Journal of Artificial Intelligence and Data Mining (JAIDM), Vol. 11, No. 3, 2023, 343-355.

Shahrood University of Technology

Journal of Artificial Intelligence and Data Mining (JAIDM)
Journal homepage: http://jad.shahroodut.ac.ir

Research paper

Hidden Pattern Discovery on Clinical Data: an Approach based on Data Mining Techniques

Meysam Roostaee[1*] and Razieh Meidanshahi[2]

1. Department of Computer Engineering, University of Mazandaran, Babolsar, Iran.
2. Department of Computer Engineering, Polytechnic University of Turin, Turin, Italy.

Article Info

Article History:
Received 27 March 2023
Revised 14 May 2023
Accepted 01 July 2023

DOI:10.22044/jadm.2023.12855.2435

Keywords:
Clinical Data, Data Mining, Unsupervised Learning, Association Rule Mining, Clustering.

**Corresponding author: m.roostaee@umz.ac.ir (M. Roostaee).*

Abstract

In this study, we sought to minimize the need for redundant blood tests in diagnosing common diseases by leveraging unsupervised data mining techniques on a large-scale dataset of over one million patients' blood test results. We excluded non-numeric and subjective data to ensure precision. In order to identify relationships between attributes, we applied a suite of unsupervised methods including preprocessing, clustering, and association rule mining. Our approach uncovered correlations that enable healthcare professionals to detect potential acute diseases early, improving patient outcomes, and reducing costs. The reliability of our extracted patterns also suggests that this approach can lead to significant time and cost savings while reducing the workload for laboratory personnel. Our study highlights the importance of big data analytics and unsupervised learning techniques in increasing efficiency in healthcare centers.

1. Introduction

The blood test is usually essential for diagnosing diseases in the early period. By a physician's order, several blood attributes in hormone, hematology, immunology, and chemistry are typically tested. Since materials and devices that are used in these experiments are expensive, these laboratory experiments are costly. Moreover, in some countries such materials are scarce; hence, high costs should be paid to import them from other countries. It is also possible to look at the problems caused by this issue from another aspect. Since physicians spend a lot of effort for interpreting these test results, their workload increases dramatically. The workload of the laboratory and routine test interpretation are the leading causes of fatigue, which, due to human limitations, can cause errors in the interpretation of blood test results [1]. Furthermore, some diagnostic and laboratory procedures are invasive, costly, and painful for patients. When we look back to the vast number of tests that have been done in all the years since databases have emerged in medical science, in some areas, we can see abandoned volumes of data stored in healthcare data repositories without any analysis usage. Furthermore, the blood test is invasive for almost every person, so doing blood experiments is painful. In fact, in nearly every case, laboratories are full of unnecessary blood samples. In contrast, some other emergency samples need to be examined immediately because they warn about severe illnesses in individuals. The researchers have shown that physicians always complain about the lack of a reliable decision-making system that could help them in their routine duties [2, 3]. According to WHO's statistics, more than 20% of emergencies can lead to case fatality due to the lack of an on-time hidden disease symptoms diagnosis [4]. Common diseases such as chronic kidney disease are the most dangerous illnesses with a straightforward relationship with diabetes that need early symptoms diagnosis; otherwise, they could be the main reason for the patient's death [5, 6].

A wide range of clinical diagnostic service tools is required to help physicians in their daily decision-making; in this case, laboratory services are used to diagnose and investigate disease. The cost of these services increases every year. This process is an excellent opportunity for equipment companies to market their products in medical centers and encourages physicians to do unnecessary tests, which leads to crowded medical centers. Consequently, individuals and the health ministry should afford excessive health expenses to import such equipment and materials. Recently, it has become an essential concern for such countries which suffer from the lack of resources for medical services [7]. Since there is no exact formula for these everyday experimental attributes, physicians have to do these repetitive tests in most cases. Thus, a precise procedure is needed to help physicians select the most appropriate attributes with the least cost and improve diagnostic performance.

This paper applied data mining techniques to medical data, which have not received enough attention. We will be looking for useful knowledge from the blood test result of patients. To do that, we employed clustering and association rule mining techniques. The clustering algorithm helps to find the relationship between different test results. It is performed by grouping normal (/abnormal) results in various clusters. The association rule mining approach is used to find valuable and interesting rules from test results. The experts then investigate the information to ensure the findings are correct and can be used in the future. The high confidence score of the extracted patterns leads to time and cost savings and reduces laboratory personnel's workloads.

The reminder of the paper is organized as what follows. Section 2 is devoted to related work in the field of pattern discovery in clinical data. Section 3 explains in detail the methodology, which contains the dataset, inclusion and exclusion criteria, and statistical analysis approaches. Then the results are shown in Section 4, and a discussion of the results is provided in Section 5. Finally, in Section 6, we conclude and present suggestions for future work.

2. Related Works

The application of data mining techniques in medical science can significantly improve healthcare services [8–10]. With the development of information and communications technology, a massive amount of data related to the blood test results of patients is stored in data stores and repositories [11]. Data mining has different algorithms to extract patterns and knowledge to help better decision-making [12–14]. Physicians have recently paid more attention to data mining and machine learning techniques to analyze the clinical data. Most studies have focused on classifying the input data into several predefined categories.

Wu *et al*. [15] utilized data mining techniques to predict the prognosis of patients suffering from rheumatoid arthritis. The study found that data mining algorithms were more effective than conventional statistical analysis methods in predicting the likely progression of the disease in these patients. This research work contributes to an expanding body of evidence supporting the potential of data mining approaches for improving the accuracy of medical prognoses, which could ultimately lead to better treatment outcomes for patients with rheumatoid arthritis.

Barrios *et al*. [16] developed several models using data mining techniques to diagnose metabolic syndrome. Their findings suggested that employing an artificial neural network with three hidden layers is a highly effective approach for recognizing the syndrome and can significantly decrease the time needed to start treatment.

Begum *et al*. [17] use data mining techniques in healthcare to diagnose thyroid diseases early. They employ various classification techniques to predict thyroid disease accurately. Their study highlights the potential for data mining to improve healthcare decision-making and patient outcomes for thyroid diseases. Thallam *et al*. [18] also developed prediction models using various data mining techniques to predict the presence of lung cancer at an early stage in a patient. This study compares different classification and ensemble models such as support vector machine, K-nearest neighbor, random forest, and voting classifier.

The paper by Ayyoubzadeh *et al*. [19] presents a comparative analysis of various data mining techniques for predicting the length of stay (LOS). The study identifies the most important factors affecting LOS as the number of para-clinical services, counseling frequency, clinical ward, doctor's specialty and degree, and cause of hospitalization. Nine classifiers were applied with and without feature selection techniques to predict the LOS, and the results showed that most models are suitable for classification, although Logistic Regression has slightly better performance in terms of accuracy.

Some approaches have applied association rule mining strategy to extract knowledge from medical data. Lakshmi and Vadivu in [20] discuss the growing interest in disease comorbidity prediction and the development of analytical tools to explore disease associations and comorbidity patterns using electronic health records (EHRs) and biological data. The study proposes a novel approach based on weighted association rule mining for predicting disease comorbidities by combining clinical data and molecular data. The results demonstrate that this system outperforms existing systems in disease comorbidity prediction, highlighting the potential of integrating multiple types of data for improved disease prediction techniques in the medical field.

Jayasri and Aruna [44] analyzed a medical dataset of diabetes patients to discover the association between diseases and their symptoms by a fusion of attention network and association rule mining. They used a map-reduce framework to implement the model. Alaiad *et al*. [45] integrated association rule mining and classification techniques to construct a system for predicting and diagnosing chronic kidney disease. Their results showed that applying an integrative approach using a combination of classification and association rule mining algorithms can improve the prediction of chronic kidney disease.

Luo *et al*. [21] applied data mining techniques to identify potential herbal pairs for treating COVID-19. Specifically, they utilized association rule mining on clinical data to identify high frequency combinations of herbs that have demonstrated efficacy in treating COVID-19 symptoms. Through their analysis, they identified Gancao-Banxia as a potential herbal pair with promising therapeutic effects for COVID-19.

Various other approaches based on data mining techniques have also been applied to clinical data [23, 46] to reduce the challenge of medical field problems, as these problems are complex that require more information and knowledge to address effectively.

Our approach to analyzing clinical data differs from most existing methods in several important ways. Firstly, we use an unsupervised learning approach, which does not require pre-labeled data or prior knowledge of potential relationships among variables. Instead, our model autonomously identifies hidden relationships among blood attributes in patients' test results. By focusing specifically on the blood test results, we are able to extract meaningful information related to underlying health conditions and potential risk factors. This is different from many other related works, which may use different types of data (such as medical images or electronic health records) or rely on supervised learning approaches. One key benefit of using an unsupervised approach is that our model is able to uncover previously unknown relationships among blood attributes. These relationships have the potential to inform the development of a decision support system, which could help human experts make more informed decisions based on a broader range of insights.

Overall, our approach represents a departure from traditional methods for analyzing clinical data. By prioritizing unsupervised learning and focusing on blood test results, we are able to uncover hidden relationships and provide new insights into patient health that could ultimately improve decision-making and patient outcomes.

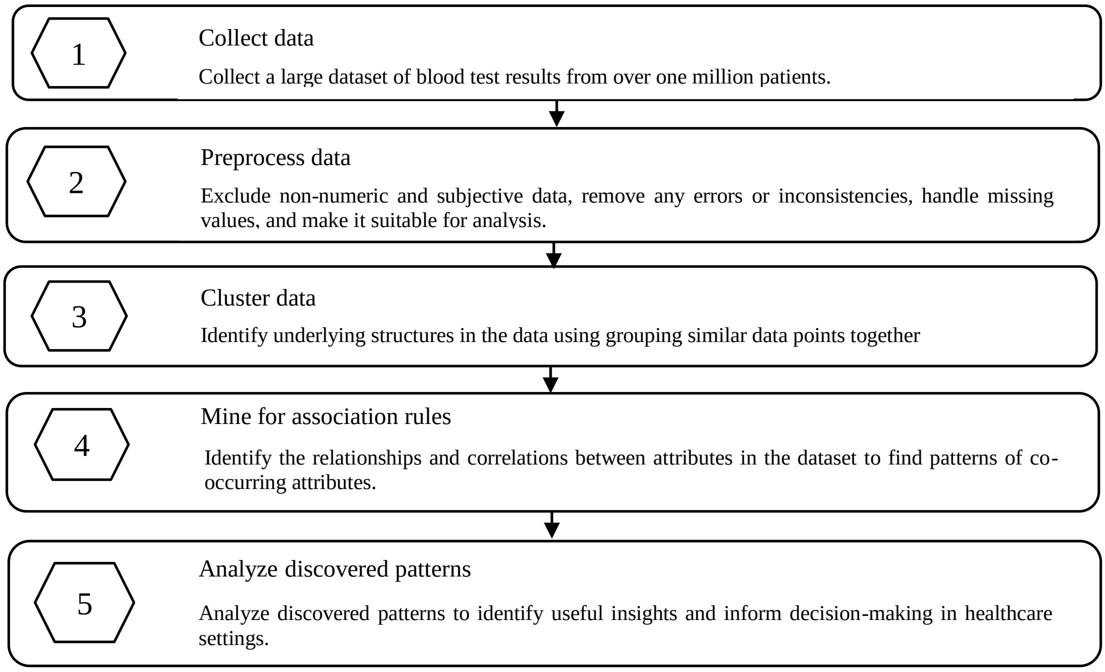

1
Collect data
Collect a large dataset of blood test results from over one million patients.
2
Preprocess data
Exclude non-numeric and subjective data, remove any errors or inconsistencies, handle missing values, and make it suitable for analysis.
3
Cluster data
Identify underlying structures in the data using grouping similar data points together
4
Mine for association rules
Identify the relationships and correlations between attributes in the dataset to find patterns of co-occurring attributes.
5
Analyze discovered patterns
Analyze discovered patterns to identify useful insights and inform decision-making in healthcare settings.

Figure 1. Overall workflow of the proposed approach.

3. Methodology

In this section, we first provide an introductory overview of the dataset used in this study. Following this, we expound upon the criteria and statistical methods that have been employed to ensure a comprehensive and rigorous analysis of the dataset. The unsupervised data mining techniques are used for preprocessing, clustering, and association rule mining on the accumulated data to extract meaningful relationships between dataset attributes. The workflow of this study is depicted in Figure 1.

3.1. Dataset

This study was based on about 1048576 blood test results from Motahari healthcare center. The Motahari healthcare center is a university-affiliated outpatient clinic and the most preferred clinic in southern Iran, with more than 1000 blood samples to be tested daily.

Data was prepared to be standard for analysis by Microsoft SQL Server. First, data cleaning was performed in consultation with physicians to remove errors, incorrect results, and additional characters. After that, the final data table was converted to a horizontal one for better detailed analysis.

According to the data assessment, some attributes were repetitive in the data table. They were attributes chosen by physicians to diagnose common diseases. Table 1 shows the list of attributes in the used dataset. In this research work, only independent attributes in blood test results were kept, and duplicate and correlated features were excluded. As a result, 16 blood attributes were used for evaluation (Table 2). The details of these attributes are provided in Table 2.

3.2. Inclusion and exclusion criteria

The chosen blood attributes are generally categorized into hormone, hematology, immunology, and chemistry groups. Since some attributes were non-numerical (i.e. categorical), they were considered subjective [24], meaning that each pathologist interprets the result from his own perspective. Thus in consultation with pathology experts, only numerical results were chosen.

As it is shown in Table 2, each blood attribute has a normal range. For example, the normal range of Creatinine is between 0.6 and 1.3. If the result is below this range, like 0.3, it is considered Low; otherwise, if the result is above the range, like 2, it is in the High range, and in this case, both the low and high ranges are in abnormal results. The Normal range was specified in consultation with pathology experts of Motahari center.

The attributes in the hormone group were used to diagnose thyroid disease. The hematology group is related to anemia disorders, and body immune disorders are considered in the immunology group. Common diseases are renal, liver disorders, and diabetes.

Table 1. List of attributes in the used dataset.

Field name	Description	Field name	Description
FNumber	File number	Age	Age
SEX	Sex	AgeType	Age type
TSH	Thyroid stimulating hormone	TSH-range	TSH result range
TG	Triglycerides	TG-range	TG result range
Chol	Cholesterol	Chol-range	Chol result range
Hb	Hemoglobin	Hb-range	Hb result range
FBS	Fast blood sugar	FBS-range	FBS result range
ANA	Anti-nuclear antibody	ANA-range	ANA result range
AntiDNA	Double-stranded DNA	AntiDNA-range	AntiDNA result range
AST	Aspartate aminotransferase	AST-range	AST result range
ALT	Alanine transaminase	ALT-range	ALT result range
ALKP	Alkaline phosphatase	ALKP-range	ALKP result range
PTH	Parathyroid hormone	PTH-range	PTH result range
Ca	Calcium	Ca-Range	Ca result range
Ph	Phosphorus	Ph-range	Ph result range
BUN	Blood urea nNitrogen	BUN-range	BUN result range
Cr	Creatinine	Cr-range	Cr result range
Albumin	Albumin	Albumin-range	Albumin result range

Table 2. Range of different blood attributes.

Test name	Group name	Normal range
TSH	Hormone	0.36-6.3
PTH	Hormone	8.3-68
Hb	Hematology	Female: 12-15.6 Male: 13.5-17.5
ANA	Immunology	x←result x < 10 Negative x > 10 Positive
AntDNA	Immunology	x←result x < 40 Negative x > 50 Positive 40 < x < 50 border line
TG	Chemistry	50-150
Chol	Chemistry	125-200
FBS	Chemistry	70-99
AST	Chemistry	1-46
ALKP	Chemistry	80-290
ALT	Chemistry	1-49
Calcium	Chemistry	8.5-10.5
Phosphorus	Chemistry	2.7-4.5
BUN	Chemistry	8-23
Creat	Chemistry	0.6-1.3
Albumin	Chemistry	3.5-5

3.3. Statistical analysis

The primary statistical analysis used in this study are preprocessing, Clustering and association rule mining, which we will describe in the following sub-sections. The variables without values are called missing values. Missing value handling is essential because missing data can lead to inaccurate analysis and modelling, resulting in biased or unreliable results.

The clinical nature of the patients was unknown based on the information in the dataset. Each row did not have any labels. Moreover, there was no output variable to predict the results or classify attributes. Therefore, there was no learning from cases that could be specified as an outcome variable. According to all these characteristics, this dataset was unsupervised; we need to employ unsupervised techniques to analyze such a dataset. It should be noted that the most popular data mining techniques for unsupervised data are clustering and association rule mining.

In this study, the statistical analysis was performed by the SPSS software system.

3.3.1. Missing value handling

Since patients who came to the Motahari clinic did not have to check all 16 attributes, the dataset had missing values. The researchers have found several techniques to complete the missing attributes in the dataset. Different kinds of techniques are used due to the nature of missing values in the dataset [25, 26].

This paper used Multiple Imputation (MI) to complete missing data. The basic idea behind MI is to simulate multiple plausible values for each missing data point based on the observed data. This creates several complete datasets where each dataset contains imputed values for the missing data points. The imputation process typically involves replacing missing values with plausible values based on other available information from the dataset. Mean imputation is one common approach to impute missing data, but many more sophisticated methods are available [28].

After imputation, an analysis is performed on each completed dataset separately using standard statistical methods. The results from all the analyses are then combined through a process known as pooling, which allows for the uncertainty due to missing data to be incorporated into the final set of estimates [27, 29].

The key advantage of MI is that it allows for the uncertainty due to missing data to be incorporated into the analysis. MI assumes that the missing data are Missing at Random (MAR), which means that the probability of a value being missing depends only on observed data, not on unobserved data [42, 43].

3.3.2. Clustering

Clustering methods group similar objects (data tuples) according to their space distance into the same cluster. The maximum or average distance between cluster objects measures the clustering quality [30, 37]. One of the most famous clustering algorithms is called k-means that tries to puts objects into k groups based on their features. It has persistent steps to minimize the sum of squares of distances between objects and cluster centroid [31]. Considering that the k-means algorithm is employed for unsupervised learning, the idea is that we use it to group normal results as well as abnormal results.

The k in this algorithm indicates the number of clusters that must be determined before any clustering take place. The works that have been done in specifying the number of k on real applications of clustering methods have shown that the "random selection" method is better than methods that include machine learning algorithms

[32, 33]. Since the random selection method does not require prior knowledge, it is simple. Moreover, the approach covers the solution space, and no computational process is needed. Consequently, random selection is used to specify a k value in the k-means algorithm to cluster data. However, in order to minimize the chances of error in choosing the value for k with random selection, we validated results with various values of k with domain experts including pathologists to find the most accurate and meaningful clustering outcomes.

In more detail, the first step involved generating k-clusters using the k-means algorithm on the entire dataset. Following this, domain experts thoroughly examined these clusters and selected those that contained the most meaningful and medically relevant attributes. From among these clusters, the experts identified the most reliable features which were then utilized to refine our analysis. Consequently, the final clusters exclusively comprise the most relevant and insightful attributes as discussed in Section 5.

3.3.3. Association rule mining

Association rule mining is another data mining technique that can be used to obtain frequent patterns and valuable rules from transactional data [34, 35]. We represent association rules as X⇒Y, where X and Y are sets of test attributes (For example, High-Cholesterol (h-ch) and High-Triglycerides (h-tg)). This technique measures the co-occurrence of X and Y and the rules that has strong interestingness, i.e. the relationship between X and Y is frequently satisfied, are considered as useful rule [36].

In this study, we employ the Apriori algorithm the most well-known association rule mining algorithm. This method uses support and confidence as the interestedness measures. The support rate, which is the fraction of transactions that contains both X and Y is defined as follows:

$$sup(X \Rightarrow Y) = \frac{sup(X \cup Y)}{|T|} \qquad (1)$$

And confidence of a rule is defined as follows:

$$conf(X \Rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)} \qquad (2)$$

where $sup(X)$ refers to the number of transactions contain X, and $sup(X \cup Y)$ is the number of transactions containing both X and Y. This equation gives the maximum likelihood of $P(Y|X)$. An item set can be used to generate a rule when it holds minimum support and minimum confidence specified by the user. However, specifying constraints for support and confidence is an issue for the user to extract valuable and meaningful rules.

When so many rules are extracted from the dataset, a high support constrain avoids the explosion of frequent items. However, it may lose valuable patterns with low support, which have high confidence. It should be noted that rules with high support are often well-known in the area, while knowledge is the underlying rare high-confidence rules that need to be mined [38].

To ensure that valuable patterns emerge, we should choose the lowest values for support. However, this strategy results in rule explosion. We can handle this issue using another measure named lift or interest. Therefore, each rule should have a positive lift. The method effectively finds rules with low support and high confidence and reduces imprecise rules [39].

Given an association rule $X \rightarrow Y$, where X is called the antecedent and Y is called the consequent, the lift measure is calculated as follows:

$$lift(X \rightarrow Y) = \frac{P(X \cup Y)}{P(X).P(Y)} = \frac{conf(X \rightarrow Y)}{sup(Y)} \qquad (3)$$

There are three different possible values for lift [40, 41]:

I. If the calculated lift value is less than 1, it means that X and Y are negatively correlated (i.e. the occurrence of X has a negative effect on the occurrence of Y).

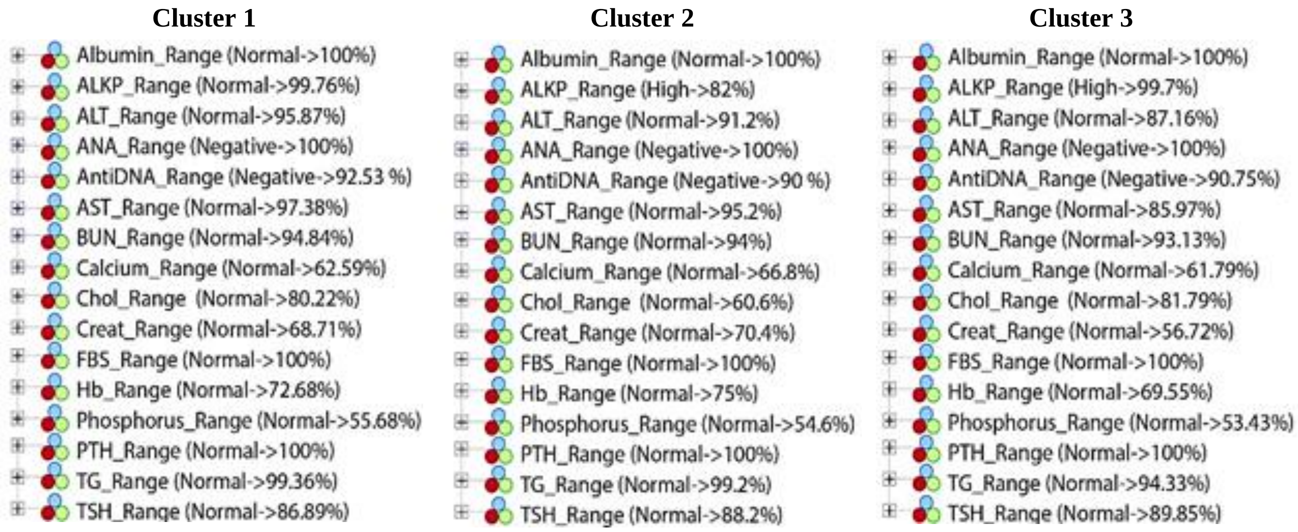
II. If the calculated lift value is 1, then X and Y are independent.

If the calculated lift value exceeds 1, X and Y are positively correlated (i.e. X and Y appear together frequently).

4. Results

This section focuses on the results obtained from applying clustering and association rule mining techniques. The clusters and rules were generated by utilizing both normal and abnormal ranges of values for various attributes. Although some mixed clusters or rules were detected, we discovered that those based solely on normal or abnormal values were more dependable. This was confirmed by experts' opinions as well as interestingness measures. Consequently, only the significant rules and clusters approved by experts are presented.

Four different groups of results are shown in the following. The first two results are related to clustering of normal and abnormal results, the rest are related to the patterns discovered by association rule mining technique.



Cluster 1

Albumin_Range (Normal->100%)
ALKP_Range (Normal->99.76%)
ALT_Range (Normal->95.87%)
ANA_Range (Negative->100%)
AntiDNA_Range (Negative->92.53 %)
AST_Range (Normal->97.38%)
BUN_Range (Normal->94.84%)
Calcium_Range (Normal->62.59%)
Chol_Range (Normal->80.22%)
Creat_Range (Normal->68.71%)
FBS_Range (Normal->100%)
Hb_Range (Normal->72.68%)
Phosphorus_Range (Normal->55.68%)
PTH_Range (Normal->100%)
TG_Range (Normal->99.36%)
TSH_Range (Normal->86.89%)

Cluster 2

Albumin_Range (Normal->100%)
ALKP_Range (High->82%)
ALT_Range (Normal->91.2%)
ANA_Range (Negative->100%)
AntiDNA_Range (Negative->90 %)
AST_Range (Normal->95.2%)
BUN_Range (Normal->94%)
Calcium_Range (Normal->66.8%)
Chol_Range (Normal->60.6%)
Creat_Range (Normal->70.4%)
FBS_Range (Normal->100%)
Hb_Range (Normal->75%)
Phosphorus_Range (Normal->54.6%)
PTH_Range (Normal->100%)
TG_Range (Normal->99.2%)
TSH_Range (Normal->88.2%)

Cluster 3

Albumin_Range (Normal->100%)
ALKP_Range (High->99.7%)
ALT_Range (Normal->87.16%)
ANA_Range (Negative->100%)
AntiDNA_Range (Negative->90.75%)
AST_Range (Normal->85.97%)
BUN_Range (Normal->93.13%)
Calcium_Range (Normal->61.79%)
Chol_Range (Normal->81.79%)
Creat_Range (Normal->56.72%)
FBS_Range (Normal->100%)
Hb_Range (Normal->69.55%)
Phosphorus_Range (Normal->53.43%)
PTH_Range (Normal->100%)
TG_Range (Normal->94.33%)
TSH_Range (Normal->89.85%)

Figure 2. Clustering results in normal range.

4.1. Clusters of normal results

Initializing the value for k had been done repetitively, and physicians checked the results of each. The best results, according to physicians' confirmation, are shown in Figure 2. As it can be seen, a normal relationship between some blood attributes emerged. These three clusters illustrate which normal attributes can occur together.

4.2. Clusters of abnormal results

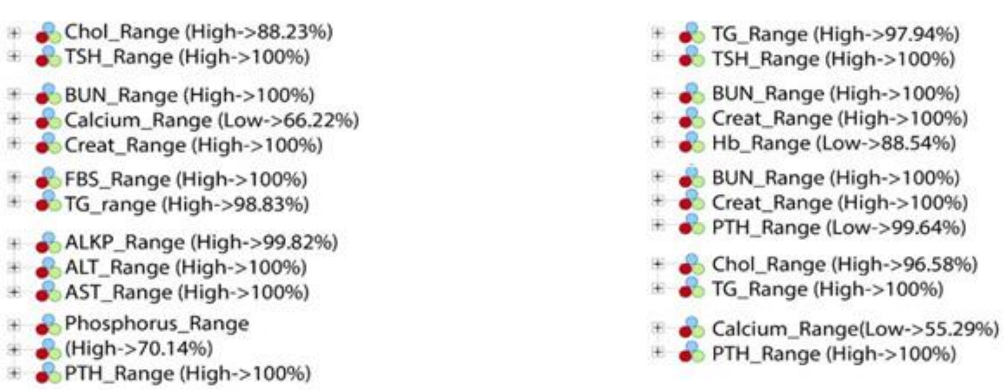
According to repetitively k selection for detailed analysis of the abnormal results, the most meaningful result in physicians' opinions was $k = 10$. As shown in Figure 3, some blood attributes are in one cluster, meaning there is a close relationship between them. Since each attribute is an indicator of a disorder in the human body, clusters indicate the occurrence of an anomaly as the side effect of the disease.

4.3. Association between normal results

In consultation with pathologist experts, association rule mining was done by $min\_sup = 40\%$ and $min\_conf = 50\%$; many rules were extracted from the dataset. Worthwhile relationships were chosen from the rules. The rules with a high percentage of support and confidence confirmed our results in the normal clusters. Table 3 shows that when an attribute is in a normal range, other attributes are in a normal range too.

4.4. Association between abnormal results

As noted, when min_sup is too high, useful rules are lost, especially in unsupervised datasets. Therefore, the results of the occurrence of diseases are much less than normal. Once high min_sup is chosen, such valuable abnormal rules would be lost. In this case, researchers suggest using lift measure in addition to min_sup and min_conf. As a result, when an attribute is in the abnormal range, it can be concluded that other attributes are also not normal. Abnormal rules extracted by association rule mining confirmed the abnormal clusters mentioned in the clusters of abnormal results (Table 4). Pathologists confirmed the valuable rules. Each cluster was meaningful when association rule mining was performed by lower min_sup and min_conf.



Cluster 4

Chol_Range (High->88.23%)
TSH_Range (High->100%)

BUN_Range (High->100%)
Calcium_Range (Low->66.22%)
Creat_Range (High->100%)

FBS_Range (High->100%)
TG_range (High->98.83%)

ALKP_Range (High->99.82%)
ALT_Range (High->100%)
AST_Range (High->100%)

Phosphorus_Range
(High->70.14%)
PTH_Range (High->100%)

Cluster 5

TG_Range (High->97.94%)
TSH_Range (High->100%)

BUN_Range (High->100%)
Creat_Range (High->100%)
Hb_Range (Low->88.54%)

BUN_Range (High->100%)
Creat_Range (High->100%)
PTH_Range (Low->99.64%)

Chol_Range (High->96.58%)
TG_Range (High->100%)

Calcium_Range(Low->55.29%)
PTH_Range (High->100%)

Figure 3. Clustering results in abnormal range.

Table 3. Association rules in normal range.

Rule number	Antecedent	Consequent	Support%	Confidence%
1	TSH_Range = Normal	TG_Range = Normal	71.78	87.45
2	TSH_Range = Normal	Chol_Range = Normal	78.35	87.25

3	BUN_Range = Normal
	Cr_Range = Normal
		Hb_Range = Normal	68.23	93.91

4	BUN_Range = Normal
	Cr_Range = Normal
		Calcium_Range = Normal	62.22	93.77

5	BUN_Range = Normal
	Cr_Range = Normal
		PTH_Range = Normal	80.96	93.81

6	FBS_Range = Normal	TG_Range=Normal	58.26	73.53
7	ANA_Range = Negative	AntiDNA_Range = Negative	47.23	91.32

8	AST_Range = Normal
	ALT_Range = Normal
		ALKP_Range = Normal	99.26	76.88

9	PTH_Range=Normal	Calcium_Range = Normal	60.50	62.35
10	PTH_Range = Normal	Phosphorus_Range = Normal	58.49	54.30
11	TG_Range = Normal	Chol_Range=Normal	71.75	83.84

Table 4. Association rules in abnormal range.

Rule number	Antecedent	Consequent	Support%	Confidence%	Lift
12	TSH_Range = High	TG_Range = High	4	97.941	24.25
13	TSH_Range = High	Chol_Range = High	4	88.231	22.05
14	BUN_Range = High Cr_Range = High	Hb_Range = Low	3	88.536	29.51
15	BUN_Range = High Cr_Range = High	Calcium_Range = Low	3	66.216	22.07
16	BUN_Range = High Cr_Range = High	PTH_Range = High	3	99.635	33.21
17	FBS_Range = High	TG_Range = High	49	98.834	2.01
18	AST_Range=High ALT_Range = High	ALKP_Range = High	13	99.816	7.67
19	PTH_Range = High	Calcium_Range = Low	55	76.03	1.38
20	PTH_Range = High	Phosphorus_Range = High	70	70.135	1.08
21	TG_Range = High	Chol_Range = High	27	96.583	3.57

5. Discussion

The current section discusses the main findings from Figures 1, 2 and Tables 3, 4. The obtained results validate the applicability of the data mining techniques in this field. In the following, we refer to rule number i as R_i.

5.1. Relationship between TSH and TG

Generally, hypothyroidism causes metabolism reduction. In this case, metabolism reduction leads to high triglycerides. The process demonstrates a straightforward relationship between the primary and secondary events. Accordingly, as R_{12} (shown in Table 4), when TSH is in the high confidence range of 97.94%, physicians can be sure that this event leads to a high degree of TG in one's body (Figure 4). The application of this rule is in the diagnosis of hypothyroidism and hypertriglyceridemia. Moreover, as R_1 (shown in Table 3), when TSH is in the normal range by the confidence of 87.45%, the TG is also in the normal range.

5.2. Relationship between TSH and Chol

Figure 5 demonstrates that the high Chol and high TSH relationship is stronger than the relationship between when these attributes are in low ranges. Moreover, a high TSH range in the body causes an increased range of cool by the confidence of 88.23% (R_{13}). Besides, by 87.25% of confidence, when TSH is in the normal range, Chol is also in the normal range (R_2). Experimental studies have shown that excessive secretion of TSH leads to a dramatic reduction of metabolism, which is one of the hypothyroidism side effects.

5.3. Relationship between TG and Chol

Hereditary disorders, poor nutrition, and body organ disorders are essential factors in a blood attribute's abnormal range. As it is shown in

Figure 6, TG and Chol changes occur simultaneously. Tables 3 and 4 indicate that when TG is in the high degree of confidence of 96.58%, Chol is also in the high range (R_{21}). In addition, by the confidence of 83.84%, these two attributes' normal ranges occur simultaneously as well **(**R_{11}**).**

5.4. Relationship between FBS and TG

Blood sugar is a critical factor that reduces the body's metabolism and increases blood lipids such as TG. This paper demonstrated that if the FBS level rises in the blood by a confidence of 98.83%, the TG will be in the high range (Figure 7 and R_{17}). Furthermore, the other relationship between these two attributes is that by the confidence of 75.53%, a normal range of FBS leads to a TG normal range (R_6).

5.5. Relationship between BUN, Cr, and Hb

Figure 8 illustrates that renal failure influences enzymatic metabolism mechanisms, including Protein synthesis disruption in the body. Consequently, due to renal failure, the body's waste such as blood urea nitrogen and creatinine increases marginally. Unfortunately, the process causes Hematopoietic System failure. In this case, the hemoglobin level declines in the blood. This rule is worthwhile for renal failure and anemia diagnosis. By the confidence of 88.54%, a high range of BUN and Cr causes Hb reduction in the body (R_{14}). Also when BUN and Cr are in a normal range by the confidence of 93.91%, Hb is also in the normal range (R_3).

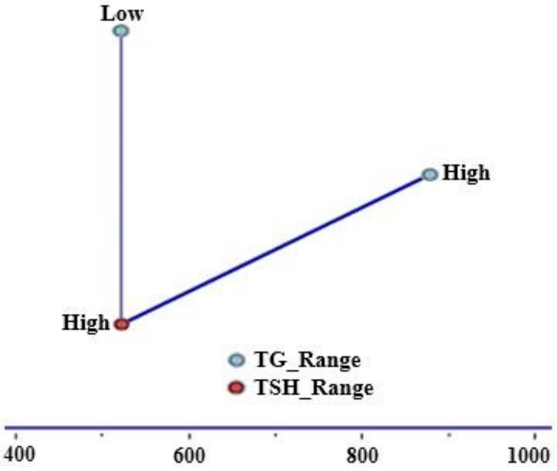


Figure 4. Link chart of TG and TSH.

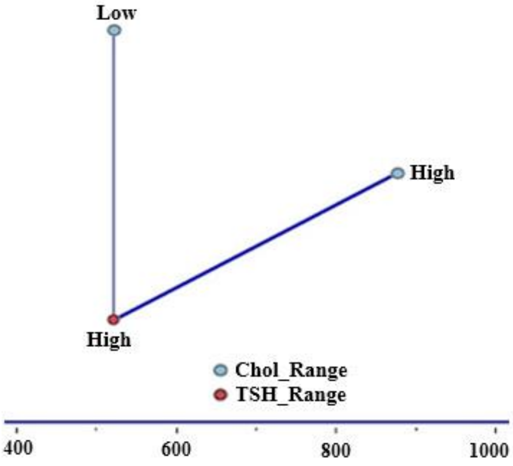


Figure 5. Link chart of TSH and Chol.

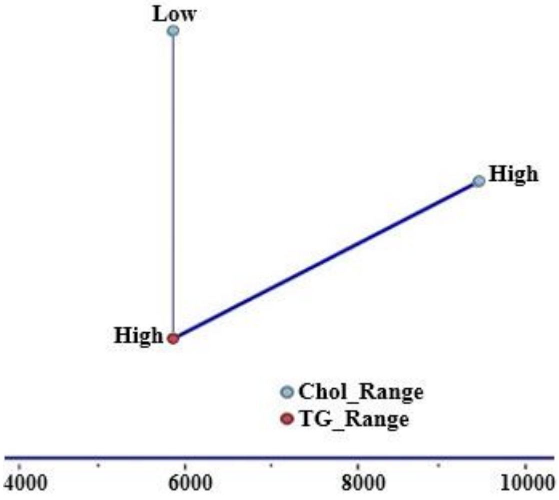


Figure 6. Link chart of Chol and TG.

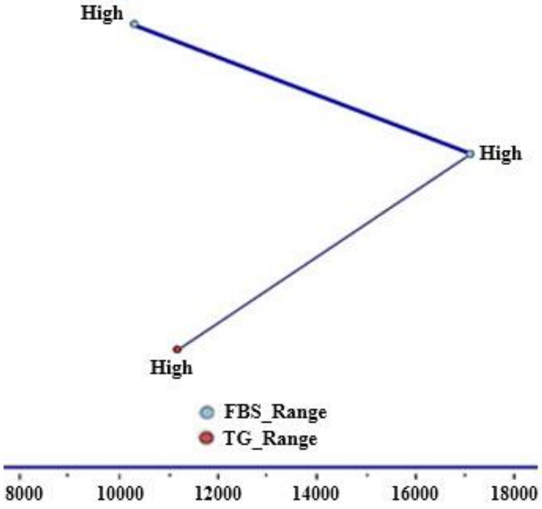


Figure 7. Link chart of FBS and TG.

5.6. Relationship between BUN, Ca, and Cr

Renal failure has many dangerous side effects in patients. When renal fail, H+ ions (acid) will increase, which leads to calcium decreasing and blood's Ph increasing (Figure 9). The rule is helpful for physicians in order to diagnosing renal failure and hypocalcemia. By the confidence of 66.22%, the high range of BUN and Cr causes the low Cr range (R_{15}). Moreover, by the confidence of 93.77%, when BUN and Cr are in the normal range, the Ca is normal, too (R_4).

5.7. Relationship between BUN, Cr, and PTH

When renal failure happens, parathyroid gland is stimulated. This stimulation leads to secrete too much parathyroid hormone (PTH) in the human's body (Figure 10). This rule can give better insight into renal failure and its side effects. High confidence of 99.64% indicates a relationship between the high ranges of these three attributes (R_{16}). In the normal range of BUN and Cr, there is a high confidence of 93.81%; furthermore, PTH is also in the normal range (R_5).

5.8. Relationship between PTH and Ca

Parathyroid gland disorders have some side effects on the human body. In such a situation, the PTH range increases in the body (Figure 11). In this case, there is a confidence of 76.03% that the Ca range is reduced in the blood (R_{19}). Moreover, when PTH is in the normal range by the confidence of 62.35%, Ca is in the normal range, too (R_9).

5.9. Relationship between PTH and Ph

Phosphorus is one of the indicators of mineral balance in the blood and bones. In many cases, its changes are unlike calcium. As evident from the results (Figure 12 and R_{20}), by the confidence of 70.14%, a high Ph range causes an increased range of PTH, which is a meaningful rule for physicians. Furthermore, by the confidence of 54.30%, these two attributes are in the normal range simultaneously (R_{10}).

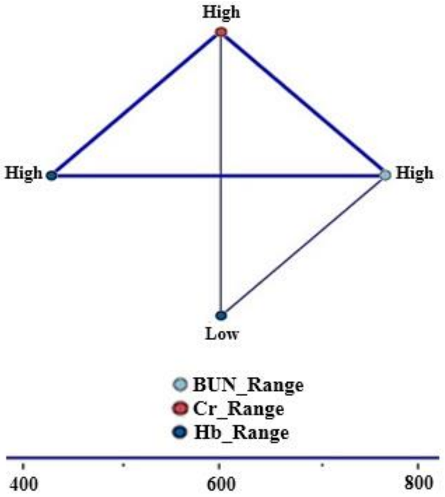


Figure 8. Link chart of BUN, Cr, and Hb.

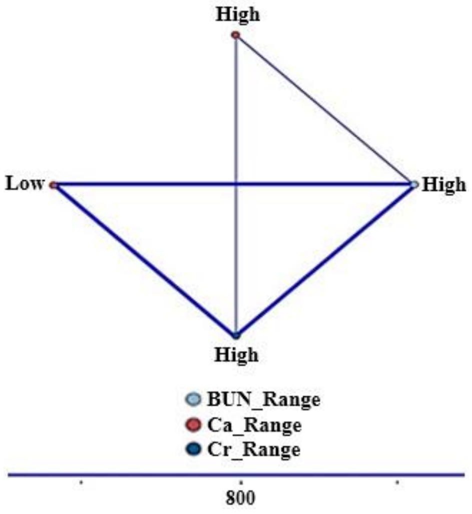


Figure 9. Link chart of BUN, Ca, and Cr.

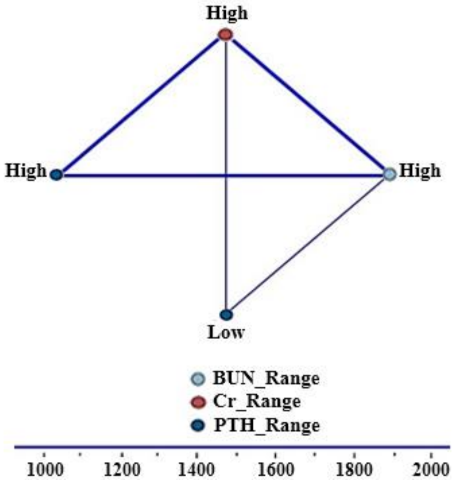


Figure 10. Link chart of BUN, Cr, and PTH.

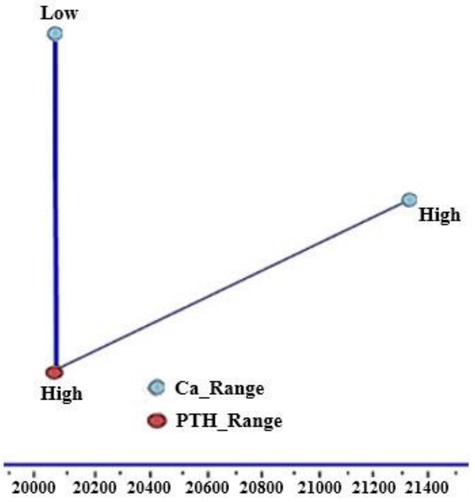


Figure 11. Link chart of Ca and PTH.

5.10. Relationship between ANA and Anti-DNA

ANA and Anti-DNA are indicators of immune health in a human's body. The negative results of these two attributes indicate healthiness in immune diseases (Figure 13). One of the most critical applications of this result is for rheumatism diagnosis. The incidence of negative consequences in these two attributes has 91.32% confidence (R_7).

5.11. Relationship between ALT, AST, and ALKP

There are too many indicators of liver function healthiness. ALT, AST, and ALKP are the most prominent indicators compared to the others. Liver failure leads to high ALT, AST, and ALKP by high confidence of 99.82% (Figure 14 and R_{18}). In their normal range, these attributes have high confidence of 99.26% (R_8).

6. Conclusion

Imagine being able to predict potential health issues before they even arise. That's exactly what we set out to do in this study. Using the power of big data analytics, we analyzed more than one million blood test results from patients to uncover hidden relationships between different blood attributes. Through cutting-edge techniques like k-means clustering and Apriori association rule mining, we were able to pinpoint which blood attributes were most closely linked to one another. And the results were staggering - not only did we identify which attributes were normal or abnormal, but we also discovered that some attributes could serve as early warning signs for potential acute diseases. Our findings can be leveraged to save costs on unnecessary testing and improve patient care. Physicians can use this information to better diagnose and treat patients, while also considering individual factors. This study represents a significant advancement in utilizing big data to enhance healthcare outcomes.

Our future work aims to use the proposed schema to solve other complex tasks, such as citation worthiness [22]. We also plan to employ advanced data mining techniques, inspired by the approach introduced in [31], to create innovative systems for analyzing clinical data.

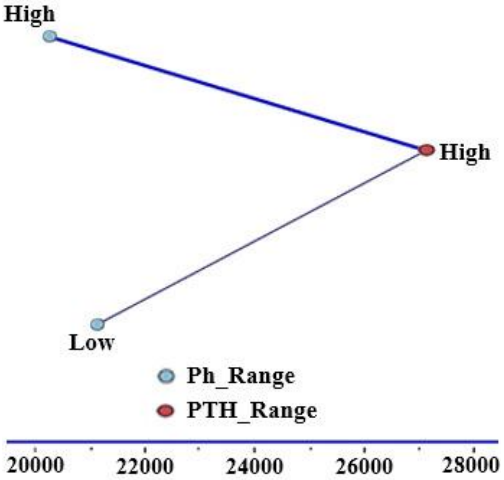


Figure 12. Link chart of Ph and PTH.

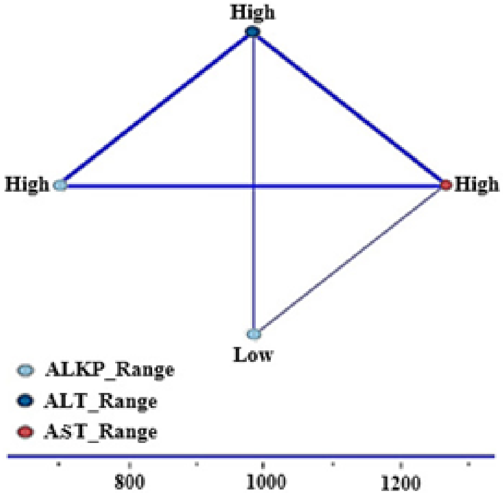


Figure 14. Link chart of ALKP, ALT, and AST.

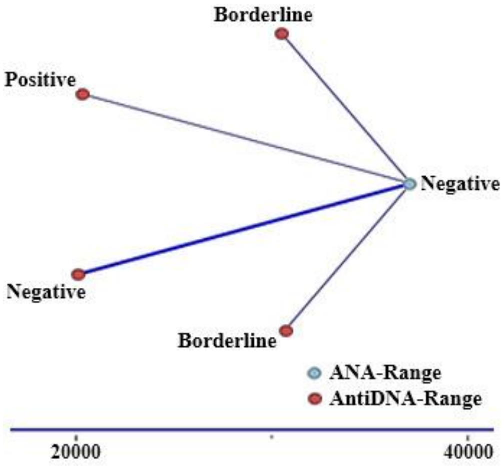


Figure 13. Link chart of ANA and anti-DNA.

Acknowledgements

The authors would like to acknowledge the continuous support of the Motahari healthcare center's employees and physicians. We also thank **Dr. Mohammad Javad Ashraf** and **Dr. Babak Shirazi** for their valuable comments, discussions, and results confirmation.

References

[1] A. Fadlelmoula, D. Pinho, V. H. Carvalho, S. O. Catarino, and G. Minas, "Fourier Transform Infrared (FTIR) Spectroscopy to Analyse Human Blood over the Last 20 Years: A Review towards Lab-on-a-Chip Devices," *Micromachines*, vol. 13, no. 2, pp. 187, Jan. 2022, doi: 10.3390/MI13020187.

[2] M. J. Sousa, A. M. Pesqueira, C. Lemos, M. Sousa, and Á. Rocha, "Decision-Making based on Big Data Analytics for People Management in Healthcare

Organizations," *Journal of medical systems*, vol. 43, no. 9, pp. 1–10, 2019, doi: 10.1007/S10916-019-1419-X/METRICS.

[3] T. Grote and P. Berens, "On the ethics of algorithmic decision-making in healthcare," *Journal of medical ethics*, vol. 46, no. 3, pp. 205–211, Mar. 2020, doi: 10.1136/MEDETHICS-2019-105586.

[4] M. J. Friedrich, "WHO's Top Health Threats for 2019," *JAMA*, Vol. 321, No. 11, pp. 1041–1041, Mar. 2019, doi: 10.1001/JAMA.2019.1934.

[5] A. C. Webster, E. V. Nagler, R. L. Morton, and P. Masson, "Chronic Kidney Disease," *Lancet*, vol. 389, no. 10075, pp. 1238–1252, Mar. 2017, doi: 10.1016/S0140-6736(16)32064-5.

[6] P. R. Pereira, D. F. Carrageta, P. F. Oliveira, A. Rodrigues, M. G. Alves, and M. P. Monteiro, "Metabolomics as a tool for the early diagnosis and prognosis of diabetic kidney disease," *Medicinal Research Reviews*, vol. 42, no. 4, pp. 1518–1544, Jul. 2022, doi: 10.1002/MED.21883.

[7] E. Paul and D. Renmans, "Performance-based financing in the heath sector in low- and middle-income countries: Is there anything whereof it may be said, see, this is new?," *The International journal of health planning and management*, vol. 33, no. 1, pp. 51–66, Jan. 2018, doi: 10.1002/HPM.2409.

[8] A. Akay and H. Hess, "Deep learning: Current and emerging applications in medicine and technology," *IEEE journal of biomedical and health informatics*, vol. 23, no. 3, pp. 906–920, May 2019, doi: 10.1109/JBHI.2019.2894713.

[9] S. Kolluri, J. Lin, R. Liu, Y. Zhang, and W. Zhang, "Machine Learning and Artificial Intelligence in Pharmaceutical Research and Development: a Review," *The AAPS Journal*, vol. 24, no. 1, pp. 1–10, Feb. 2022, doi: 10.1208/S12248-021-00644-3/FIGURES/5.

[10] F. Rabbi, S. R. Dabbagh, P. Angin, A. K. Yetisen, and S. Tasoglu, "Deep Learning-Enabled Technologies for Bioimage Analysis," *Micromachines*, Vol. 13, Page 260, vol. 13, no. 2, p. 260, Feb. 2022, doi: 10.3390/MI13020260.

[11] Z. Lv and L. Qiao, "Analysis of healthcare big data," *Future Generation Computer Systems*, Vol. 109, pp. 103–110, Aug. 2020, doi: 10.1016/J.FUTURE.2020.03.039.

[12] M. S. Islam, M. M. Hasan, X. Wang, H. D. Germack, and M. Noor-E-alam, "A Systematic Review on Healthcare Analytics: Application and Theoretical Perspective of Data Mining," *Healthcare*, vol. 6, no. 2, p. 54, 2018, doi: 10.3390/HEALTHCARE6020054.

[13] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," *Computational and structural biotechnology journal*, vol. 15, pp. 104–116, Jan. 2017, doi: 10.1016/J.CSBJ.2016.12.005.

[14] R. Rastogi and M. Bansal, "Diabetes prediction model using data mining techniques," *Measurement: Sensors*, vol. 25, p. 100605, Feb. 2023, doi: 10.1016/J.MEASEN.2022.100605.

[15] C. T. Wu, C. L. Lo, C. H. Tung, and H. L. Cheng, "Applying Data Mining Techniques for Predicting Prognosis in Patients with Rheumatoid Arthritis," *Healthcare*, vol. 8, no. 2, p. 85, Apr. 2020, doi: 10.3390/HEALTHCARE8020085.

[16] M. Barrios, M. Jimeno, P. Villalba, and E. Navarro, "Novel Data Mining Methodology for Healthcare Applied to a New Model to Diagnose Metabolic Syndrome without a Blood Test," *Diagnostics*, vol. 9, no. 4, p. 192, Nov. 2019, doi: 10.3390/DIAGNOSTICS9040192.

[17] A. Begum and A. Parkavi, "Prediction of thyroid Disease Using Data Mining Techniques," *2019 5th International Conference on Advanced Computing and Communication Systems, ICACCS 2019, pp. 342–345, Mar. 2019*, doi: 10.1109/ICACCS.2019.8728320.

[18] C. Thallam, A. Peruboyina, S. S. T. Raju, and N. Sampath, "Early Stage Lung Cancer Prediction Using Various Machine Learning Techniques," *Proceedings of the 4th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2020, pp. 1285–1292, Nov. 2020*, doi: 10.1109/ICECA49313.2020.9297576.

[19] S. M. Ayyoubzadeh et al., "A study of factors related to patients' length of stay using data mining techniques in a general hospital in southern Iran," *Health information science and systems*, vol. 8, pp. 1–11, 2020, doi: 10.1007/S13755-020-0099-8/METRICS.

[20] K. S. Lakshmi and G. Vadivu, "A novel approach for disease comorbidity prediction using weighted association rule mining," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–8, Jan. 2019, doi: 10.1007/S12652-019-01217-1/METRICS.

[21] W. Luo *et al*., "Clinical data mining reveals Gancao-Banxia as a potential herbal pair against moderate COVID-19 by dual binding to IL-6/STAT3," *Computers in biology and medicine*, vol. 145, p. 105457, Jun. 2022, doi: 10.1016/J.COMPBIOMED.2022.105457.

[22] M. Roostaee, "Citation Worthiness Identification for Fine-Grained Citation Recommendation Systems," *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, vol. 46, no. 2, pp. 353-365, 2022, doi: 10.1007/s40998-021-00472-3.

[23] N. Almugren, N. Alrumayyan, R. Alnashwan, A. Alfutamani, I. Al-Turaiki, and O. Almugren, "The effect of vitamin B12 deficiency on blood count using data mining," *Advances in Intelligent Systems and Computing*, vol. 753, pp. 234–245, 2018, doi: 10.1007/978-3-319-78753-4_18/COVER.

[24] I. K. Park and G. S. Choi, “Rough set approach for clustering categorical data using information-theoretic dependency measure,” *Information Systems*, vol. 48, pp. 289–295, Mar. 2015, doi: 10.1016/J.IS.2014.06.008.

[25] A. Holzinger, M. Dehmer, and I. Jurisica, “Knowledge Discovery and interactive Data Mining in Bioinformatics - State-of-the-Art, future challenges and research directions,” *BMC Bioinformatics*, vol. 15, no. 6, pp. 1–9, May 2014, doi: 10.1186/1471-2105-15-S6-I1/FIGURES/2.

[26] W. C. Lin and C. F. Tsai, “Missing value imputation: a review and analysis of the literature (2006–2017),” *Artificial Intelligence Review*, vol. 53, no. 2, pp. 1487–1509, Feb. 2020, doi: 10.1007/S10462-019-09709-4/METRICS.

[27] A. Elasra, “Multiple Imputation of Missing Data in Educational Production Functions,” *Computation*, vol. 10, no. 4, p. 49, Apr. 2022, doi: 10.3390/COMPUTATION10040049/S1.

[28] S. K. Paul, J. Ling, M. Samanta, and O. Montvida, “Robustness of Multiple Imputation Methods for Missing Risk Factor Data from Electronic Medical Records for Observational Studies,” *Journal of Healthcare Informatics Research*, vol. 6, pp. 385–400, 2022, doi: 10.1007/S41666-022-00119-W/METRICS.

[29] P. C. Austin, I. R. White, D. S. Lee, and S. van Buuren, “Missing Data in Clinical Research: A Tutorial on Multiple Imputation,” *Canadian Journal of Cardiology*, vol. 37, no. 9, pp. 1322–1331, Sep. 2021, doi: 10.1016/J.CJCA.2020.11.010.

[30] T. T. D. Nguyen, L. T. T. Nguyen, Q. T. Bui, U. Yun, and B. Vo, “An efficient topological-based clustering method on spatial data in network space,” *Expert Systems with Applications*, vol. 215, p. 119395, Apr. 2023, doi: 10.1016/J.ESWA.2022.119395.

[31] M. W. Li, D. Y. Xu, J. Geng, and W. C. Hong, "A hybrid approach for forecasting ship motion using CNN–GRU–AM and GCWOA," *Applied Soft Computing*, 114, 108084, 2022, doi: 10.1016/j.asoc.2021.108084.

[32] P. Fränti and S. Sieranoja, “How much can k-means be improved by using better initialization and repeats?,” *Pattern Recognition*, Vol. 93, pp. 95–112, Sep. 2019, doi: 10.1016/J.PATCOG.2019.04.014.

[33] M. E. Celebi, H. A. Kingravi, and P. A. Vela, “A comparative study of efficient initialization methods for the k-means clustering algorithm,” *Expert Systems with Applications*, vol. 40, no. 1, pp. 200–210, Jan. 2013, doi: 10.1016/J.ESWA.2012.07.021.

[34] X. Liu, L. Zheng, W. Zhang, J. Zhou, S. Cao, and S. Yu, “An Evolutive Frequent Pattern Tree-based Incremental Knowledge Discovery Algorithm,” *ACM Transactions on Management Information Systems*, vol. 13, no. 3, Feb. 2022, doi: 10.1145/3495213.

[35] Y. B. Abushark, “An intelligent feature selection approach with systolic tree structures for efficient association rules in big data environment,” *Computers and Electrical Engineering*, Vol. 101, p. 108080, Jul. 2022, doi: 10.1016/J.COMPELECENG.2022.108080.

[36] A. Cecconi, G. De Giacomo, C. Di Ciccio, F. M. Maggi, and J. Mendling, “Measuring the interestingness of temporal logic behavioral specifications in process mining,” *Information Systems*, vol. 107, p. 101920, Jul. 2022, doi: 10.1016/J.IS.2021.101920.

[37] M. R. Keyvanpour, Z. K. Zandian, and N. Mottaghi, “BRTSRDM: Bi-Criteria Regression Test Suite Reduction based on Data Mining,” Journal of AI and Data Mining, vol. 11, no. 2, pp. 161-168, Apr. 2023, doi: 10.22044/jadm.2023.12208.2374.

[38] K. Hu, L. Qiu, S. Zhang, Z. Wang, and N. Fang, “An incremental rare association rule mining approach with a life cycle tree structure considering time-sensitive data,” *Applied Intelligence*, pp. 1–25, Aug. 2022, doi: 10.1007/S10489-022-03978-3/METRICS.

[39] M. Tandan, Y. Acharya, S. Pokharel, and M. Timilsina, “Discovering symptom patterns of COVID-19 patients using association rule mining,” *Computers in biology and medicine*, vol. 131, p. 104249, Apr. 2021, doi: 10.1016/J.COMPBIOMED.2021.104249.

[40] J. Hong, R. Tamakloe, and D. Park, “Application of association rules mining algorithm for hazardous materials transportation crashes on expressway,” *Accident Analysis & Prevention*, vol. 142, p. 105497, Jul. 2020, doi: 10.1016/J.AAP.2020.105497.

[41] F. Zahedi and M.-R. Zare-Mirakabad, “Employing data mining to explore association rules in drug addicts,” *Journal of AI and Data Mining*, vol. 2, no. 2, pp. 135–139, Jul. 2014, doi: 10.22044/JADM.2014.308

[42] J. W. Lee and O. Harel, "Incomplete clustering analysis via multiple imputation," *Journal of Applied Statistics*, 1-18, 2022, doi: 10.1080/02664763.2022.2060952.

[43] M. Quartagno and J. R. Carpenter, "Substantive model compatible multilevel multiple imputation: A joint modeling approach," *Statistics in medicine*, vol. 41, no. 25, pp. 5000-5015, 2022, doi: 10.1002/sim.9549.

[44] N. P. Jayasri and R. Aruna, “Big data analytics in health care by data mining and classification techniques,” *ICT Express*, vol. 8, no. 2, pp. 250–257, Jun. 2022, doi: 10.1016/J.ICTE.2021.07.001.

[45] A. Alaiad, H. Najadat, B. Mohsen, and K. Balhaf, “Classification and Association Rule Mining Technique for Predicting Chronic Kidney Disease,” *Journal of Information & Knowledge Management*, vol. 19, no. 1, Mar. 2020, doi: 10.1142/S0219649220400158.

[46] B. Andrianto, Y. K. Suprapto, I. Pratomo, and I. Irawati, “Clinical decision support system for typhoid fever disease using classification techniques,” *Proceedings - 2019 International Seminar on Intelligent Technology and Its Application, ISITIA 2019, pp. 248–252, Aug. 2019,* doi: 10.1109/ISITIA.2019.8937286.

کشف الگوی پنهان در داده‌های بالینی: رویکردی مبتنی بر تکنیک‌های داده کاوی

میثم روستائی[1،*] و راضیه میدانشاهی[2]

[1] **گروه مهندسی کامپیوتر، دانشکده مهندسی و فناوری، دانشگاه مازندران، بابلسر، ایران.**

[2] **گروه مهندسی کامپیوتر، دانشکده مهندسی کنترل و کامپیوتر، دانشگاه پلی تکنیک تورین، تورین، ایتالیا.**

ارسال ۲۰۲۳/۰۳/۲۷؛ بازنگری ۲۰۲۳/۰۵/۱۴؛ پذیرش ۲۰۲۳/۰۷/۰۱

چکیده:

در این مطالعه، با استفاده از تکنیک‌های داده کاوی بدون نظارت بر روی مجموعه داده‌ای بزرگ از نتایج آزمایش خون بیش از یک میلیون بیمار، سعی در به حداقل رساندن نیاز به انجام تست‌های خون اضافی در تشخیص بیماری‌های رایج داشتیم. برای اطمینان از دقت مدل، داده‌های غیرعددی و توصیفی کنار گذاشته شدند. به منظور شناسایی روابط بین ویژگی‌ها، مجموعه‌ای از روش‌های بدون نظارت از جمله پیش پردازش، خوشه‌بندی و استخراج قوانین انجمنی به کار گرفته شد. رویکرد پیشنهادی روابطی را کشف کرد که پزشکان را قادر می‌سازد بیماری‌های حاد احتمالی را در مراحل اولیه شناسایی کنند و در نتیجه، بهبود نتایج بیمار و کاهش هزینه‌ها را ممکن می‌سازد. قابلیت اطمینان الگوهای استخراج‌شده نیز نشان می‌دهد که این رویکرد می‌تواند منجر به صرفه‌جویی قابل توجه در زمان و هزینه و در عین حال کاهش بار کاری برای پرسنل آزمایشگاه شود. مطالعه حاضر، اهمیت تجزیه و تحلیل داده‌های بزرگ و تکنیک‌های یادگیری بدون نظارت را در افزایش کارایی مراکز بهداشتی برجسته می‌کند.

کلمات کلیدی: داده‌های بالینی، داده کاوی، یادگیری بدون نظارت، کشف قوانین انجمنی، خوشه‌بندی.