Knowledge Discovery from Healthcare Electronic Records for Sustainable Environment

Naeem Ahmed Mahoto, Asadullah Shaikh, Mana Saleh Al Reshan, Muhammad Ali Memon and Adel Sulaiman

1 Department of Software Engineering, Mehran University of Engineering & Technology, Jamshoro 76062, Pakistan; naeemmahoto@gmail.com
2 College of Computer Science and Information Systems, Najran University, Najran 61441, Saudi Arabia; msareshan@nu.edu.sa (M.S.A.R.); aaalsulaiman@nu.edu.sa (A.S.)
3 Department of Information Technology, University of Sindh, Jamshoro 76090, Pakistan; muhammad.ali@usindh.edu.pk
* Correspondence: asshaikh@nu.edu.sa

Abstract: The medical history of a patient is an essential piece of information in healthcare agencies, which keep records of patients. Due to the fact that each person may have different medical complications, healthcare data remain sparse, high-dimensional and possibly inconsistent. The knowledge discovery from such data is not easily manageable for patient behaviors. It becomes a challenge for both physicians and healthcare agencies to discover knowledge from many healthcare electronic records. Data mining, as evidenced from the existing published literature, has proven its effectiveness in transforming large data collections into meaningful information and knowledge. This paper proposes an overview of the data mining techniques used for knowledge discovery in medical records. Furthermore, based on real healthcare data, this paper also demonstrates a case study of discovering knowledge with the help of three data mining techniques: (1) association analysis; (2) sequential pattern mining; (3) clustering. Particularly, association analysis is used to extract frequent correlations among examinations done by patients with a specific disease, sequential pattern mining allows extracting frequent patterns of medical events and clustering is used to find groups of similar patients. The discovered knowledge may enrich healthcare guidelines, improve their processes and detect anomalous patients’ behavior with respect to the medical guidelines.

Keywords: knowledge discovery; data mining; sequential pattern; clustering; association analysis

1. Introduction

A healthcare system is comprised of institutions, people and resources and aims to deliver health services to meet the health demands of the population. It is, in reality, a complex fusion of human-centered activities, which increasingly depend upon Information Technology (IT) and knowledge [1]. The digital age has rapidly increased the amount of data storing the medical history of patients, making healthcare systems overwhelmed with data [1–3]. However, the advancement in technology and a few software applications/tools have helped to manage, evaluate and analyze the collected healthcare data [4]. The analysis of such big data, which are often high-dimensional and sparse, is critical and physicians often do not have adequate tools for extracting useful information [5].

Data mining helps in finding hidden relationships and global patterns, which exist in large databases [6]. Among the different data mining techniques, association analysis is used to extract correlations among data in medical records, such as to discover the co-occurrence of treatments or drugs taken by patients [7,8], to detect the diseases affecting patients depending on the value of specific attributes [9] or to predict disease risk levels [10].

Another frequent problem is to extract sequences of medical diagnostic events or examinations done by patients. A diagnostic examination is a medical test, which has been
done by the patient for a certain pathology that is normally carried out by the medical team in accordance with the advice of a medical expert. The patterns of such biochemical variables or treatments undergone by patients with specific diseases [11,12] are used to see if the extracted sequences are coherent with the guidelines or represent anomalies to be further investigated. These issues are addressed by the sequential pattern mining algorithms. A third problem is to discover a group of similar entities (e.g., patients) based on certain similarity measures, such as finding patients with similar gait patterns [13], similar disease sub-type [14] or similar examination frequencies [15]. The data mining technique used to address this problem is clustering.

The aim of this paper is twofold: (1) provide a summary of recent studies which apply data mining techniques to healthcare data; (2) demonstrate a case study of knowledge discovery from healthcare data. This study applied three data mining techniques for discovering knowledge from real healthcare electronic records: (i) association analysis; (ii) sequential pattern mining; (iii) clustering. The extracted knowledge represents the medical pathways of diabetic patients. Therefore, the main contribution of the study is the demonstration of the reverse process of determining medical pathways through medical examinations of patients. Tracing the actual medical pathways followed by patients helps in evaluating healthcare services. Besides, actual medical pathways may help in getting an in-depth understanding of coherent, non-coherent and complicated cases of patients. In particular, the extracted pathways fall into three categories: (1) correct pathways; (2) new or alternative pathways; (3) erroneous pathways. The extracted actual pathways of the patient are termed as correct pathways when these are coherent with medical guidelines. New or alternative pathways are the ones which are commonly followed by patients and may not be available in the guidelines due to rare or specific diseases. The erroneous pathways are non-coherent with medical guidelines. The possible reason behind such pathways could be incorrect data entry and/or the patient may not have followed proper treatment.

The rest of the paper is organized as follows. Section 2 provides the related work which applied data mining techniques to extract medical knowledge from data. Section 3 recalls the basic concepts of data mining techniques. Section 4 is devoted to the demonstration of knowledge discovery and analysis of a real diabetic dataset. Finally, Section 5 concludes the work.

2. Related Work

The healthcare sector generates large quantities of data everywhere in the world. The advancement of technologies has provided new means of data storage and data analysis, which are greatly adopted by the healthcare sector, especially in telemedicine. Telemedicine has become an excellent opportunity for the Internet of Things (IoT). IoT is also playing a vital role in modern healthcare. It is a healthcare monitoring system consisting of wearable sensors able to recognize human activity with data mining techniques. Subasi et al. [16] proposed a data mining approach for human activity recognition. The proposed approach used the body motion data of humans performing physical activities. The accuracy of the proposed approach is measured to be 99%. They performed data analysis by applying different data mining classification techniques to the IoT dataset. The dataset was collected from wearable sensors and placed on 10 subjects during 12 different physical activities. As a result, researchers found that the algorithms Random Forest and SVM produced the highest accuracy on average with 99.89% accuracy compared to the other six algorithms. The results of this study provide an indication that both Random Forest and SVM techniques can be applied in human activity recognition, which can be advantageous to the healthcare teams. However, the number of subjects in this study was only 10, which is considerably low and may affect the results. The study did not justify the results and explain how they obtained the results or why the results were obtained in the adopted way? The researchers also did not mention the tools used to implement the different data mining techniques. Finally, they did not compare the results to other studies where the same algorithms were implemented.
Image segmentation is used in the analysis and storage of medical images in telemedicine. Medical image analysis is used for clinical monitoring, patient monitoring and preventive systems, such as driver tiredness analysis [17]. The study proposed a novel framework to analyze big datasets related to healthcare. The system is a support vector machines classification algorithm that aims at more efficient diagnosis for type II diabetes. The proposed algorithm tries to maximize the classification rate. The results of the study present better efficiency measures compared to the exciting algorithms. Further, the proposed system showed better values in terms of RMSE and MAPE measurements compared to the other two algorithms. However, the study compared the new algorithm with only two other algorithms in terms of RMSE and MAPE. Adding more algorithms to the comparison could strengthen the results. Additionally, the paper presents the efficiency kernels for the mB-SVM but does not mention the efficiency kernels for the other algorithms. The study also compared the results with other algorithms, but these algorithms were implemented with different datasets. The tools utilized to implement the results were not clearly stated in the study. Furthermore, more specifications about the dataset that was used in this study could be beneficial. The study showed an improvement compare to other techniques; however, it did not present how significant was the performance improvement or justify why the improvement took place. Finally, the paper does not specify the amount of data used in training and testing, which could be helpful to comprehend the analysis.

Apart from that, these data analysis techniques use data mining and machine learning techniques to find the desired and relevant medical data more rapidly and precisely, which can then be further used for predictive systems. Data mining models are classified into three categories: predictive, descriptive and perspective. Predictive data mining models are used in businesses to identify the risk associated with a particular set of conditions. These risks are assessed well before making the right decision. In addition, predictive data mining models improve efficiency by analyzing the previous performances for multiple tasks that demonstrate a particular behavior in the future. Descriptive algorithms represent some essential property of the data in unsupervised learning, while in supervised learning this technique typically adopts a model to foresee the value. Perspective data mining models deal with huge data volume that measures efficiency, which is strongly concerned in database perspectives. The types of data considered for analysis are web/social-media data, sensor data, biometric data, big transaction data and human-generated data. Different data mining techniques taken from the literature are regression, association, classification, clustering, anomaly detection, data warehousing and sequential pattern mining [18]. Application areas in healthcare consider clinical decision support, healthcare administration, privacy and fraud detection, mental health, public health, pharmacovigilance, heart disease, etc. [19].

A wide range of literature addresses the effectiveness of several data mining techniques for knowledge discovery in healthcare data. The following sections give an overview of recent studies carried out focusing on the three data mining techniques, namely association analysis, sequential pattern mining and clustering. This study aimed at demonstrating that evaluation of actual medical pathways is possible in the healthcare system. The extracted actual medical pathways may enrich the existing guidelines as well as understanding the deviation in the actual medical pathways than that of existing guidelines. Thus, it can be considered a subset of process mining, which is a process management technique analyzing business processes based on event logs about data, control, process, organization and social structures. In addition, the conformance analysis technique belongs to the family of process mining techniques. Therefore, the work under discussion fits the conformance analysis as it extracts actual medical pathways to understand the deviation (if any) in the actual pathways from existing guidelines. The extracted pathways also help in understanding complicated cases of patients. For example, if a patient suffers from several other diseases, then the actual medical pathways of such patient would more likely deviate from standard pathways of diabetic patients.
2.1. Correlations through Association Analysis

The association rules belong to the descriptive model of data mining, which finds the relationships among non-related/random data by combining data mining with statistical analysis. These relationships are established on similar occurrences of data items concerning similar occurrences of events. ARM [20] is interactive web-based software that allows the mining of the following association rules methods focusing on the medical record: filtered association, a priori, frequent pattern growth, predictive a priori, generalized sequential patterns, hotspot and tertius, among others. ARM develops user-friendly software that allows different groups of people to interact. ARM can, therefore, be of great use in the medical field where it is used to identify and analyze hidden patterns. The hidden patterns will help identify the causes of certain diseases such as high blood pressure and the necessary treatment plan. The software also makes it easy to identify which disease a patient could be suffering from since the pattern could be similar to that of another patient who has already been treated. ARM also helps with the retrieval of patient data, especially where the medical facility is large and serves thousands of patients in a day. Retrieval of data can be hectic when the software being used is not up to date. Loss of patient data means that the doctor will have to start the diagnosis afresh, hence the need for effective software.

Association analysis is used to find co-occurrence and correlations among healthcare data. For example, by exploiting association rule mining, Nuwangi et al. [21] studied how different risk factors contribute towards diabetes and how diabetes contributes to other diseases in a dataset of patients’ records. In this way, they identified some relationships among diabetes, wheeze and edema diseases. Another application of association rules was conducted by Chen et al. [22] to acquire viable side effects at different points in a pregnancy because of being exposed to multiple drugs. They found interesting associations between drugs and side effects, e.g., exposure to citalopram or sedative medicine late in pregnancy is correlated with a higher risk of pre-term birth.

Krysiak-Baltyn et al. [23] exploited association rule mining for finding associations in infertility-related data of military conscripts. Such data contain both categorical and continuous variables generated from biological measurements. Several interesting correlations were extracted from this dataset, such as free androgen index or smoking during pregnancy and less likely eating organic food and alcohol intake. The approach proposed in [24] reveals the potentiality of early detecting breast cancer by extracting association rule from digital mammograms. The authors of [25,26] exploited association rules to extract correlation among lung cancer patient’s data.

Contributing factors to heart disease are investigated utilizing association rules in [27,28], while the correlations among electroencephalographic signals and arousal occurrences are detected in [29]. The approaches based on association rule are presented in [9] for finding the frequent occurrences of diseases which affect patients, in [30] for automatically discovering infection and in [31] for assessing the frequency of drug prescription along with antacids.

2.2. Medical Pathways through Sequential Pattern Mining

Since patients’ records usually contain a date or a timestamp, sequential pattern mining can be applied to extract sequences of medical events. Some of the studies analyzed trends or patterns of medical diagnostic examinations over time to fetch sequences that may help in detecting a critical event (e.g., thrombosis [32]). An analysis of time sequences through biochemical variables is reported in [12]. In particular, extraction of frequent sequential patterns is carried out to determine the positive influence of individual variables towards the effectiveness of photophoretic therapy in liver transplants.

The authors of [33,34] focused on identifying the flow of patients in different wards or hospitals and optimizing healthcare resources. The time dependency pattern mining approach is reported in [35], where sequential pattern mining is exploited to extract hospital paths to be able to predict them for new patients at their admission. In [36], a temporal
pattern mining framework is proposed, which finds predictive patterns for the detection of adverse medication conditions linked with diabetes. Sequential pattern mining has also been exploited to analyze the patterns of diseases for patients who revisited a hospital to determine anticipated diseases due to revisiting patients [37].

Exarchos et al. [38] obtained proteins through sequential patterns, which later assisted in classifying the unknown proteins. Ryan [39] investigated sequences related to health behaviors of the village patients and extracted sequential patterns of treatments. Lasker [40] focused on the identification of patients’ disease through sequential symptoms. Concaro et al. [41] reported the analysis of sequential patterns in discovering frequent sequences of diagnoses shared by American hospitals. Li et al. [42] proposed an algorithm for mining risk pattern sets based on the Anti-monotone property in medical data.

Sequential pattern mining was also applied to medical examination log data in identifying medical pathways for diabetic patients [43]. Similarly, the medical pathways followed by colon cancer patients [44] and pregnant women [7] were also investigated. A method is proposed in [45] for interactive pattern mining from clinical data. The approach comprises of three components: (1) visual query module; (2) pattern mining module; (3) interactive visualization module. Clinical pathways were analyzed by Huang et al. [46] using the process mining technique. In the study, real-world datasets were evaluated for six specific diseases including lung cancer, breast cancer and colon cancer. The patterns of sleep problems are extracted from electroencephalographic signals in [29]. Locally frequent patterns for diseases are extracted in [47].

2.3. Discovering Groups through Clustering

In medical records, one major focus of medical decision support systems is to identify patients with similar medical history or records for effective diagnosis of other patients employing clustering techniques. It is related to compound information of multiple similar events occurrences of patients, clinical patterns associated with these events. Pokharel et al. [48] used the Temporal Tree Technique (TTT) to gather that compound information and used Sequential Pattern Mining (SPM) to discover more patterns. For example, a multistage algorithm for clustering Electrocardiography (ECG) data is proposed in [49], and long-term biomedical time-series signals (e.g., ECG) are grouped using an approach proposed in [50]. Ultimately, the analysis of that compound information will help in patient’s several health-related issues, even predicting future issues. Rani and Kautish [51] proposed association clustering and time series-based data to develop an early warning system for patients with diabetes. Taking into account several diabetes-related parameters such as glucose, blood pressure, insulin intake, age, skin thickness, exercise, food consumption, family history and nature of work, an actual diabetes predictive system is developed. This system uses ANN and time series-based prediction machine learning techniques. Heart disease is one of the major causes of death throughout the world. Zriqat et al. [52] developed a predictive system for the early detection of heart disease. They used SVM, KNN and Naïve Bayes supervised machine learning algorithms to predict heart disease. These were implemented in R Language, and the algorithms were measured on the accuracy parameter. Among these algorithms, the Naïve Bayes algorithm provided promising results with 86.6% accuracy. The cardiac abnormalities have been identified by Expected Maximization (EM) [53]. Clustering techniques were also exploited for identifying groups of patients having similar treatment patterns by, e.g., the authors of [15,54].

Another predictive analysis was done on breast cancer by Chaurasia et al. [55]. Exploiting the hierarchical agglomerative clustering method in [56], patients have been clustered. The authors applied three machine algorithms Naïve Bayes, RBF network and J48, on a dataset of 683 breast cancer patients. The same accuracy parameter is used for the comparison of the results. The results indicate that Naïve Bayes predicts with an accuracy of 97.36%, while RBF came second, followed by J48.
The diabetic patients are classified in [57] by hybrid methods that incorporate the k-means algorithm. Hirano et al. [58] discussed the potentiality of clustering methods in medical datasets. K-means clustering helps in building an analytical model for patient flow in hospitals [59].

3. Fundamental Concepts of Data Mining Techniques

In this section, the basic concepts and definitions needed to describe the three mentioned data mining techniques (i.e., association analysis, sequential pattern mining and clustering) are reported.

3.1. Association Analysis

This methodology is applied to a transactional dataset to discover interesting relationships among attributes. The extracted relationships are in the form of frequent item sets and frequent association rules. The fundamental terminologies used in the association analysis are defined in the definitions.

The evaluation of the association rules is performed using well-established indices namely, association rule support, association rule confidence and association rule lift. The association rule support simply represents the frequency of the antecedent and consequent in a given dataset. The association rule confidence, however, provides the probability that the consequent may likely happen in the presence of antecedent. For example, if an association rule $X \Rightarrow Y$ has confidence value 85%, it means in 15% of relationships in the given dataset $X$ does not imply to $Y$, while in 85% of relationships $X$ implies $Y$ holds true. Association rule lift is an interesting indexing measure that shows the correlation between antecedent and consequent either positive (greater than 1 value), negative (less than 1 value) or neutral (value 1). For example, consider in a given dataset that the frequency of an item $X$ is 60% (i.e., $\text{supp}(X) = 60\%$). Similarly, the occurrence of another item $Y$ is 50% (i.e., $\text{supp}(Y) = 50\%$) and both items together appear 50%, i.e., $\text{supp}(X \cup Y) = 50\%$. The lift value for the association rule $X \Rightarrow Y$ equals 1.66. The greater than 1 lift value means that correlation between $X$ and $Y$ is much higher due to the fact every time $Y$ appears in a given dataset is together with $X$ item. However, item $X$ appears 10% of the time when it is not together with item $Y$.

**Definition 1 (Item set).** Let the set of attributes $I = \{i_1, i_2, \ldots, i_n\}$ be called items and $T = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions (or rows) in a given dataset. Each row (i.e., transaction) in $T$ possesses a unique identifier and comprised of certain items from $I$. An item set $X$ is any set of items belonging to $I$, i.e., $X \subseteq I$.

**Definition 2 (Item set support).** The support value, denoted as $\text{supp}(X)$, of an item set $X$ is defined as the number of transactions in a given dataset containing $X$: $\text{supp}(X) = |\{t_i | X \subseteq t_i, t_i \in T\}|$.

**Definition 3 (Closed Itemset).** An item set $I$ is called a closed item set as long as there exists no superset $I' \supset I$, such that $\text{supp}(I') = \text{supp}(I)$ [60].

**Definition 4 (Association rule).** An association rule represents implication among sets of items, and it is expressed in the form $A \Rightarrow B$, where $A$ and $B$ are item sets and $A \cap B = \emptyset$. $A$ is called the body or antecedent of the rule, while $B$ is called the head or consequent. The strength of an association rule is measured by three parameters, i.e., the support, confidence and lift [6].

**Definition 5 (Association rule support).** Let $A \Rightarrow B$ be an association rule. The support of the rule, denoted as $\text{supp}(A \Rightarrow B)$, is the joint probability of a transaction containing both $A$ and $B$, given as $\text{supp}(A \Rightarrow B) = P(A \cap B) = \text{supp}(A \cup B)$. 


Definition 6 (Association rule confidence). Let $A \Rightarrow B$ be an association rule. The confidence of the rule is the conditional probability that a transaction contains $B$, given that it contains $A$, given as $\text{conf}(A \Rightarrow B) = P(A|B) = \frac{P(A \land B)}{P(A)} = \frac{\text{supp}(A \cup B)}{\text{supp}(A)}$.

Definition 7 (Association rule lift). Consider $A \Rightarrow B$ represents an association rule. The lift of the rule measures dependency between the antecedent (i.e., $A$) and the consequent (i.e., $B$) and is given by: $\text{lift}(A \Rightarrow B) = \frac{\text{conf}(A \Rightarrow B)}{\text{supp}(B)} = \frac{\text{supp}(A \cup B)}{\text{supp}(A) \text{supp}(B)}$. Lift values range within $(0, +\infty)$. Values approaching 1 mean the antecedent (i.e., $A$) and the consequent (i.e., $B$) are independent. A lift value greater than 1 reflects a correlation of the antecedent and the consequent in appearing together more often than expected. A lift value less than 1 indicates the presence of negative correlations between $A$ and $B$.

Definition 8 (Closed association rule). An association rule $A \Rightarrow B$ is called a closed association rule if $A$ and $B$ are closed item sets.

3.2. Sequential Pattern Mining

Sequential pattern mining aims at finding frequent sub-sequences in a given set of sequences (i.e., sequence database) with a support value satisfying the user-specified minimum support threshold [6]. The basic concepts of sequential pattern mining are defined in the definitions.

A closed sequence removes the redundancy and, thus, reduces the number of patterns. For example, consider a sequence $S_a = \{e_1\} \{e_2\}$ having support value 20% in a given dataset and another sequence $S_b = \{e_1\} \{e_2, e_3\}$ with 20% support value. The sequence $S_b$ is closed sequence because $S_a$ is contained within $S_b$. In other words, there is no super-sequence of $S_b$ with the same support of 20%.

Definition 9 (Sequence). A sequence $S$ is an ordered list of item sets $S = < I_1 I_2 ... I_k >$, where each item set refers to a specific time such that $t_1 < t_2 < \ldots < t_k$.

Definition 10 (Sequence length). The length of a sequence, $|S|$, is given by the number of item sets of the sequence. A sequence that contains $L$ item sets is called L-sequence.

Definition 11 (Subsequence). A sequence $S = < s_1 s_2 \ldots s_n >$ being part of another sequence $SS = < s_1 s_2 \ldots s_m >$ (m ≥ n) as long as there have been integers $i_1 < i_2 < \ldots < i_n$ such that $S_1 \subseteq SS_{i_1}, S_2 \subseteq SS_{i_2}, \ldots, S_n \subseteq SS_{i_n}$. $S$ is called a subsequence of $SS$ and $SS$ is called a supersequence of $S$.

Definition 12 (Sequence support). The support of a sequence, $\text{supp}(S)$, is given by the number of transactions of the dataset containing $S$ as subsequence.

Definition 13 (Closed sequence). A sequence $C$ is called a closed sequence as long as no supersequences of $C$ with the same support exist.

3.3. Clustering

Clustering is an approach to evaluate the similarity among the objects of a dataset based only on data characteristics without a priori information [6]. Clustering algorithms create a set of groups, named clusters, in which similar objects are kept in the same group, while dissimilar ones are kept in different groups. Clustering algorithms may be categorized into three main categories depending on the technique on which the clusters are created: partitioning, hierarchical and density-based. For each algorithm, it is possible to exploit several similarity measures to compute the similarity between the objects. Furthermore, several indices are available to evaluate cluster results in terms of intra-cluster homogeneity and inter-cluster heterogeneity.
3.3.1. Clustering Techniques

Partitioning clustering (e.g., K-means [61]) dissects a given set of data objects into non-overlapping groups (i.e., clusters), such that each object is more similar (nearer) to the centroid of the cluster compared to the centroid of any other cluster. These methods find spherical-shaped clusters and have the advantage of being simple and efficient in terms of execution time, but they are not suitable for finding non-spherical clusters and clusters with different sizes and densities. Furthermore, they are highly sensitive to outliers.

Different from partitioning clustering, hierarchical clustering allows clusters to have subclusters. For instance, an agglomerative, also known as the divisive approach, generates a hierarchical set of clusters and produces a tree of nested clusters called a dendrogram. Hierarchical clustering is often used when the underlying application requires a taxonomy creation, and this clustering approach often reaches better results compared with partitioning clustering. Their main disadvantage is they are highly expensive in terms of computational and storage requirements, thus they are mostly used on small datasets.

In density-based clustering (e.g., DBSCAN [62]), a cluster is considered as a dense area of data objects, which are surrounded by a low-density area. This type of clustering is less sensitive towards outliers and can identify non-spherical-shaped clusters. These algorithms usually do not need the number of clusters as an input parameter but only the minimum number of points to form a cluster and the maximum allowed distance between two objects in the same cluster.

3.3.2. Proximity Functions

Different measures can be adopted to compute the distance between objects during clustering, depending on the nature of their dimensions [6].

Let \( p \) and \( q \) be two vectors of \( m \) dimensions. The Euclidean distance is often used for a data point in Euclidean space, and it is computed as follows:

\[
d_E(p, q) = \sqrt{\sum_{i=1}^{m} (p_i - q_i)^2}
\]

(1)

A variation of this measure is the Manhattan distance, which computes the distance to get from one object to another if a grid-like path is followed:

\[
d_M(p, q) = \sum_{i=1}^{m} |p_i - q_i|
\]

(2)

The Chebyshev distance computes the distance between two objects as the greatest of the differences along all their dimensions:

\[
d_{Ch}(p, q) = \max_{i=1}^{m} |p_i - q_i|
\]

(3)

The Hamming distance counts the number of dimensions at which the corresponding values are different:

\[
d_H(p, q) = \sum_{i=1}^{m} H(p_i, q_i)
\]

(4)

where \( H(p_i, q_i) = 0 \leftrightarrow p_i = q_i \) and \( H(p_i, q_i) = 1 \leftrightarrow p_i \neq q_i \).

The cosine similarity measures the cosine of the angle formed by the two objects considered as vectors:

\[
\text{cosine}(p, q) = \frac{\sum_{i=1}^{m} p_i q_i}{\sqrt{\sum_{k=1}^{m} p_k^2} \sqrt{\sum_{k=1}^{m} q_k^2}}
\]

(5)

If the two vectors have the same orientation, the cosine similarity is equal to 1; if they are at 90 degrees, they have a similarity of 0; if they are diametrically opposed, they have a similarity of −1, independently of their magnitude. The cosine similarity is commonly
used in text mining, where each term is assigned a different dimension and a document is characterized by a vector where the value of each dimension corresponds to the weighted number of times that term appears in the document [63]. Since the cosine is a measure of similarity, the following formula can be applied to have the corresponding distance:

$$d_{Co}(p, q) = \arccos(\cosine(p, q))$$ (6)

3.3.3. Evaluation Indices

To evaluate clustering results, several indices can be exploited. One of the most used is the silhouette [64], which evaluates if an object has been assigned to the correct cluster. Its values range in $[-1, 1]$, and it is computed for each object $i$ of the dataset as follows:

$$\text{Silhouette}_i = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$ (7)

where $a(i)$ is the average distance of $i$ with the other objects of the same cluster. The term $b(i)$ is the smallest of the average distances from the objects to the other clusters. A value near 1 means that the object is placed in the right cluster, while a value near $-1$ indicates that there is another cluster where the object would be better placed instead of the current one. The average value of all silhouettes of objects represents the silhouette of the whole result.

Other measures have been proposed to quantify the inter-cluster homogeneity and the inter-cluster heterogeneity [65]. They are defined as follows:

$$\text{Homogeneity}_c = \frac{2}{n_c(n_c - 1)} \sum_{i,j \in C, i < j} s(i, j)$$ (8)

$$\text{Heterogeneity}_c = \sum_{i \in C} d(i, z_c)$$ (9)

where $n_c$ is the number of elements in cluster $C$, $s(i, j)$ and $d(i, j)$ are the similarity and tand of the distance functions computed between objects $i$ and $j$ and $z_c$ is the element representative of cluster $C$ (e.g., the centroid). The homogeneity increases if the solution improves, while the heterogeneity increases if the solution gets worse. The homogeneity and the heterogeneity of the whole clustering are the average values of all clusters.

4. Knowledge Discovery from Real Healthcare Data—A Case Study

This section is devoted to the applications of the data mining techniques previously described. The knowledge discovery from a real healthcare dataset is shown in Figure 1. The knowledge discovery process is comprised of three major steps: data preparation, knowledge discovery and results evaluation. Data preparation steps are prerequisites of the knowledge discovery steps. While results evaluation step is the evaluation of the extracted knowledge. The result evaluation is carried out in two ways: (1) evaluation indices; (2) medical expert. The evaluation index is used according to the data mining technique applied and medical experts evaluated the results in accordance with medical guidelines. The subsequent sections describe the healthcare electronic data used in the study, while the other three subsections report the results of the data mining techniques applied to the dataset to derive medical knowledge. The experiments were performed on a computer with a processor 3 GHz Dual-Core Intel Core i7 and 8 GB RAM. The data mining techniques were implemented in Java programming language due to the fact that the enquired data pre-processing in the transformation of sequence database from the transactional dataset was not supported within existing tools such as WEKA, R and other such tools/application. Besides, the clustering evaluation index also needed to be implemented. However, any other tool/application can also be used in applying data mining algorithms.
Figure 1. Discovering knowledge from healthcare data.

4.1. Healthcare Electronic Records

Healthcare systems record for each diabetic patient the examinations performed together with their date. From this data, interesting information can be extracted. For example, with the association analysis, the examinations frequently performed on the same day by patients can be detected. This analysis allows healthcare systems to control the examination process because the detected examination sets can deviate from predefined guidelines, for example, if different or additional exams have been carried out, or, if some exams are missing, medical experts may take necessary and timely measures. In addition, the sequential pattern mining can be profitably used to extract sequences of examination sets, and thus compare the medical pathways followed by patients with the theoretical ones. The pathways representing non-compliance may be determined due to patient negligence in strictly following medical treatments, incorrect procedures during record-keeping or the existence of different guidelines for specific cases. Finally, in order to group similar patients, clustering techniques can be adapted to divide the data into smaller and meaningful sets that are easier to analyze.

To perform such analyses, we collected electronic records of diabetic patients from an agency (the name is kept secret due to privacy), and we arranged them in the form of a transactional patient dataset, reporting examinations done for each patient with time, as shown in Table 1. The healthcare agency covers several medical hospitals digitally connected with a central database. The diagnostic examinations are carried out by the advice of the medical expert; thus, the complete process is in accordance with medical guidelines. Each record of the transactional dataset (cf. Table 1) represents patient-id, date and diagnostic examination(s). The dataset includes 96,000 medical records of the diagnostic examinations performed by 6350 diabetic patients. The number of distinct diagnostic examinations is 160. Table 1 reports an example of this dataset. The adopted approach requires attributes such as patient-id, date and diagnostic examination for pre-processing. Different healthcare agencies may follow different data representation formats, which does not affect the pre-processing step in which unnecessary attributes are cleaned. This dataset was further transformed into a sequence dataset, where each row contains the sequence of examination sets performed by the patient, i.e., the set of examinations done on the same date. Table 2 shows the sequence dataset corresponding to the transactional patient dataset in Table 1. For example, Sequence 1 in Table 2 shows that Patient 1 went
through two diagnostic examinations (i.e., glucose and capillary blood) on the same day. Sequence 3 reflects that Patient 3 went through one diagnostic examination (i.e., urine test) on a single day and two more diagnostic examinations (i.e., venous blood and glucose) on some other day. The sequence database does not consider the temporal sequences, i.e., sequence with a timestamp, in this study. For example, a patient is required to be diagnosed for a certain examination at a given time. This aspect of timing is not considered in this study. However, the order of the sequence is considered. If a patient performed diagnostic in a different order than that of guidelines, then the pathways of such patient would be extracted as non-coherent with guidelines. This sequence dataset (i.e., Table 2) is used for the knowledge discovery in terms of association analysis, sequential pattern mining and clustering.

Table 1. Diabetic patient dataset.

| Patient ID | Date             | Examination    |
|------------|------------------|----------------|
| 1          | 17 February 2007 | Glucose        |
| 1          | 17 February 2007 | Capillary blood|
| 2          | 25 July 2007     | Glucose        |
| 3          | 8 January 2007   | Urine Test     |
| 3          | 15 February 2007 | Venous blood   |
| 3          | 15 February 2007 | Glucose        |
| ...        | ...              | ...            |

Table 2. Sequence dataset.

| Patient ID | Examination Sequence                                        |
|------------|-------------------------------------------------------------|
| 1          | {Glucose, Capillary blood}                                  |
| 2          | {Glucose}                                                   |
| 3          | {Urine Test} {Venous blood, Glucose}                        |
| ...        | ...                                                         |

4.2. Association Analysis

We applied the association analysis to extract the frequent examination sets and frequent correlation between examination sets. To this aim, closed association rules [66] were generated from the considered dataset. The Java implementation of the closed association rules was provided by Philippe Fournier-Vigier [67], and it was properly modified to include the lift computation.

By setting a minimum support threshold of 10%, the total number of extracted item sets (i.e., closed item sets) is 2231. From the analysis of frequent closed item sets, it emerged that the extracted sets are generally in accordance with the medical guidelines, even though some anomalies were detected. The diagnostic examinations that are the base tests and routinely repeated by diabetic patients remained at the top extraction. These results, as shown in the top of Table 3, help in monitoring the concentration of sugar in the blood. Particularly, glucose, venous and capillary blood samples and urinalysis are the top examinations with higher than 75% occurrences. Despite higher frequency than that of other examinations, these examinations have been performed significantly less often than expected. Overall, 15% patients did not undergo the blood glucose level examination, revealing an anomaly or some problem in patient management. Thus, it needs further investigation for possible causes and proper treatment. By carefully analyzing these sets, another anomaly emerged. According to medical guidelines [68], glucose level analysis requires an association with at least one of these diagnostic examinations: venous blood, capillary blood or urine. Most of the discovered examination sets verify this constraint. For example, {Glucose, Urinalysis} (74.8% of patients), {Glucose, Capillary Blood} (74.4%) and {Glucose, Venous blood} (71.0%) were detected, which confirms the coherence with medical guidelines. However, among 5.56% of patients, glucose level examinations were
not detected associated with the three examinations. These results deviate from the medical guidelines and highlight an error. Medically, it is impossible to measure glucose levels without urine or blood samples. The possible reason behind such erroneous results may be incorrect data entry, which may have relied only on a subset of diagnostic examinations instead of a the complete set of performed examinations.

Examinations sets represent the examinations that are frequently done together by patients. The analysis of such sets is useful to have a view of the behavior of patients and the examinations frequently done by them. However, correlations may exist among examinations that are not so frequent. To discover such information, the closed association rules are extracted. Association rules are used to evaluate how the presence of a set of examinations is dependent on the presence of another set of examinations, and, in addition to support, they have confidence and lift values. By setting a minimum support threshold of 25% and a minimum confidence value of 96%, 92 rules were extracted. Examples of such rules are reported in Table 3, sorted by decreasing lift values.

It can be noticed that rules with higher lift are usually the ones with lower support. In these rules, examinations more specific than the ones extracted with the item sets emerged. For example, the rule \{Glucose, Venous blood, ALT, Hemoglobin\} ⇒ \{AST\} states that almost all (98%) patients who did the glucose, venous blood, ALT and hemoglobin examinations also did the AST examination. The lift value represents the factor by which the confidence of the rule exceeds the expected confidence, in this case 3.27. This rule represents the behavior of patients suffering from liver complications and thus performing the Alanine Transaminase (ALT) and the Aspartate Transaminase (AST) examinations, which are usually done together. The rule \{Venous blood, Triglycerides, Total cholesterol, HDL cholesterol\} ⇒ \{Hemoglobin\} states that almost all the patients who did the venous blood, triglycerides, total cholesterol and HDL cholesterol examinations, also did the hemoglobin one. This rule may represent the behavior of patients who suffer from cardiovascular complications of diabetes. In fact, low levels of HDL cholesterol and high level of triglycerides increase the risk for heart disease. These examinations are associated with hemoglobin because this examination may also reveal cardiovascular complications. In fact, hemoglobin is measured primarily to identify the average plasma glucose concentration over prolonged periods, and high amounts of hemoglobin indicate poorer control of blood glucose levels.

Table 3. Extracted closed examination sets and closed association rules.

| Closed Frequent Item sets | Support |
|---------------------------|---------|
| \{Glucose\}               | 84.8%   |
| \{Venous blood\}          | 79.2%   |
| \{Capillary blood\}       | 75.0%   |
| \{Urinalysis\}            | 74.9%   |
| \{Glucose, Urinalysis\}   | 74.8%   |
| \{Glucose, Capillary blood\} | 74.4%   |
| \{Glucose, Urinalysis, Capillary blood\} | 73.9% |
| \{Glucose, Venous blood\} | 71.0%   |
| \{Glucose, Urinalysis, Capillary blood, Venous blood\} | 57.4% |
| \{Hemoglobin\}            | 46.4%   |
| \{Hemoglobin, Venous blood\} | 43.0%   |
| \{Cholesterol\}           | 36.0%   |
| \{Triglycerides\}         | 35.7%   |
| \{HDL cholesterol\}       | 35.4%   |
| \{Cholesterol,Triglycerides,HDL cholesterol\} | 33.7% |

| Closed Association Rules | Support | Confidence | Lift |
|--------------------------|---------|------------|------|
| \{Venous blood, AST, HDL cholesterol\} ⇒ \{ALT, Total cholesterol\} | 25% | 98% | 3.66 |
| \{Glucose, Venous blood, ALT, Hemoglobin\} ⇒ \{AST\} | 26% | 98% | 3.27 |
| \{Venous blood, Hemoglobin, HDL cholesterol\} ⇒ \{Triglycerides, Total cholesterol\} | 33% | 97% | 2.72 |
| \{Venous blood, Triglycerides, Total cholesterol, HDL cholesterol\} ⇒ \{Hemoglobin\} | 33% | 96% | 2.03 |
| \{Glucose, Capillary blood, Venous blood\} ⇒ \{Urine Test\} | 64% | 100% | 1.32 |
| \{Capillary blood, Venous blood\} ⇒ \{Urine Test\} | 65% | 99% | 1.32 |
| \{Glucose, Urine Test\} ⇒ \{Capillary blood\} | 73% | 98% | 1.32 |
| \{Urine Test, Venous blood\} ⇒ \{Glucose\} | 65% | 99% | 1.17 |
4.3. Sequential Pattern Mining

Table 4 reports the frequent examination sequences, which were detected (26,838 sequences) using BIDE algorithm [69] with a minimum support threshold of 25%. It was analyzed that glucose level has been examined twice for 58% of patients, thrice for 32% of patients and four times for 15% of patients during a one-year period. These sequence patterns are in line with medical knowledge, although the frequency of the sequences remained less than expected. This can highlight the fact that many patients do not access the public service but prefer to perform examinations privately.

Regarding this specific disease, it is normal that no other sequences than the ones controlling the glucose level are frequent. In fact, medical guidelines do not prescribe a sequence of examinations as in other cases such as the examinations to perform in pregnancy. Instead of considering the whole dataset, it is also possible to extract sequences only from patients suffering from a specific disease complication. In this case, a subset of patients is selected from the original dataset to restrict the analysis. For example, a severe deterioration of diabetes is the damage of the eye retina (retinopathy). Retinal photo-coagulation refers to a therapy performed for repairing retina lacerations. To analyze the frequency of sequences characterizing patients who performed the retinal photo-coagulation therapy, it was observed that the 140 patients (i.e., 2.204% of the total) performed this treatment at least once. Evidence of multiple treatments was analyzed because the frequent sequences included repetition of the therapy: two times (1.13% of the total) or three times (0.56% of the total). These sequences are coherent with the medical knowledge because a patient with proliferative retinopathy regularly stands in need of multiple therapy treatments. Surprisingly, 5.56% (i.e., 353 patients) performed at least once glucose level examination without any association of three examinations: (a) venous blood; (b) capillary blood; (c) urine. The glucose level can be measured by analyzing at least one of these three examinations associated with glucose. This reflects an erroneous pathway.

Table 4. Frequent examination sequences.

| Examination Sequences                                            | Frequency (%) |
|------------------------------------------------------------------|---------------|
| {Glucose}|{Capillary blood}                                                | 53            |
| {Venous blood}|{Glucose}                                                        | 53            |
| {Venous blood}|{Capillary blood}                                                | 48            |
| {Venous blood}|{Glucose, Capillary blood, Urine}                                | 48            |
| {Capillary blood}|{Glucose}|{Glucose}                                                        | 26            |
| {Glucose, Urine}|{Glucose}|{Glucose, Capillary blood}                                      | 25            |
| {Venous blood}|{Glucose, Capillary blood}|{Venous blood}        | 20            |
| {Capillary blood}|{Glucose, Venous blood}|{Glucose}                                                        | 20            |
| {Capillary blood, Venous blood}|{Glucose}|{Glucose}                                                        | 17            |
| {Haemoglobin}|{Glucose}|{Venous blood}                                                   | 16            |

4.4. Clustering

Instead of manually selecting a subset of data on which to perform the analysis, clustering algorithms can be exploited to automatically divide patients into groups, based on the similarity of their performed examinations. Among clustering techniques, DBSCAN demonstrated noteworthy characteristics for clustering patients, because it is less sensitive towards noisy data and outliers. In addition, prior information about a number of clusters is not required for DBSCAN. The discovery of patient clusters using DBSCAN from the healthcare diabetic dataset was implemented with the help of RapidMiner [70].

To apply the clustering, the dataset was represented according to the Vector Space Model (VSM) [71]. A patient is mapped as a vector in the examination space, and each vector element correlates with a different examination. The value of a vector element is the number of times an examination is done by a patient, divided by the number of
patients who have undergone that examination. This representation is known as The Term Frequency (TF)—Inverse Document Frequency (IDF) in text mining [6].

To measure the similarity between patients, we exploited the cosine similarity measure between the weighted examination frequency vectors. To select the value of the DBSCAN parameters (the minimum number of elements in each cluster \(\text{MinPts}\) and the maximum distance among elements of the same cluster \(\text{Eps}\)), we performed a set of experiments varying the parameters and measuring the average (Avg.) silhouette and standard deviation (STD) of silhouette values, as shown in Figure 2.

![Figure 2. Cluster evaluation index—silhouette.](image)

The value \(\text{MinPts} = 30\) was determined with the help of k-nearest neighbor distance (k-dist), which provided fewer data as outliers. However, the \(\text{Eps}\) value determination is a challenge. Therefore, several experiments were performed to determine the correct value of \(\text{Eps}\), which allowed better clustering. As illustrated in Figure 2, different \(\text{Eps}\) values produced different numbers of clusters, while \(\text{MinPts}\) remained 30. The selection of \(\text{Eps}\) depends on the better cluster index value (i.e., silhouette). For example, \(\text{Eps} = 0.2\) resulted 10 clusters, Avg. silhouette equal to 0.763 and standard deviation of 0.427. Likewise, \(\text{Eps} = 0.4\) produced 10 clusters, Avg. silhouette = 0.77 and standard deviation of 0.216. The highest silhouette (i.e., Avg silhouette = 0.813 and standard deviation of 0.035) was achieved with \(\text{Eps} = 0.3\) and \(\text{MinPts} = 30\), thus we analyzed the 11 clusters generated by this configuration. The obtained 11 clusters can be classified into two groups: Group 1 includes the clusters comprising of patients with standard examinations of glucose level (cluster \(C_1\)–\(C_5\)) and Group 2 includes the clusters containing patients presenting diabetes complications (\(C_6\)–\(C_{11}\)). This labeling of clusters into two categories was done through the advice of the medical expert, who assisted experimental results according to medical guidelines. The outliers (i.e., patients having significant deviation in examination history) were not included in any clusters. Figure 3 reports, for each group of clusters, the set of examinations done by patients in these clusters. In the following, the main characteristics of each cluster are reported.
Figure 3. Clustering results representing different subgroups of patients.

- **Group 1, C₁–C₅ clusters:** This group of clusters contained the two largest clusters (C₁ and C₂) having patients who have been examined for standard routine tests. In addition to routine examinations, patients of C₄ had a specialistic visit for the detection of diabetes complications. The patients of C₄ and C₅ merely visited for routine check-ups. These patients in C₄ and C₅ have undergone private diagnostic examinations and reported results in the healthcare agency.

- **Group 2, C₆–C₁₁ clusters:** It was analyzed that the patients in clusters C₆–C₁₁ have been tested additionally for diabetes complications in: (i) eye (C₆); (ii) cardiovascular system (C₇); (iii) both eye and cardiovascular system (C₈); (iv) carotid (C₉); (v) limb (C₁₀). Finally, Cluster C₁₁ comprised patients who underwent tests for liver, kidneys and cardiovascular. Besides, standard routine diagnostic examinations have also been observed in C₆–C₁₁ to be comparatively less than those of clusters C₁–C₅.

These detected 11 clusters contain around 50% of patients of the considered dataset. The remaining patients are labeled as outliers. To deeply investigate the outliers, the same clustering approach can be reapplied to the outlier set to discover new clusters. The clusters achieved in this second step contain additional examinations of the diabetic complication, thus representing patients who underwent more critical and rarer examinations. The clustering outcomes help in understanding and differentiating standard routine diabetic patients, complicated cases and those who have gone through rare diagnostic examinations.

### 5. Conclusions and Future Work

The different data mining techniques are applied in knowledge discovery from real healthcare electronic records of diabetic patients. The experimental results show the effectiveness of the three data mining techniques (i.e., association rule mining, sequential pattern mining and clustering) exploited in this study. In particular, sequential pattern mining rebuilt the treatment procedures from the considered healthcare data. Association rule mining potentially uncovered the associations between diagnostic examinations and their interdependency on each other. Moreover, the clustering technique deeply investi-
gated the dataset and detected several subgroups with homogeneous medical treatments present in the dataset. The detected subgroups highlighted the different complications being seen within diabetic treatments. The discovered information may be applied in existing guidelines to enrich them, and, at the same time, it may help healthcare management to utilize their resources efficiently since experimental results define the dependency of examinations and possible severe conditions of the diabetic patients. This work is only limited to diagnostic examinations and three techniques were applied. As future work, we intend to work on time series of sequential patterns, i.e., what patterns are observed with respect to time and age factors that must be considered for further analysis.

**Author Contributions:** Conceptualization, N.A.M. and A.S. (Asadullah Shaikh); methodology, A.S. (Asadullah Shaikh) and M.S.A.R.; software, N.A.M.; validation, M.A.M., A.S. (Adel Sulaiman) and A.S. (Asadullah Shaikh); formal analysis, M.S.A.R. and A.S. (Asadullah Shaikh); investigation, N.A.M.; resources, A.S. (Asadullah Shaikh); data curation, M.A.M.; writing—original draft preparation, N.A.M. and A.S. (Asadullah Shaikh); writing—review and editing, A.S. (Adel Sulaiman); visualization, M.S.A.R.; supervision, A.S. (Adel Sulaiman); project administration, M.M and A.S. (Asadullah Shaikh); funding acquisition, A.S. (Asadullah Shaikh). All authors have read and agreed to the published version of the manuscript.

**Funding:** Authors would like to acknowledge the support of the Deputy for Research and Innovation-Ministry of Education, Kingdom of Saudi Arabia for this research through a grant (NU/IFC/INT/01/008) under the institutional Funding Committee at Najran University, Kingdom of Saudi Arabia.

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

**References**

1. Tien, J.M.; Goldschmidt-Clermont, P.J. Healthcare: A complex service system. *J. Syst. Sci. Syst. Eng.* 2009, 18, 257–282. [CrossRef]
2. El-Sappagh, S.H.; El-Masri, S.; Riad, A.M.; Elmogy, M. Data Mining and Knowledge Discovery: Applications, Techniques, Challenges and Process Models in Healthcare. *Int. J. Eng. Res. Appl.* 2013, 3, 900–906.
3. Schmidt, S.; Vuillermin, P.; Jenner, B.; Ren, Y.; Li, G.; Chen, Y.P.P. Mining Medical Data: Bridging the Knowledge Divide. In Proceedings of the eResearch Australasia, Melbourne, Australia, 28 September–3 October 2008; pp. 1–10.
4. Simon, S.; Kaushal, R.; Cleary, P.; Jenter, C.; Volk, L.; Orav, E.; Burdick, E.; Poon, E.; Bates, D. Physicians and electronic health records: A statewide survey. *Arch. Intern. Med.* 2007, 167, 507. [CrossRef] [PubMed]
5. Prather, J.C.; Lobach, D.F.; Goodwin, L.K.; Hales, J.W.; Hage, M.L.; Hammond, W.E. Medical data mining: Knowledge discovery in a clinical data warehouse. In Proceedings of the AMIA Annual Fall Symposium, Nashville, TN, USA, 25–29 October 1997; pp. 101–105.
6. Sumathi, S.; Sivanandam, S. *Introduction to Data Mining and Its Applications*; Springer: Berlin/Heidelberg, Germany, 2006; Volume 29, ISSN 1860-9503.
7. Antonelli, D.; Baralis, E.M.; Chiusano, S.A.; Mahoto, N.A.; Bruno, G.; Petrigni, C. Extraction of medical pathways from electronic patient records. In *Medical Applications of Intelligent Data Analysis: Research Advancements*; IGI Global: Hershey, PA, USA, 2012; pp. 273–289.
8. Lakshmi, K.; Kumar, G.S. Association rule extraction from medical transcripts of diabetic patients. In Proceedings of the 2014 Fifth International Conference on the Applications of Digital Information and Web Technologies (ICADIWT), Chennai, India, 17–19 February 2014; pp. 201–206.
9. Ilayaraja, M.; Meyyappan, T. Mining medical data to identify frequent diseases using Apriori algorithm. In Proceedings of the 2013 International Conference on Pattern Recognition, Informatics and Medical Engineering (PRIME), Tamilnadu, India, 21–22 February 2013; pp. 194–199.
10. Khaing, H.W. Data mining based fragmentation and prediction of medical data. In Proceedings of the 2011 3rd International Conference on Computer Research and Development (ICCRD), Shanghai, China, 11–13 March 2011; Volume 2, pp. 480–485.
11. Antonelli, D.; Bruno, G.; Chiusano, S. Anomaly detection in medical treatment to discover unusual patient management. *IIE Trans. Healthc. Syst. Eng.* 2013, 3, 69–77. [CrossRef]
12. Berlingerio, M.; Bonchi, F.; Giannotti, F.; Turini, F. Mining clinical data with a temporal dimension: A case study. In Proceedings of the 2007 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2007), Fremont, CA, USA, 2–4 November 2007; pp. 429–436.
13. Sawacha, Z.; Guarneri, G.; Avogaro, A.; Cobelli, C. A new classification of diabetic gait pattern based on cluster analysis of biomechanical data. *J. Diabetes Sci. Technol.* 2010, 4. [CrossRef] [PubMed]
14. van Rooijen, S.M.; Colas, F.; Martínez-Martín, P.; Visser, M.; Verbaan, D.; Marinus, J.; Chaudhuri, R.K.; Kok, J.N.; van Hilten, J.J. Clinical subtypes of Parkinson’s disease. *Mov. Disord.* 2011, 26, 51–58. [CrossRef]
15. Antonelli, D.; Baralis, E.; Bruno, G.; Cerquitelli, T.; Chiusano, S.; Mahoto, N. Analysis of diabetic patients through their examination history. *Expert Syst. Appl.* **2013**, *40*, 4672–4678. [CrossRef]

16. Subasi, A.; Radhwan, M.; Kurdi, R.; Khateeb, K. IoT based mobile healthcare system for human activity recognition. In Proceedings of the 15th Learning and Technology Conference (L&T), Jeddah, Saudi Arabia, 25–26 February 2018; pp. 29–34.

17. Kumar, S.R.; Gayathri, N.; Muthuramalingam, S.; Balamurugan, B.; Ramesh, C.; Nallakuruppan, M. Medical big data mining and processing in e-healthcare. In *Internet of Things in Biomedical Engineering*; Elsevier: Amsterdam, The Netherlands, 2019; pp. 323–339.

18. Rose, K. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems. *Proc. IEEE* **1998**, *86*, 2210–2239. [CrossRef]

19. Singh, P.; Singh, S.; Pandi-Jain, G.S. Effective heart disease prediction system using data mining techniques. *Int. J. Nanomed.* **2018**, *13*, 121–124. [CrossRef] [PubMed]

20. Perçin, İ.; Yağın, F.H.; Güldoğan, E.; Yoloğlu, S. ARM: An Interactive Web Software for Association Rules Mining and an Application in Medicine. In Proceedings of the 2019 International Artificial Intelligence and Data Processing Symposium (IDAP), Malaty, Turkey, 21–22 September 2019; pp. 1–5.

21. Nuwangi, S.; Oruthotaarachchi, C.; Tilakaratna, J.; Caldera, H. Usage of association rules and classification techniques in knowledge extraction of diabetes. In Proceedings of the 2010 6th International Conference on Advanced Information Management and Service (IMS), Seoul, Korea, 30 November–2 December 2010; pp. 372–377.

22. Chen, Y.; Pedersen, L.; Chu, W.; Olsen, J. Drug exposure side effects from mining pregnancy data. *ACM SIGKDD Explor. Newsl.* **2007**, *9*, 22–29. [CrossRef]

23. Krysiak-Baltyn, K.; Nordahl Petersen, T.; Audouze, K.; Jørgensen, N.; Ångquist, L.; Brunak, S. Compass: A hybrid method for clinical and biobank data mining. *J. Biomed. Inform.* **2014**, *47*, 160–170. [CrossRef]

24. Mohanty, A.K.; Senapati, M.R.; Lenka, S.K. An improved data mining technique for classification and detection of breast cancer from mammograms. *Neural Comput. Appl.* **2013**, *22*, 303–310. [CrossRef]

25. Qiang, Y.; Guo, Y.; Li, X.; Wang, Q.; Chen, H.; Cuic, D. The diagnostic rules of peripheral lung cancer preliminary study based on data mining technique. *J. Nanjing Med. Univ.* **2007**, *21*, 190–195. [CrossRef]

26. Nuwangi, S.; Oruthotaarachchi, C.; Tilakaratna, J.; Caldera, H. Usage of association rules and classification techniques in knowledge extraction of diabetes. In Proceedings of the 2010 6th International Conference on Advanced Information Management and Service (IMS), Seoul, Korea, 30 November–2 December 2010; pp. 372–377.

27. Chen, Y.; Pedersen, L.; Chu, W.; Olsen, J. Drug exposure side effects from mining pregnancy data. *ACM SIGKDD Explor. Newsl.* **2007**, *9*, 22–29. [CrossRef]

28. Nahar, J.; Imam, T.; Tickle, K.S.; Chen, Y.P.P. Association rule mining to detect factors which contribute to heart disease in males and females. *Expert Syst. Appl.* **2013**, *40*, 1086–1093. [CrossRef] [PubMed]

29. Shmiel, O.; Shmiel, T.; Dagan, Y.; Teicher, M. Data mining techniques for detection of sleep arousals. *J. Neurosci. Methods* **2009**, *179*, 331–337. [CrossRef] [PubMed]

30. Brossett, S.; Sprague, A.; Jones, W.; Moser, S. A data mining system for infection control surveillance. *Methods Inf. Med.* **2000**, *39*, 303–310. [CrossRef] [PubMed]

31. Chen, T.J.; Chou, L.F.; Hwang, S.J. Application of a data-mining technique to analze coprescription patterns for antacids in Taiwan. *Clin. Ther.* **2003**, *25*, 2453–2463. [CrossRef]

32. Jensen, S. Mining medical data for predictive and sequential patterns. In Proceedings of the Fifth European Conference on Principles and Practice of Knowledge Discovery in Databases, Freiburg, Germany, 3–5 September 2001; pp. 1–10.

33. Dart, T.; Cui, Y.; Chatellier, G.; Degoulet, P. Analysis of hospitalised patient flows using data-mining. *Stud. Health Technol. Inform.* **2002**, *95*, 263–268.

34. Rossille, D.; Cuggia, M.; Arnault, A.; Bouget, J.; Le Beux, P. Managing an emergency department by analysing HIS medical data: A focus on elderly patient clinical pathways. *Clin. Med. J.* **2009**, *11*, 139–146. [CrossRef]

35. Lin, F.; Chou, S.; Pan, S.; Chen, Y. Mining time dependency patterns in clinical pathways. *Int. J. Med. Inform.* **2001**, *62*, 11–25.

36. Batal, I.; Fradkin, D.; Harrison, J.; Moerchen, F.; Hauskrecht, M. Mining recent temporal patterns for event detection in multivariate time series data. In Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Beijing, China, 12–16 August 2012; pp. 280–288.

37. Choi, K.; Chung, S.; Rhee, H.; Suh, Y. Classification and sequential pattern analysis for improving managerial efficiency and providing better medical service in public healthcare centers. *Healthc. Inform. Res.* **2010**, *16*, 67–76. [CrossRef]

38. Exarchos, T.P.; Papaloukas, C.; Lampros, C.; Fotiadis, D.I. Mining sequential patterns for protein fold recognition. *J. Biomed. Inform.* **2008**, *41*, 165–179. [CrossRef] [PubMed]

39. Ryan, G.W. What do sequential behavioral patterns suggest about the medical decision-making process?: Modeling home case management of acute illnesses in a rural Cameroonian village. *Soc. Sci. Med.* **1998**, *46*, 209–225. [CrossRef]

40. Lasker, G.E. Application of sequential pattern-recognition technique to medical diagnostics. *Int. J. Bio-Med. Comput.* **1970**, *1*, 173–186. [CrossRef]

41. Concaro, S.; Sacchi, L.; Bellazzi, R. Temporal data mining methods for the analysis of the AHRQ archives. *Proc. Am. Med. Inform. Assoc.* **2007**, *1–23*. Available online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.131.6417&rep=rep1&type=pdf (accessed on 2 July 2021).
42. Li, J.; Fu, A.; Fahey, P. Efficient discovery of risk patterns in medical data. *Artif. Intell. Med.* 2009, *45*, 77–89. [CrossRef]

43. Baralis, E.; Bruno, G.; Chiusano, S.; Domenici, V.C.; Mahoto, N.A.; Petrigl, C. Analysis of medical pathways by means of frequent closed sequences. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*; Springer: Berlin/Heidelberg, Germany, 2010; Volume 6278, pp. 418–425.

44. Antonelli, D.; Baralis, E.; Bruno, G.; Chiusano, S.; Mahoto, N.A.; Petrigl, C. Analysis of diagnostic pathways for colon cancer. *Flex. Serv. Manuf. J.* 2012, *24*, 379–399. [CrossRef]

45. Gotz, D.; Wang, F.; Perer, A. A methodology for interactive mining and visual analysis of clinical event patterns using electronic health record data. *J. Biomed. Inform.* 2014, *48*, 148–159. [CrossRef]

46. Huang, Z.; Lu, X.; Duan, H. On mining clinical pathway patterns from medical behaviors. *Artif. Intell. Med.* 2012, *56*, 35–50. [CrossRef]

47. Khaleel, M.A.; Pradhan, S.K.; Dash, G. Finding Locally Frequent Diseases Using Modified Apriori Algorithm. *Int. J. Adv. Res. Comput. Commun. Eng.* 2013, *2*, 3792–3797.

48. Pokharel, S.; Zuccon, G.; Li, Y. Representing EHRs with Temporal Tree and Sequential Pattern Mining for Similarity Computing. In Proceedings of the International Conference on Advanced Data Mining and Applications, Foshan, China, 12–14 November 2020; pp. 220–235.

49. Abawajy, J.; Kelarev, A.; Chowdhury, M. Multistage approach for clustering and classification of ECG data. *Comput. Methods Programs Biomed.* 2013, *112*, 720–730. [CrossRef]

50. Wang, J.; Liu, P.; She, M.F.H.; Nahavandi, S.; Kouzani, A. Biomedical time series clustering based on non-negative sparse coding and probabilistic topic model. *Comput. Methods Programs Biomed.* 2013, *111*, 629–641. [CrossRef]

51. Rani, S.; Kautish, S. Association Clustering and Time Series Based Data Mining in Continuous Data for Diabetes Prediction. In Proceedings of the 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 14–15 June 2018; pp. 1209–1214.

52. Zriqat, I.A.; Altamimi, A.M.; Azzeh, M. A Comparative Study for Predicting Heart Diseases Using Data Mining Classification Methods. *arXiv 2017*, arXiv:1704.02799.

53. Sufi, F.; Khalil, I.; Mahmood, A.N. A clustering based system for instant detection of cardiac abnormalities from compressed ECG. *Expert Syst. Appl.* 2011, *38*, 4705–4713. [CrossRef]

54. Mahoto, N.A.; Shaikh, F.K.; Ansari, A.Q. Exploitation of Clustering Techniques in Transactional Healthcare Data. *Mehran Univ. Res. J. Eng. Technol.* 2014, *33*, 77–92.

55. Chaurasia, V.; Pal, S.; Tiwari, B. Prediction of benign and malignant breast cancer using data mining techniques. *J. Algorithms Comput. Technol.* 2018, *12*, 119–126. [CrossRef]

56. Buczak, A.L.; Moniz, L.J.; Feighner, B.H.; Lombardo, J.S. Mining electronic medical records for patient care patterns. In Proceedings of the IEEE Symposium on Computational Intelligence and Data Mining (CIDM), Nashville, TN, USA, 30 March–2 April 2009; pp. 146–153.

57. Karegowda, A.G.; Jayaram, M.; Manjunath, A. Cascading k-means clustering and k-nearest neighbor classifier for categorization of diabetic patients. *Int. J. Eng. Adv. Technol.* 2012, *1*, 147–151.

58. Hirano, S.; Sun, X.; Tsumoto, S. Comparison of clustering methods for clinical databases. *Inf. Sci.* 2004, *159*, 155–165. [CrossRef]

59. Isken, M.; Rajagopalan, B. Data mining to support simulation modeling of patient flow in hospitals. *J. Med. Syst.* 2002, *26*, 179–197. [CrossRef] [PubMed]

60. Wang, J.; Han, J.; Pei, J. Closet+: Searching for the best strategies for mining frequent closed itemsets. In Proceedings of the Ninth International Conference on Knowledge-Based and Intelligent Information and Engineering Systems; Springer: Berlin, Germany, 2002; Volume 2434, pp. 77–89. [CrossRef]

61. Juang, B.H.; Rabiner, L. The segmental K-means algorithm for estimating parameters of hidden Markov models. *IEEE Trans. Acoust. Speech Signal Process.* 1990, *38*, 1639–1641. [CrossRef]

62. Ester, M.; Kriegel, H.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, OR, USA, 2–4 August 1996; Volume 1996, pp. 226–231. [CrossRef]

63. Xia, H.; Wang, S.; Yoshida, T. A modified ant-based text clustering algorithm with semantic similarity measure. *J. Syst. Sci. Syst. Eng.* 2006, *15*, 474–492. [CrossRef]

64. Rousseuw, P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 1987, *20*, 53–65. [CrossRef]

65. Sharan, R.; Maron-Katz, A.; Shamir, R. CLICK and EXPANDER: A system for clustering and visualizing gene expression data. *Bioinformatics* 2003, *19*, 1787–1799. [CrossRef]

66. Szathmary, L. Symbolic Data Mining Methods with the Coron Platform. Ph.D. Thesis, University Henri Poincare, Nancy, France, 2006.

67. Fournier-Viger, P.; Gomariz, A.; Soltani, A.; Lam, H.; Gueniche, T. SPMF: A Sequential Pattern Mining Framework. 2014. Available online: http://www.philippe-fournier-viger.com/spmf/ (accessed on 2 July 2021).

68. Audet, A.M.; Greenfield, S.; Field, M. Medical practice guidelines: Current activities and future directions. *Ann. Intern. Med.* 1990, *113*, 709–714. [CrossRef] [PubMed]

69. Wang, J.; Han, J. BIDE: Efficient Mining of Frequent Closed Sequences. In Proceedings of the 20th International Conference on Data Engineering (ICDE ’04), Boston, MA, USA, 2 April 2004; pp. 79–90.
70. Rapid Miner Project. The Rapid Miner Project for Machine Learning. 2013. Available online: http://rapid-i.com/ (accessed on 2 July 2021).
71. Dierk, S.F. The SMART retrieval system: Experiments in automatic document processing Gerard Salton, Ed. (Englewood Cliffs, N.J.: Prentice Hall, 1971, 556 pp., $15.00). IEEE Trans. Prof. Commun. 1972, PC-15, 17. [CrossRef]