*

$$D^k(p) = \frac{\sum_{q_i \in N_k(p)} \text{distance}(p, q_i)}{|N_k(p)|} \tag{7}$$

Definition 8 ($DB(k, r)$ outliers). *If $|N(\mathrm{p}, \mathrm{r})| \leq k$, then the data point p is considered to be an outlier of $DB(k, r)$.*

Definition 9 (Top-n nearest neighbor outliers). *If $p \in D$, then the top n data points with the largest value $D^k(p)$ are the top-n outliers.*

The flow chart of the outlier detection procedure of KNN average distance is shown in Figure 3. The KNN algorithm has many advantages such as clear concept and easy implementation. As a non-parametric classification technology, it is a good supplement to the statistical-based outlier algorithm presented in Section 4. It can achieve good classification accuracy for unknown and non-normally distributed data, and detected outliers also have good local outlier significance.

To sum up, the advantages of the traditional KNN algorithm are mainly:

1. Simple and easy to implement;
2. No need to rely on other data;
3. Only related to *k* nearest neighbors, avoiding the imbalance problem caused by the amount of sample data.

The shortcomings are also obvious. The shortcomings of the traditional KNN method mainly include:

1. The value of *k* is difficult to determine;
2. Processing is slow;
3. Very dependent on training data.

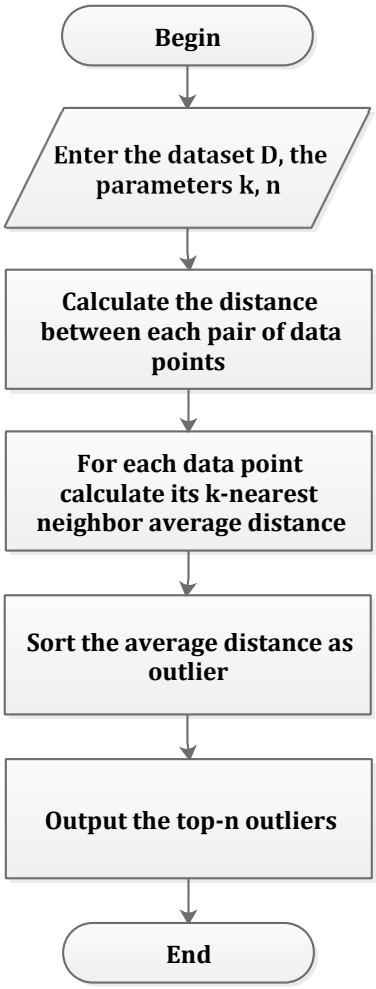

Figure 3. Outlier detection algorithm based on KNN.

6. An Improved KNN-Based Outlier Detection Algorithm

In this section, we propose some improvements to the basic outlier detection algorithm based on KNN (presented in Section 5). We use the statistics-based outlier detection algorithm to detect, to ensure that the comparison conditions based on statistics and KNN are in the same dimension. The goal of the algorithm is to output the top-n outliers of the dataset D.

The main improvements are as follows. A new parameter m is introduced to greatly reduce the computation amount of the distance between two points. Firstly, the improved algorithm analyzes the dataset and performs a pre-judgment pruning. Then, the remaining sub-datasets are sorted and classified. Finally, the corresponding pruning operation is performed on the classification results to achieve the purpose of reducing the time complexity of the algorithm.

Definition 10 ($DB(m,k,r)$). *In the dataset D, select m points nearby, and k is the number of neighbors, r is the threshold from the object O to the kth neighbor, if the distance is greater than r, then The object O is a candidate outlier of $DB(m,k,r)$. The improvement of the algorithm is mainly in two aspects. One is to set an area range for pairwise distance calculation according to Equation (8).*

$$m = \left\lfloor \frac{N_{all}}{H_{\text{hospital}}} \right\rfloor \tag{8}$$

where N_{all} and H_{hospital} are the number of all data points and the number of all hospitals in the dataset D, respectively. Therfore, m represents the average number of data points in the hospital in

the entire dataset. The practical significance of this is that the number of data points to be compared should be as many as possible within the scope of the hospital.

Another improvement is to prune the dataset, which mainly includes three steps, namely pre-judgment pruning, sub-dataset sorting, and reducing the search for k-nearest neighbors through pruning conditions. The specific steps are as follows.

(1) Pre-judgment pruning: Check and classify the input dataset D_i according to some conditions, and divide it into many sub-classes. There must be some outliers that do not exist in the classification result set, and they will be clipped so that subsequent operations on this dataset will not be performed to reduce the size of the initial dataset. The specific idea is: select an initial point, and then continue to review other points in the dataset near this point to determine whether the distance from the entire point to the center point is greater than $r/2$, if not, then merge this point into this cluster. Otherwise, this point is used as the center of the new dataset, and a count binary array is set at the same time to record the number of data points in each cluster and the corresponding point. After all the data points are scanned, the sub-datasets larger than k are removed, and those smaller than k are left.

Proof. Take $r/2$ as the radius of the circle and o as the center of the circle, and let o be the class's midpoint data point and the number of data points in the class be at least k, here is assumed to be p and q, then Equations (9) and (10) are satisfied:

$$\text{distance}(p, q) \leq r \tag{9}$$

$$|N(p, r)| \geq k \tag{10}$$

Therefore, p is not an outlier, i.e., there are no outliers in this subdataset. □

(2) Sort sub-datasets: In the first step of pre-judgment pruning, the dataset is divided into $D = \{D_1, D_2, D_3, \ldots, D_n\}$, where n is the number of datasets left in the previous step. Here, the center point values of the sub-datasets are sorted, and divided into four clusters according to the quartiles. The following are the definitions of some specific points. $|C_i|$ represents the total number of data points in the class, C_iR indicates the radius of the corresponding dataset, that is, the distance from the center point of the cluster to the farthest point. Finally, we perform an ascending sort according $Density_i$, where the density of the cluster $Density_i = |C_i|/C_iR$. Density is a good indicator for the spareness of a dataset.

(3) Calculate outlier: The KNN algorithm based on nested loop detects outliers in data. To calculate the distance between this data point and all n data points, it is necessary to calculate the k-nearest neighbors of each data point, so the time complexity reaches $O(n^2)$. In our work, we only calculate the distance with m points near it, in order to reduce the distance calculation between data points. The calculation is performed by the following two clipping conditions.

Definition 11 (Outlier dynamic threshold cutvalue). *This parameter is set to a dynamically variable threshold (0 by default). When the number of outlier candidate sets is greater than top-n, then the threshold is set to the minimum value among the n candidate outliers.*

Clipping condition 1: When $D^x(p) \leq$ cutvalue, the search for the k-nearest neighbors of this point does not need to be carried out, and this point must not be an outlier. $D^x(p)$is the average distance value of the k-nearest neighbors that the data point p has found, and the cutvalue is the threshold for judging whether it is an outlier.

Proof. Suppose that when searching for its k-nearest neighbors for a data point p, the point with the farthest distance will be replaced by the point with the closest distance, and the outlier degree of the average distance of k-nearest neighbors will continue to decrease as the search progresses. If the average distance of the found k-nearest neighbors of p is

already less than or equal to cutvalue, then the k-nearest neighbors of this point p will become smaller and smaller. So, its outlier degree will only be smaller than cutvalue. In other words, if $x(x \geq k)$ distance judgment, and the outlier degree at this time is already smaller than the cutvalue threshold, then there is no need to calculate the distance between p and other points in the dataset. □

Clipping condition 2: When the minimum value among n potential outliers is assigned to cutvalue, if $D^k(q) + \text{distance}(p,q) < \text{cutvalue}$, then p cannot be an outlier; where p is an unknown data point, q is an already calculated data point, $D^k(q)$ represents the k-nearest neighbor average distance of point q, $\text{distance}(p,q)$ is the distance between p and q, and cutvalue is the threshold for judging whether it is an outlier.

Proof. Let p and q two points and the three nearest neighbors of q form three triangles. Suppose $k = 3$, a, b, and c are three nearest neighbors of q. According to the triangle inequality, the following results can be obtained:

$$\text{distance}(q,a) + \text{distance}(p,q) > \text{distance}(p,a) \tag{11}$$

$$\text{distance}(q,b) + \text{distance}(p,q) > \text{distance}(p,b) \tag{12}$$

$$\text{distance}(q,c) + \text{distance}(p,q) > \text{distance}(p,c) \tag{13}$$

When we sum Equations (11)–(13), we obtain the following equation $\forall a_i \in N(q,3)$.

$$\sum_{i=1}^{3} \text{distance}(q,a_i) + \text{distance}(p,q) > \sum_{i=1}^{3} \text{distance}(p,a_i) \tag{14}$$

Extending Equation (14) to k, we obtain Equation (15) $\forall a_i \in N(q,k)$:

$$\sum_{i=1}^{k} \text{distance}(q,a_i) + k \times \text{distance}(p,q) > \sum_{i=1}^{k} \text{distance}(p,a_i) \tag{15}$$

Divide the left and right sides of Equation (14) by k to obtain Equation (16):

$$\frac{1}{k}\sum_{i=1}^{k} \text{distance}(q,a_i) + \text{distance}(p,q) > \frac{1}{k}\sum_{i=1}^{k} \text{distance}(p,a_i) \tag{16}$$

In addition, due to Equation (17):

$$D^k(q) = \frac{1}{k}\sum_{i=1}^{k} \text{distance}(q,a_i); \forall a_i \in N(q,k) \tag{17}$$

Then we can obtain Equation (18), $\forall a_i \in N(q,k)$:

$$D^a(q) + \text{distance}(p,q) > \frac{1}{k}\sum_{i=1}^{k} \text{distance}(q,a_i) \tag{18}$$

The nearest neighbors of p and q are generally different, so for data point p, the calculation is shown in Equation (19):

$$\frac{1}{k}\sum_{i=1}^{k} \text{distance}(q,a_i) > \frac{1}{k}\sum_{i=1}^{k} \text{distance}(p,a_i) \tag{19}$$

Therefore, the following equation can be obtained:

$$D^k(q) + \text{distance}(p,q) > D^k(p) \tag{20}$$

In addition, from Equation (20), it can be known that :

$$D^k(q) + \text{distance}(p, q) < \text{cutvalue} \tag{21}$$

Then we can obtain the following conclusion:

$$D^k(p) < \text{cutvalue} \tag{22}$$

It can be seen from the above evidence that D^k of the candidate outlier is already greater than $D^k(q)$, and the search for top-n does not need to be performed anymore. Therefore, when $D^k(q) + \text{distance}(p, q) < \text{cutvalue}$, there is no need to search the k-nearest neighbors of p. □

6.1. Algorithm Steps

The selection of m and r values are already presented in Definitions 10 and 11. In the following, we briefly describe the k value selection. The principle of k value is generally to obtain the best choice through a large amount of test training data. We tried different k values, based on the healthcare data of thousands of known outlier results obtained for a certain disease in a particular region. As shown in Table 2, k represents the number of neighbors, and t is the strictness of judging whether outliers are really outliers.

Table 2. The number (n) and proportion (%) of outliers under different k and t combinations.

k	$t = 1$		$t = 2$		$t = 3$		$t = 4$		$t = 5$	
	n	%	n	%	n	%	n	%	n	%
$k = 9$	102	4.22%	124	5.13%	147	6.08%	178	7.36%	221	9.14%
$k = 13$	48	1.99%	56	2.32%	65	3.69%	71	4.44%	88	5.24%
$k = 17$	33	1.37%	39	1.61%	48	2.99%	61	3.52%	78	4.73%
$k = 21$	16	0.66%	23	0.95%	34	2.11%	55	2.68%	59	3.44%

From Table 2, we remark that when k does not change, the number of outliers increases with the increase in t, and when t does not change, the number of outliers decreases with the increase in k. It shows that the stricter the conditions, the more difficult it is to establish a connection, and the more outliers there are. The more the number of k-nearest neighbors, the higher the fault tolerance rate and the greater the possibility of establishing a connection. For this reason, the entire algorithm has fewer outliers.

In this paper, we consider the average daily expenses, days of hospitalization, and drug proportions. These three dimensions have a significant impact on the evaluation of healthcare quality.

According to Yu et al. [18], when k increases the computation time increases exponentially. Therefore, the selection of k should be as small as possible, and try not to exceed 20. As shown in Table 3, it is one of the training results. The total number of outliers in the training sample is 151. When $k = 17$, it is optimal to comprehensively consider the proportion of outliers, the number of outliers, the accuracy rate and the false alarm rate. An outlier is defined as a point with at least two outliers in the three dimensions.

Table 3. Outlier detection analysis for different k values.

k	Outlier Detection		Accuracy	False Alarm Rate
	n	%		
$k = 11$	62	2.10%	32%	0.51%
$k = 12$	71	2.41%	37%	0.54%
...	...	...	...	...
$k = 17$	142	4.81%	84%	0.88%
$k = 20$	220	5.57%	87%	3.20%

6.2. Algorithm Implementation

The specific steps of the improved KNN algorithm are described in detail in Algorithms 1–3. The improved KNN algorithm is mainly divided into three parts: pre-judgment and pruning (Algorithm 1), sorting sub-datasets (Algorithm 2), and calculating outliers (Algorithm 3).

Algorithm 1 Pre-judgment and pruning

Input: dataset D, the distance value m, the number of neighbors k, and the distance threshold r
Output: pruned dataset $D1$

$D1 = D$
Select the first point p in $D1$ and put it into the cluster center point set O, O_i represents the subset of data centered at point p
for each data point q in $D1$ **do**
  Calculate the distance from q to the midpoint of O
  if the distance $> r/2$ **then**
    Put q into the new center point set O
  else
    Put q into the cluster set O_i
  end if
end for
for each data point q in $D1$ **do**
  Classify and divide q into several small clusters
  if the number of data points in the cluster is $> k$ **then**
    Cut it out
  end if
end for
return the pruned dataset $D1$

Algorithm 2 Sub-datasets sorting

Input: $D1 = \{D_1, D_2, D_3, \ldots, D_n\}$ generated by Algorithm 1
Output: Sorted sub-dataset $C = \{C_1, C_2, C_3, C_4\}$

Sort the data in $D1$ by the center point value
Take the quartiles to divide the entire data subset into a class
$a = 4$, *because the number of data points in each subset is not uncertain, so the number of data points varies widely.*
Calculate $Density_i = |C_i|/C_iR$ for each class C_i
Sort the sub-datasets in ascending order according to Density
return The sorted sub-dataset C

As shown in Figure 4, assuming X and Y are the two classes after the initial classification, the density of the two categories is obvious. If you select X with a smaller density to calculate first, you can quickly increase the threshold. When the calculation of the Y class is performed again, the search of k-nearest neighbors can be terminated in advance according to the pruning conditions in Algorithm 3. Therefore, unnecessary distance calculations and time complexity are reduced.

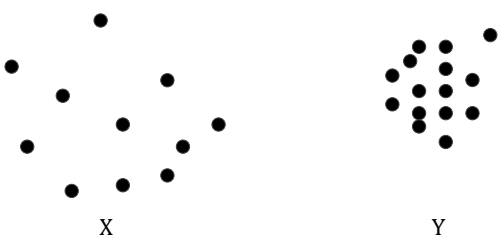

Figure 4. Example of two classes X and Y with different densities.

Algorithm 3 Outlier degree computation

Input: sub-dataset C, the number of distances to be calculated m, the nearest neighbor data k, the number of outliers n
Output: the set of outliers O

for each C_i in C **do**
  for each data point p in C_i **do**
    p's nearest neighbors $Neig(p)$ are empty
    for each data point p in C_i; $p \neq q$ **do**
      if $|Neig(p)| < k$ OR distance$(q, p) <$ Maxdistance$(p, Neig(p))$ **then**
        Update the set of neighbors of p
        // Maxdistance(p, D) *returns the maximum distance between p and objects in D*
      end if
      if $|O| < n$ **then**
        Put p in the outlier set O
      end if
      if $|O| = n$ **then**
        Update the outlier detection set O
        Update the pruning threshold:
        cutvalue $=$ MinThreshold(O)
        //MinThreshold(O) returns the minimum value of outliers in D.
      end if
    end for
  end for
end for
return the set of outliers O

6.3. Algorithm Analysis

In the design of the healthcare quality assessment model based on big data, multi-node calculation processing is carried out based on the Hadoop platform. Since only m data points to be detected are compared, the datasets of the improved KNN algorithm are relatively independent. Therefore, the outlier detection of the dataset can be placed on different nodes. This paper only needs to summarize the detection results of each outlier.

Figure 5 shows the multi-node outlier detection processing flow. Under the premise that the temporal and spatial attributes of adjacent data points are similar, after running on multiple running nodes, the outliers of each dimension are judged on a single node, and then aggregated, which also significantly speeds up the processing speed.

This section mainly analyzes the time complexity of the proposed approach. The algorithm is divided into three parts:

1. Pre-judgment pruning: in order to reduce the data set scale.
2. Sort the dataset by density.
3. Calculate the outlier degree for each point in the dataset, and output the outlier result set.

Let N is the total number of data points contained in the dataset, and D is the dimension of the dataset. The first part of the algorithm only scans the dataset twice, so the time complexity is $O(ND)$. The second part adopts the quick-sort algorithm with the best sorting performance. After calculating the average value of the sub-dataset, the time complexity of the sorting algorithm here is $O(K \log k)$, where K represents the number of result subsets after the first part of the pruning. In the third part, when the sorted dataset in the second part is relatively uniform, the time complexity is still $O(N^2D)$, which means that no pruning operation is performed. However, for the case where the data density distribution after sorting is very different, with the rapid determination of the outlier threshold cutvalue, the corresponding pruning conditions can be applied. Such scans are relatively few, and the time complexity of the third part will be much lower than $O(N^2D)$.

In summary, the worst time complexity of the entire algorithm is still $O(N^2D)$, but in the actual experimental test, the average time complexity performance will be greatly improved.

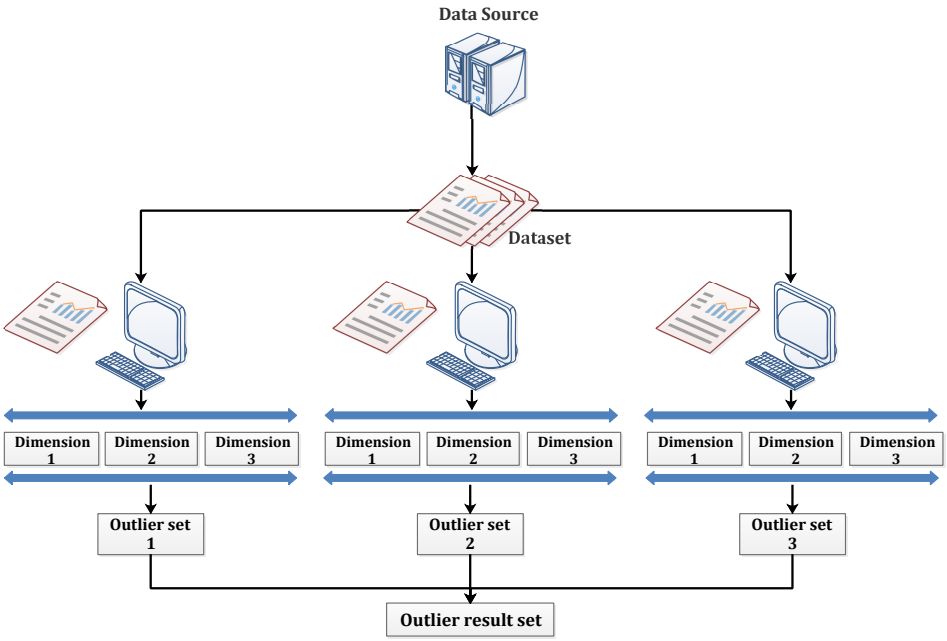

Figure 5. Multi-node outlier detection processing flow chart.

7. Experimental Results

The experimental conditions of this paper are based on 3 CPU i5, 1 TB hard disk, 8 GB RAM. The traditional and the improved KNN algorithms are compared in terms of correct rate, false positive rate, false negative rate and operation efficiency, and finally the availability of the final outlier indicator is verified.

7.1. Data Preparation

Data preparation refers to a series of processes before the data enters the model, as shown in Figure 6. The processing process in this paper includes three processes: data integration, data noise cleaning, and data preprocessing. The three processes of data preparation are briefly described below.

7.1.1. Data Integration

The most important data source in this paper is the medical data of a region. The medical data of the past five years are integrated with some other data sources, in an Oracle database. The unified management in the storage module of the big data healthcare quality system in this paper greatly facilitates the management and use of data. This step is the basis of all the following parts.

7.1.2. Data Noise Cleaning

Although each piece of data used in this paper comes from real medical records, it is well known that for medical data, especially for data collected from different hospitals and medical service points. There must be a lot of "dirty data" such as conflicting letters, missing fields, and abnormal inputs. We strictly handle these noises before entering the model.

The data cleaning step, although the technical content is not high, is more about the understanding and analysis of the business and the judgment of common sense. However, it has a very important role. During the research process of this paper, understanding the relevant knowledge of medicine has spent a large part of the time of the whole research.

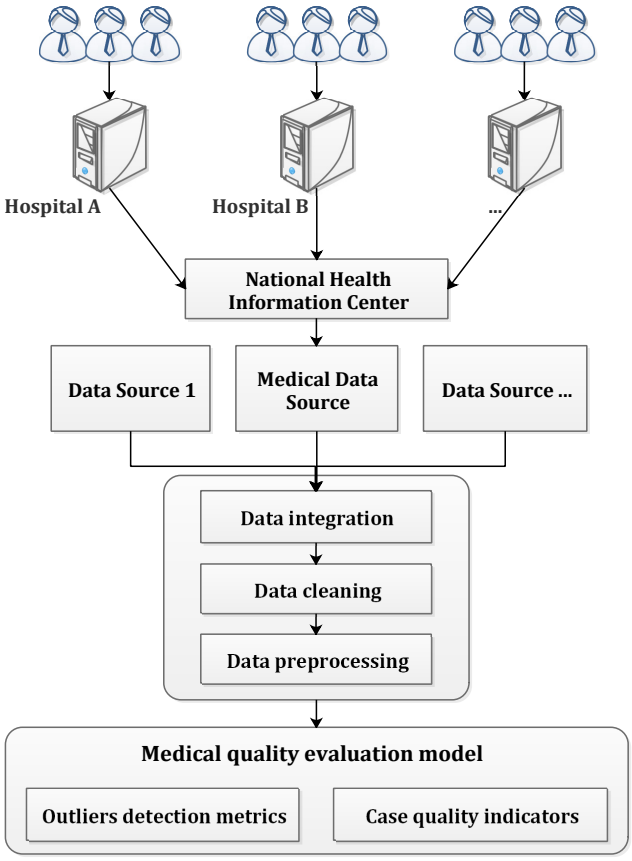

Figure 6. Data preparation schematic.

7.1.3. Data Preprocessing

After data integration and data cleaning, the accuracy is guaranteed and a real and effective dataset is obtained, but this part of the data is basically in the format of the data source, which is far from the data format required by the model. The most important goal of the preprocessing process is to organize the data source into the data format required by our healthcare quality assessment model.

7.2. Experimental Data

The majority of the experimental data come from medical data of a specific region in Saudi Arabia from 2013 to 2018, with a total of millions of medical records, and a TB-level data volume, which basically meets the requirements of big data in terms of data volume. The medical data is fairly extensive, and the parameters for analysis are reasonably precise. In this paper, we selected data related to two diseases A and B. We evaluated the healthcare quality of the hospitals in a region that offer the treatment of these diseases. In the previous section, we performed a large number of processing operations on the data to make the data format meets the input of the outlier detection algorithm. The main data used can be described as follows:

- VN: the visit number.
- Treatment hospital number: the ID number of the hospital.
- Treatment hospital name: the name of the hospital.
- Treatment hospital level: the hospital level according to the Saudi Arabia hospitals classification.
- Total expenses (in SAR): the value of the amounts spent on this visit.
- Drug fee (in SAR): the total cost of the medication in this visit.
- Admission date: the date of admission of the patient.
- Discharge date: the date of discharge of the patient.

- Length of hospitalization: it is a calculated field, which is equal to the difference between admission and discharge dates.
- Average daily hospitalization cost (in SAR): it is a calculated field, which is equal to the ratio of the total expenses to the length of stay.

As shown in Table 4, some examples of the test data are shown. Because it involves a lot of personal privacy, a part of the content containing obvious personal information has been processed. It can be seen that its dimensions are relatively rich. During the project process, the experimental data were also encrypted with the MD5 algorithm.

Table 4. Example of test data.

VN	Hospital ID	Hospital Name	Hospital Level	Total Expenses	Drug Fee	Admission Date	Discharge Date	Length of Hospit.	Avg. Daily Hospit. Costs
001	01	A	0101	9743.25	1823.09	21/6/2017	1/7/2017	10	974.33
002	02	B	0101	27,008.58	13,479.3	5/7/2016	30/7/2016	25	1080.34
003	01	A	0101	11,103.84	3879.94	21/6/2016	21/6/2016	12	925.32
004	03	C	0101	11,456.93	3905.71	1/11/2015	8/11/2015	7	1636.70
005	01	A	0101	10,042.20	2159.3	24/1/2017	1/2/2017	8	1255.28
006	04	D	0101	10,701.37	3984.56	6/5/2015	17/5/2015	11	972.85
007	05	E	0101	9094.31	2320.84	16/8/2017	26/8/2017	10	909.43
008	06	F	0101	10,213.69	3106.09	18/9/2016	26/9/2016	8	1276.71
009	01	A	0101	10,259.82	2925.72	16/4/2016	24/4/2016	8	1282.48
010	01	A	0101	10,856.24	3225.48	10/9/2017	18/9/2017	8	1357.03
...	...	...	...	...	...	...	...	...	...

The data hospital level, hospital name, and admission date are sorted in turn, which basically ensures that the time and space dimensions between adjacent points remain similar, as shown in Figure 7.

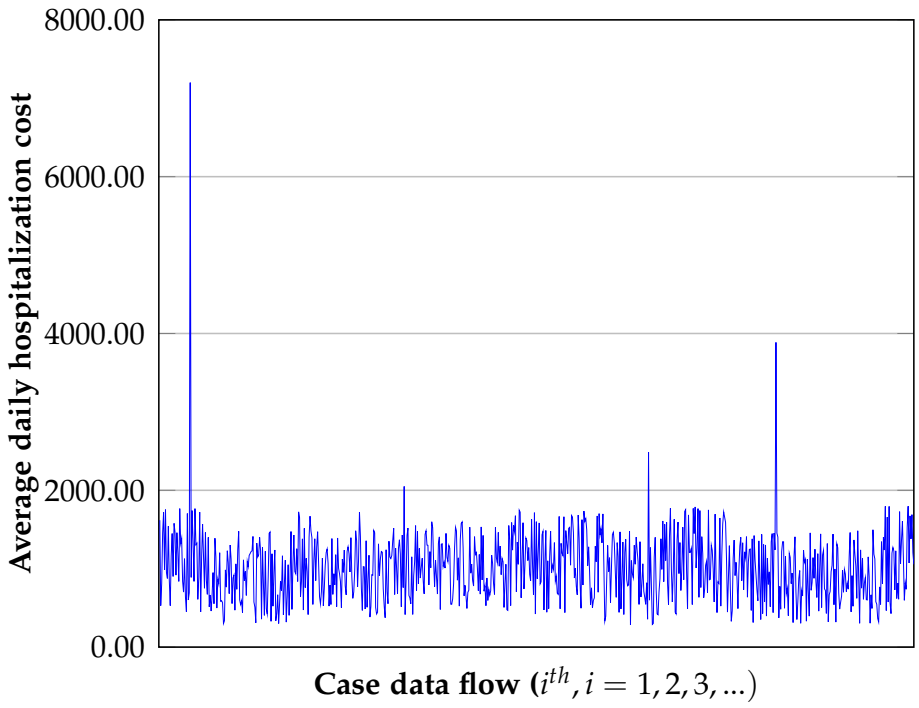

Figure 7. Data flow diagram of a dimension of data.

7.3. Analysis of Results

Due to the traditional status and significance of statistics in the medical field, it is difficult for all parties to accept the outlier detection algorithm completely separated from

statistics. As a result, this paper's outlier detection algorithm for the healthcare quality assessment model is subdivided into two parts. The outlier detection approach based on statistics is one component, while the improved KNN outlier detection algorithm is the other. According to the outlier detection results of the two, an outlier index based on hospital units is obtained, and the data outliers are obtained from the data.

From the perspective of judging the quality of healthcare in hospitals, the total number of outlier sets of these two types of outlier detection algorithms in this paper is set to be the same. In this way, it is possible to compare the similarities and differences between the proportion of outliers in each hospital based on statistics and the proportion of outliers in hospitals based on KNN when the number of outliers is uniform.

7.3.1. Accuracy Analysis

For accuracy analysis, two related indicators are generally used, one is the false positive rate or false alarm rate FAR (False Alarm Rate), and the other is the correctness CR (Correct Rate). The false negative rate or reporting rate OR (Omission Rate) is a commonly used indicator.

The false positive rate FAR is defined as the ratio of the set of data points X_{false} that are actually non-outliers but falsely reported as outliers to the X_{true} of the set of non-outlier data points in the dataset, that is $\text{FAR} = X_{\text{false}} / X_{\text{true}}$.

The definition of correct rate CR refers to the ratio of the number of outlier records X_{total} detected correctly by the algorithm to the total number of outliers in the dataset, i.e., $\text{CR} = X_{\text{correct}} / X_{\text{total}}$.

The false negative rate OR is defined as the percentage of the number of data points that are actually outliers but not detected, to the total number of outliers in the dataset, that is, $\text{OR} = X_{\text{false}} / X_{\text{total}}$. From another perspective, $\text{OR} = 1 - \text{CR}$.

The accuracy analysis and comparison in this paper is based on the analysis and research of two cases, case A and case B. Among them, the amount of data in case A is relatively small, with thousands of cases (4431 cases). This part of the data has been reviewed by experts in a certain hospital. Combining medical-related knowledge, a strict judgment is made on whether the cases are outliers. Therefore, the information such as outliers of disease A are known, and the accuracy of disease A is mainly analyzed. Disease B has a large amount of data, with hundreds of thousands of cases. It is considered to be a data set of unknown outliers, and the operation efficiency is analyzed.

As shown in Figures 8 and 9, top-n is the specified number of outliers. With the increase in n, the fault tolerance rate of the algorithm is improved, so that the accuracy can be rapidly improved. Overall, the improved KNN algorithm is slightly better than the traditional KNN algorithm in terms of accuracy. There is no obvious change before top-n = 120, but after top-n is greater than 120, the improved KNN algorithm performs better than the traditional KNN algorithm in terms of accuracy and false positive rate.

In this paper, because the statistical-based outlier detection method and the neighbor-based outlier detection method are to be divided and compared, the number of outlier sets of the two is the same for comparison. From Figures 8 and 9, it can be seen that when top-n is 200, it is more accurate when it accounts for about 5% of the total data volume N. Therefore, the method based on statistics here selects twice the standard deviation, that is, when the outlier probability is 4.56%, we compare the outlier detection algorithms based on statistics, based on traditional KNN, and based on improved KNN. From Figure 10, it can be seen that in general, the improved KNN algorithm performs better in accuracy analysis under the same conditions.

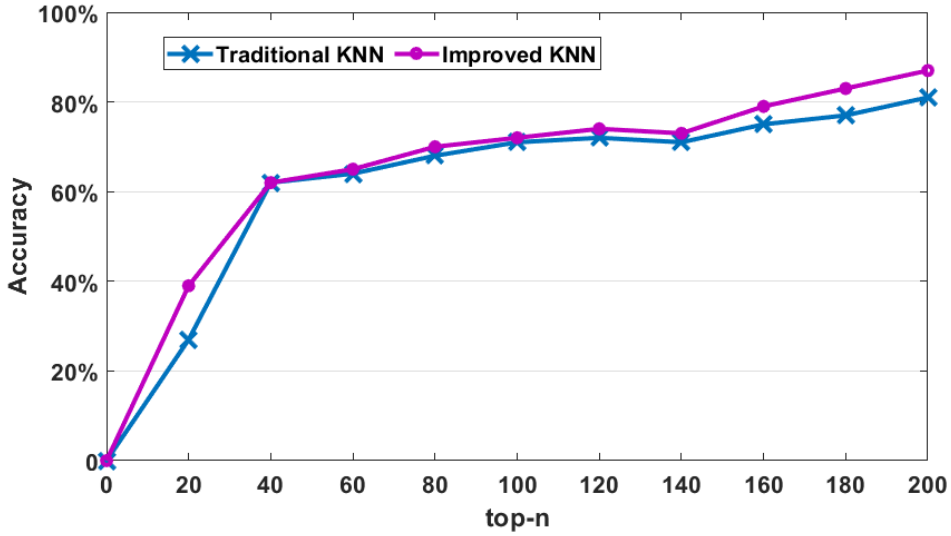

Figure 8. Accuracy rate variation.

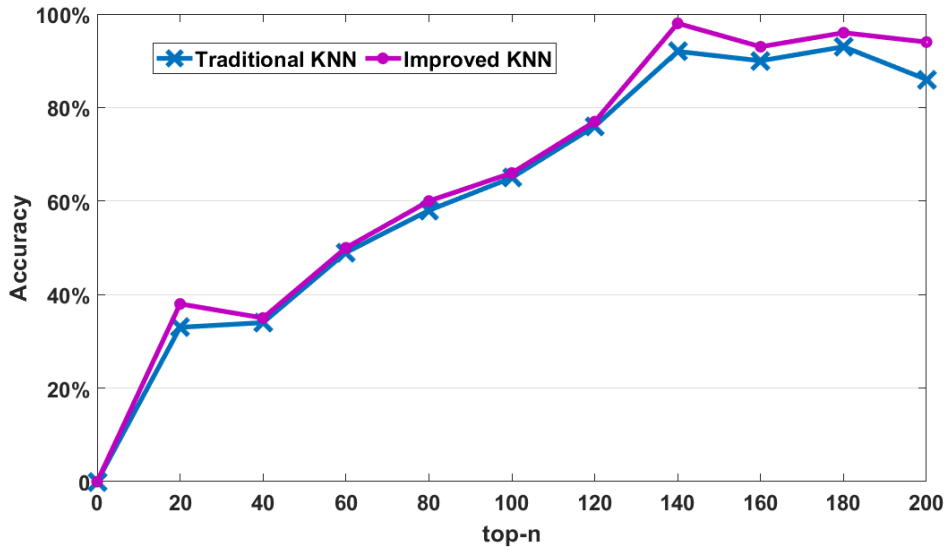

Figure 9. False alarm rate variation.

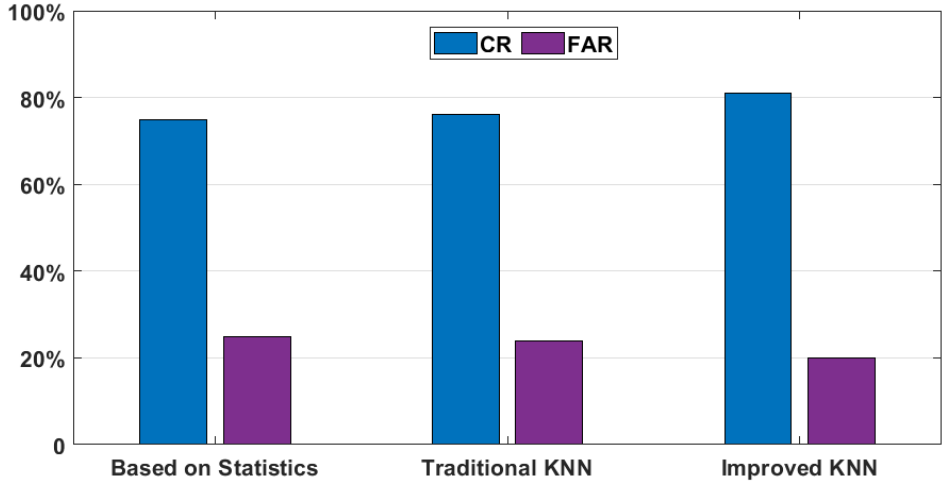

Figure 10. Comparison of accuracy rate and false negative rate of disease A.

7.3.2. Operational Efficiency Analysis

The following is mainly to analyze the operation efficiency of disease B with a large amount of data.

As shown in Figure 11, for disease B, the amount of data is relatively small. In this paper, 20,000 pieces of data are used as the incremental value of the amount of data on the abscissa. Here, $k = 17$ obtained from training, $n = 4.56\% \times N$ (where N is the total number of data points), and the traditional KNN-based and improved KNN-based outlier detection algorithms are run on the data set. It can be seen that with the increase in the amount of data, the improved KNN algorithm performs well in terms of operating efficiency. With the increase of the amount of data, the improvement is more obvious.

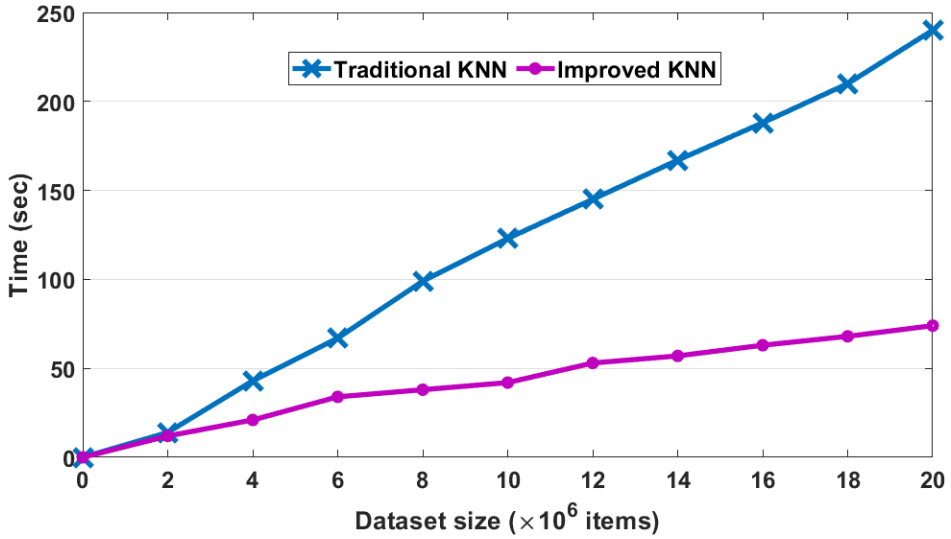

Figure 11. Time performance for disease B.

7.3.3. Outlier Analysis

Since the statistical-based outlier detection algorithm is based on the global scope of detection, it generally causes a large proportion of outliers in a certain part (hospital). Of course, this also has practical significance. As shown in Figure 12, this paper briefly illustrates the proportion of outlier results used in outlier detection. Among them, the proportion of outlier results ODR (Outlier Detection Rate) $= X_{\text{last}}/X_{\text{all}}$, in X_{last} is the number of outliers detected by the algorithm in the hospital, X_{all} is the number of all data in the hospital. It can be seen that for the top-10 with the highest proportion selected by the global statistical outlier detection algorithm, the proportion of outliers based on neighbors in each area (hospital) are relatively average, which can better reflect the A concept of regional outliers.

The hospital ranking determined by the outlier index can reflect the ranking of a hospital's outlier rate among all hospitals in a practical sense. Practice has proved that the top-13 hospitals in the outlier index selected in this paper are the same as the National Health Information Center in the ranking of all hospitals. After strict medical analysis of 13 hospitals with management loopholes and fraud, the disease has been successfully hit in 7 of them, which is of great significance in practical application.

This paper conducts experiments on both the statistics-based outlier detection algorithm and the improved KNN-based outlier detection algorithm. It is of practical significance to use the ratio of the hospital outliers of the two as the outlier indicator. The accuracy and running time of the improved KNN algorithm and the traditional KNN algorithm are compared. It can be seen that on the basis of a slight improvement in accuracy, the running time is greatly reduced, and the time complexity is significantly reduced.

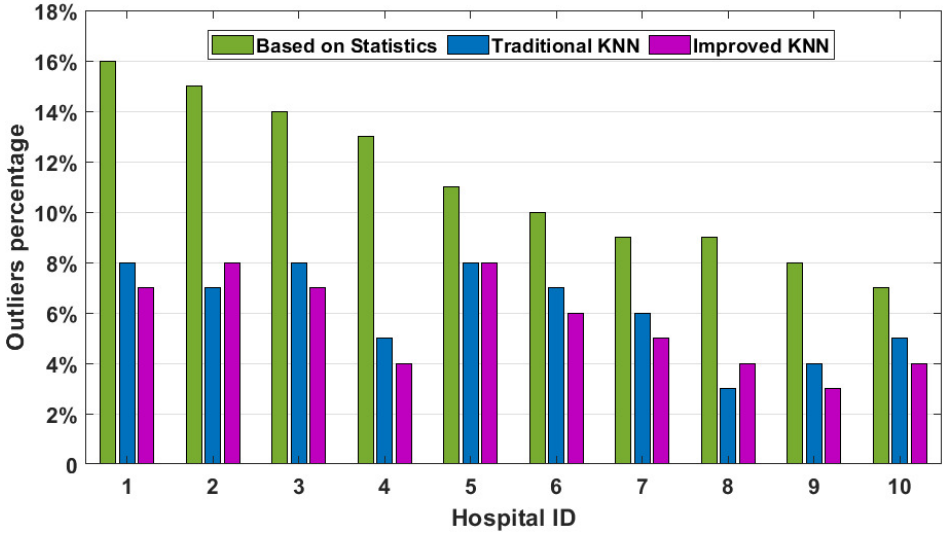

Figure 12. The proportion of abnormal points in disease A for the top-10 hospitals.

8. Conclusions

For the establishment of outlier indicators, healthcare data is the real data of each item. After noise processing, the availability of the data is very high and the outliers themselves contain a lot of useful information. Outlier detection based on these data can explain the healthcare quality of the hospital from the perspective of data outliers. In this paper, the statistical outlier detection algorithm based on the Gaussian model and the ratio of the outlier proportion of each hospital obtained by the improved KNN algorithm are used as the outlier index in this work. Experimental results have proven that the proposed evaluation method can reflect a certain degree of healthcare quality, especially for the detection of hospitals with fraudulent finance and medical loopholes.

The outlier indexes in the proposed model have practical significance, but there are still some shortcomings and areas for improvement, mainly in the following aspects:

- The application of data mining ideas to the medical industry still needs a long way to go, and the non-interoperability of data mining knowledge and medical knowledge has seriously hindered the development of medical big data.
- The relatively high rate of missing fields has a relatively influence on the evaluation results and slow down the data preprocessing stage.
- Future work concerns integrating more data in order to improve the accuracy and the usability of the evaluation model. These data can be medical safety and quality data, including surveys of staff and patients, and data related to hospital systems and procedures.

Author Contributions: N.A. carried out the data collection, participated in the proposed solution and drafted the manuscript. M.A.R. conceived the proposed algorithm, and participated in its design and helped to draft the manuscript. A.A. participated in the design of the study and performed the performance analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used and analyzed during the current study are available from the corresponding author on reasonable request.

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

Abbreviations

The following abbreviations are used in this manuscript:

ODR	Outlier Detection Rate
KNN	*k*-Nearest Neighbors
FAR	False Alarm Rate
CR	Correct Rate
IQR	Inter Quartile Range
QR	Quartile Range
ECRI	Excellent Cases Rate Indicator
OR	Outlier Index
BNQP	Baldrige National Quality Program
HCQI	Health Care Quality Indicator Project
PATH	Performance Assessment Tool for Quality Improvement in Hospitals

References

1. Chen, M.; Mao, S.; Liu, Y. Big data: A survey. *Mob. Netw. Appl.* **2014**, *19*, 171–209. [CrossRef]
2. Murdoch, T.B.; Detsky, A.S. The inevitable application of big data to health care. *J. Am. Med. Assoc.* **2013**, *309*, 1351–1352. [CrossRef] [PubMed]
3. Merelli, I.; Pérez-Sánchez, H.; Gesing, S.; D'Agostino, D. Managing, Analysing, and Integrating Big Data in Medical Bioinformatics: Open Problems and Future Perspectives. *BioMed Res. Int.* **2014**, *2014*, 134023. [CrossRef] [PubMed]
4. Raghupathi, W.; Raghupathi, V. Big data analytics in healthcare: Promise and potential. *Health Inf. Sci. Syst.* **2014**, *2*, 3. [CrossRef] [PubMed]
5. Carinci, F.; Gool, K.V.; Mainz, J.; Veillard, J.; Pichora, E.C.; Januel, J.M.; Arispe, I.; Kim, S.M.; Klazinga, N.S. Towards actionable international comparisons of health system performance: Expert revision of the OECD framework and quality indicators. *Int. J. Qual. Health Care* **2015**, *27*, 137–146. [CrossRef] [PubMed]
6. Owens, D.K.; Lohr, K.N.; Atkins, D.; Treadwell, J.R.; Reston, J.T.; Bass, E.B.; Chang, S.; Helfand, M. AHRQ Series Paper 5: Grading the strength of a body of evidence when comparing medical interventions-Agency for Healthcare Research and Quality and the Effective Health-Care Program. *J. Clin. Epidemiol.* **2010**, *63*, 513–523. [CrossRef] [PubMed]
7. Phillips, D.M. JCAHO Pain Management Standards Are Unveiled. *J. Am. Med. Assoc.* **2000**, *284*, 428. [CrossRef] [PubMed]
8. Thomson, R.; Taber, S.; Lally, J.; Kazandjian, V. UK Quality Indicator Project® (UK QIP) and the UK independent health care sector: A new development. *Int. J. Qual. Health Care* **2004**, *16*, i51–i56. [CrossRef] [PubMed]
9. Quan, H.; Sundararajan, V.; Halfon, P.; Fong, A.; Burnand, B.; Luthi, J.C.; Saunders, L.D.; Beck, C.A.; Feasby, T.E.; Ghali, W.A. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. *Med. Care* **2005**, *43*, 1130–1139. [CrossRef] [PubMed]
10. Hillestad, R.; Bigelow, J.; Bower, A.; Girosi, F.; Meili, R.; Scoville, R.; Taylor, R. Can electronic medical record systems transform health care? Potential health benefits, savings, and costs. *Health Aff.* **2005**, *24*, 1103–1117. [CrossRef] [PubMed]
11. Knorr, E.M.; Ng, R.T.; Tucakov, V. Distance-based outliers: Algorithms and applications. *VLDB J.* **2000**, *8*, 237–253. [CrossRef]
12. Petrovskiy, M.I. Outlier detection algorithms in data mining systems. *Program. Comput. Softw.* **2003**, *29*, 228–237. [CrossRef]
13. Christy, A.; Gandhi, M.G.; Vaithyasubramanian, S. Cluster based outlier detection algorithm for healthcare data. *Procedia Comput. Sci.* **2015**, *50*, 209–215. [CrossRef]
14. van Capelleveen, G.; Poel, M.; Mueller, R.M.; Thornton, D.; van Hillegersberg, J. Outlier detection in healthcare fraud: A case study in the Medicaid dental domain. *Int. J. Account. Inf. Syst.* **2016**, *21*, 18–31. [CrossRef]
15. Domingues, R.; Filippone, M.; Michiardi, P.; Zouaoui, J. A comparative evaluation of outlier detection algorithms: Experiments and analyses. *Pattern Recognit.* **2018**, *74*, 406–421. [CrossRef]
16. Jyothi, P.N.; Lakshmi, D.R.; Rao, K.V.R. A supervised approach for detection of outliers in healthcare claims data. *J. Eng. Sci. Technol. Rev.* **2020**, *13*, 204–213. [CrossRef]
17. Shao, M.; Qi, D.; Xue, H. Big data outlier detection model based on improved density peak algorithm. *J. Intell. Fuzzy Syst.* **2021**, *40*, 6185–6194. [CrossRef]
18. Yu, Y.; Miao, D.Q.; Liu, C.H.; Wang, L. An improved KNN algorithm based on variable precision rough sets. *Moshi Shibie Yu Rengong Zhineng/Pattern Recognit. Artif. Intell.* **2012**, *25*, 617–623.