EFFECTIVE PATTERN RECOGNITION USING MODIFIED PREFIX SPAN ALGORITHM

[1]Komal Sidhu, [2]Ritika, [3]Baljinder Singh

[1]Department of Computer Science and Engineering,
[1]Beant College of Engineering and Technology,Gurdaspur,India

***Abstract*:** The phenomenal advances in health and biotechnology have been produces huge amount of data like clinical data and high throughput information that makes Electronic Health records (EHRs) expansive and complex. For handling this AI and data mining techniques have been utilizes along with health services. Today Data mining is utilized to detect abnormal pattern using various informational datasets along with machine learning algorithms. There are many techniques available which is utilized for diagnosis of diabetes abnormal patterns like FP growth, Apriori, FP-Tree algorithm. These techniques discover unknown patterns or relationships from large amount of data and these are utilized for making decisions for preventive and suggestive medicine. The main disadvantage of these techniques is it discovers fewer patterns. In this paper we proposed prefix span along with MPV algorithm that discover more patterns to detect abnormal patterns accurately. The results will help in predicting quicker and more accurate abnormal patterns so that it lead timely treatment of the patients.

***Keywords*—Data Mining, Diabetes Abnormal patterns Diagnosis and Treatment, Data mining techniques.**

I. INTRODUCTION

1.1 Data Mining

Data Mining has quickly grown with the presence of the wonder BIG Data. For sure, numerous associations have begun to digitize their records, and have changed their paper-based frameworks to electronic frameworks. This change conveys a few advantages to the associations, among them time funds, a superior administration and a more tightly checking making the assignments less demanding. One of the immediate results of this change is the visit gathering of significant Data. While the Data's holders started to stress over the capacity of Data, they understood the benefits they can take from it. The Data gathered can be considered as another unformatted of structure (Raw Data) which needs to be filtered. Handling Data give a superior quality Data which contribute in request to make choice in data selection[1]. Moreover, Healthcare elements likewise choose electronic frameworks, by utilizing different strategies, among them, Electronic health Record (EHR) or Electronic Medical Records (EMR) frameworks. It implies the executing EHR frameworks, leads to an immense measure of Data gathered by doctor's facilities, centers and other health care suppliers. At that point, the vast majority of these Datasets are most certainly not extremely very much organized and fitting for explanatory purposes. In expansion, health care Data are generally extremely perplexing and difficult to investigate. For instance the US Healthcare framework alone as of now achieved 150 Exabyte (1 Exabyte = 8388608 Terabit) five years prior. This pattern is because of the way that multi scale Data created from people is consistently expanding, especially with the new high-throughput sequencing stages, continuous imaging, and purpose of care gadgets, also as wearable figuring and versatile health care innovations. As needs be, Data Mining has gotten a great deal of consideration on account of its solid capacity of separating Data from Data, furthermore, winds up noticeably prevalent in Healthcare field by dint of its productive diagnostic procedure for recognizing obscure and significant Data in health care Data[2], [3].

1.2 Data Mining and Health Related Models

The upsides of Data Mining for health care and the reasons make Data Mining critical to be considered in health care Data examination. Data mining health care Dataset with missing values is considered to be analyzed initially through Support vector machine and accuracy is analyzed and after that ARIMA with KNN and Euclidean distance is used for rectification and analysis purpose[5]–[7][8][9]. Accuracy is observed in both the cases to prove worth of the study.

Data mining approaches is the base of this literature. Analysis of existing literature provide base for proposed literature. [10] Reviewed various models and methods used within data mining. Data mining techniques development from 2005 to 2015 is reviewed and application in regards to health care is proposed. [3] Suggests the integration of medical data with data mining strategies used to form medical information system. Patient medical condition can be analyzed along with future prediction about patient's health. Hidden possibilities can be extracted using unlimited data mining techniques to make accurate health forecast. [11] Proposed multilayer perceptron in order to analyze big data corresponding to health care. As literature deals with health care of patients hence high degree of accuracy is desired. To accomplish the desired goal comparison of SVM and multilayer perceptron on health care data set is made. Results of SVM in terms of classification are better as compared to multilayer perceptron. It suggests data mining techniques used for analysis of diabetics. Support Vector Machine (SVM) is used for this purpose. Genetic approach is also analyzed for diabetic's dataset in the field of data mining. Results of SVM are obtained to be better. [12] Suggests five J.48 classifiers to predict hypertension and eight other abnormal patterns. Prediction accuracy is obtained and compared against naïve Bayes approach. Results in terms of J.48 are obtained to be better. [6] Suggests hybrid approach for health care to predict abnormal pattern using Big data. Pruning based KNN is used for this purpose which used density based clustering based method integrated with KNN approach. Local outlier factor of PB-KNN is better as compared to KNN. [13] Proposes SVM and neural network techniques for skin lesion detection in human body. Segmentation along with classification is performed in order to detect the abnormal pattern. [7]predict heart abnormal pattern are

primary cause of death among humans in last decade. Data mining techniques are used in order to detect and predict heart abnormal pattern efficiently. [2] Proposes a mechanism through which information about patient coming for checkup at hospital is stored and algorithm is applied in order to perform predictions. Data mining algorithm considered in this approach is naïve Bayes. Accuracy of prediction is obtained is significant in this case.[14] Suggested intelligent heart abnormal patterns prediction system. Decision tree, naïve Bayes and neural network technique are used for accurate analysis and prediction of abnormal patterns.

1.3 Pattern Mining

Frequent pattern mining is a heavily researched area in the field of data mining with wide range of applications. Mining frequent patterns from large scale databases has emerged as an important problem in data mining and knowledge discovery community, the most popular algorithm for pattern mining is without a doubt Apriori (1993).

- **Apriori Algorithm**

It is designed to be applied on a transaction database to discover patterns in transactions made by customers in stores. But it can also be applied in several other applications. A transaction is defined a set of distinct items (symbols). Apriori takes as input (1) a *minsup* threshold set by the user and (2) a transaction database containing a set of transactions. Apriori outputs all frequent itemsets, i.e. groups of items shared by no less than *minsup* transactions in the input database. For example, consider the following transaction database containing four transactions. Given a *minsup* of two transactions, frequent itemsets are “bread, butter”, “bread milk”, “bread”, “milk” and “butter”.

T1:bread,butter,spinach
T2:butter,salmon
T3:bread,milk,butter
T4: cereal, bread, milk

Fig.1 a transaction database

Apriori can also apply a post-processing step to generate “association rules” from frequent itemsets, which I will not discuss here. The Apriori algorithm has given rise to multiple algorithms that address the same problem or variations of this problem such as to (1) incrementally discover frequent itemsets and associations , (2) to discover frequent subgraphs from a set of graphs, (3) to discover subsequences common to several sequences, etc.

- **FP-Tree Algorithm**

An algorithm to Frequent Sequence Mining is the FP-TREE (Sequential Pattern Discovery using Equivalence classes) algorithm. It uses a vertical id-list database format, where we associate to each sequence a list of objects in which it occurs. Then, frequent sequences can be found efficiently using intersections on id-lists. The method also reduces the number of databases scans, and therefore also reduces the execution time.

The first step of FP-TREE is to compute the frequencies of 1-sequences, which are sequences with only one item. This is done in a single database scan. The second step consists of counting 2-sequences. This is done by transforming the vertical representation into a horizontal representation in memory, and counting the number of sequences for each pair of items using a bi-bi dimensional matrix. Therefore, this step can also be executed in only one scan. Subsequent n-sequences can then be formed by joining (n-1)-sequences using their id-lists. The size of the id-lists is the number of sequences in which an item appears. If this number is greater than minsup, the sequence is a frequent one. The algorithm stops when no frequent sequences can be found anymore. The algorithm can use a breadth-first or a depth-first search method for finding new sequences

- **PrefixSpan**

The PrefixSpan (Prefix Projected Sequential pattern Mining) algorithms proposed by Jian Pei, Jiavei Han and Helen Pinto is the only projection based algorithms in all the sequencing pattern mining algorithms. It is more efficient than the algorithm like apriori, freespan, SPADE, FP Growth. This algorithm discovers the frequent items by scanning the sequence database once. The database is projected into many smaller databases according to the frequent items. By recursively growing subsequence fragment in every projected database, It found the complete set of sequential pattern. The main idea behind the prefixspan algorithm to successfully discovered patterns is employing the divide-and-conquer strategy.

II. LITERATURE SURVEY

Eenan(2009) proposed[10] a non-homogeneous mark over model is used to identify the chronic abnormal patterns in the patients. The algorithm uses global optimization that efficiently identify the number of frequent pathway required to analyses the patient. The result shows that the proposed methodology probability is better than existing ones but this approach can be extended using admission scheduling policy.

Béchet et al. (2012) proposed paper presents the sequential pattern mining [6] to discover the rare abnormal patterns within human body where experiments are conducted using data mining tool WEKKA. This show betterment in percentage for classification accuracy.

Abbasghorbani and Tavoli (2015) proposed the analysis of various pattern mining techniques [1] are done and also the features of all the algorithms. It introduced various minimizing support counting which is used for minimizing search space. We have generated small search space which will include earlier candidate sequence pruning then database is analyzed and compression technique is used to analyze.

Ghosh et al. (2015) proposed [12] a technique that extracts sequential patterns from hypotensive patient groups. These patterns are further utilized to inform medical decisions and randomized clinical trials. It further extended by including various clinical features and also include some sequential patterns. It also does not considered missing value during the preprocessing phase.

Alzahrani (2016) proposed[4] data mining method for abnormal patterns prediction for this purpose sequential data mining is used in order to accomplish this data preprocessing mechanism is applied. After applying preprocessing mechanism the attributes will be analyzed this will be done using passes on medical data. The first pass determines whether support for each abnormal patterns is present or not at the end of this phase the frequent abnormal patterns within the database will be identified, a counter will be maintained to count the occurrence of each abnormal patterns within the dataset. Next phase determines the second sequence of abnormal pattern present within the dataset. The overall process yield the abnormal pattern which can cause the occurrence of other abnormal pattern. The abnormal patterns resulting in another abnormal patterns is termed as candidate generation. And for declaring that it is generated from the previous level Pruning is used.

Patel and Chaudhari (2016) gives analysis of various sequential pattern mining algorithm [20] are done. It discover the various challenges in these algorithm and improved the performance by proposing constraints in patterns. It enhances existing CAI prefix span algorithm by introducing time constraints. The comparative results shows it is better and we can also further enhance it by applying efficient constraints.

Uragaki et al. (2016) proposed an approach [27] that verified and recommend clinical pathway to the patients it utilizes sequential pattern mining that handles the record between various time intervals . the proposed methodology uses the actual logs of patients that would further analyze these pattern using T-prefix span algorithm .but it will be necessary to introduce faster mining algorithm that are not in proposed methodology.

Zhang et al. (2016) proposed a technique named [29] ConSgen that are used to identify the contiguous sequential generator and also minimize the redundant patterns, It utilizes the divide conquer technique to find the sequential generator with contiguous constraints. But it does not considered the gapped alignments and also not discovered the binding sites.

Ahmed (2017) proposed [2] an application that utilizes the data mining technique to predict the heart abnormal patterns. Also it guide the patient to take treatment at early stage. But is completely dependent upon patient input and does not considered predefined dataset values. It also not utilizes the missing value that are essential to predict abnormal pattern.

Alamanda et al. (2017) sequence pattern mining is proposed[3] in order to detect the time duration used for promotion .the sequence or pattern is checked from within the database . The weight of each sequence in each database is achieved from the interval of the successive element in the sequence and the mining is performed on the basis of weight considering time interval. Time interval based pattern is used in this case. In preprocessing missing values are not considered.

Chen et al. (2017) proposed [8] a pattern growth method is used that analyze the medical database to specify the combination of chronic abnormal patterns it introduce prefix span algorithm that identify all possible patterns in the images but it constrained only specific abnormal patterns and can further improved for efficient search, it shows the results in terms of HTN and DP abnormal pattern.

Cheng et al. (2017) proposed [9] a sequential mining approach for early assessment of chronic abnormal patterns. The clinical database is considered .A dataset of patients derived from Taiwan, it derives richest of risk patterns. Data preprocessing as performed to rectify the problem if found but missing values are not considered .sequential pattern mining is used to observe the risk pattern and generate the result. The problem with this approach is that no precautions have been suggested. The classification accuracy is 80% further improvement in classification is needed. The chronic abnormal patterns is analyzed in this paper built in over the existing problem.

Kunjir et al. (2017) enhancement [16] which can be improved proposed multiclass Naïve Bayes algorithm is used for prediction of particular abnormal patterns but training it on set of data before implementation. This is downloaded from UCI repository work. The proposed system can help doctors to take clinical decisions where traditional decision support system fails, J47 algorithm is also used for proving the worth of study of accuracy in heart abnormal patterns, breast cancer and diabetes approaches 83% by using this approach. This accuracy requires in future.

Sarac and Seker (2017) proposed a mining technique[22]that are used to reduce the complexity and cost of the data storage. It divide chunks into separate parts and regression analysis are to be done to analysis the trial variable and samples dataset. But it does not considered separate chunks for feature analysis and separate storage reservoir also not utilized.

Zihayat et al. (2017) identified [30] a problem of top –k utility based regulation pattern which is used to find out meaning in biology. Firstly proposed a utility model called TU-SEQ which is used to find top –K high utility gene regulation sequential patterns. It is considering the relation between the various patterns and interactions in biological studies.

III. PROPOSED METHODOLOGY

The proposed algorithm uses the prefix span algorithm for determining patterns which can be grouped together to form clusters. Pre-processing mechanism includes most probable value replacement with the missing value.

Algorithm:

- Input: Dataset
- Output: Classification Accuracy, Abnormal patterns Prediction
- Input Dataset

  Data=$Dataset_i$

  Where I are the number of rows within the dataset
- Apply Pre-processing mechanism to resolve the missing values

  MPV=mean(Values($Person_{idi}$=dataset($person_{idi}$)))
- Repeat while all the missing values are tackled

  If ($Missing_i$)

  $Missing_i$=MPV

  End of if

  End of loop
- Apply Pre-fix span algorithm for pattern growth determination
- Form clusters

  Repeat until values in dataset are examined

  If($Datset_{iValue}$==$Dataset_{i+1value}$)

  $Cluster_i$=$Datset_{iValue}$

  End of if

  I=i+1

  End of loop
- Predict abnormal patterns looking at the pattern clusters
- Result: Accuracy, Abnormal patterns

IV. RESULTS AND PERFORMANCE ANALYSIS

The performance of the system is analyzed by the use of parameters such as accuracy, specificity and sensitivity.

Accuracy is obtained by subtracting the actual result from the approximate result. In terms of predictions accuracy is obtained as

$$Accuracy = Correct_{predictions} / Total_{predictions} \quad (4.1)$$

Sensitivity is obtained by dividing number of positive predictions to the total true positive rate.

$$Sensitivity = Correct_{positive\_predictions} / Total_{positives} \quad (4.2)$$

Specificity is another parameter used to evaluate correctness of the proposed system. It is given as under

$$Specificity = TN / (TN / FP) \quad (4.3)$$

Where TN=True Negatives

FP=False Positives

The abnormal patterns detection and prediction is given though accurate classification, result in terms of plots is given as under.

Results in terms of **number of abnormal patterns** discovered by varying the size of dataset tuples is more in case of proposed approach as compared to existing approach. Pattern discovered in the proposed and existing are listed through the following structure

Table 4.1: Number of abnormal patterns discovered

Dataset Size	Parameter	FP-growth	Prefix Span
5 Rows with 55 attributes	Number of Abnormal Patterns	20	35
10 rows with 100 attributes	Number of Abnormal Patterns	40	55

The result in terms of patterns is more in case of Prefix span algorithm as compared to FP growth algorithm but optimality will be tested only through the parameters such as accuracy, specificity and sensitivity.

Result comparison in terms of accuracy, sensitivity and specificity is given as under

Table 4.2: Parameters observation in terms of accuracy, sensitivity and specificity

Dataset Size	Parameters	FP-growth (%)	Prefix Span (%)
5 Rows with 55 attributes	Accuracy	77	85
	Specificity	75	84
	Sensitivity	79	84
10 rows with 100 attributes	Accuracy	77	85
	Specificity	79	86
	Sensitivity	78	87
20 rows with 200 attributes	Accuracy	78	86
	Specificity	79	87
	Sensitivity	78	87

Classification accuracy of proposed system appears to be more as compared to existing techniques. Multiple class prediction mechanism showing higher accuracy proving the worth of study.

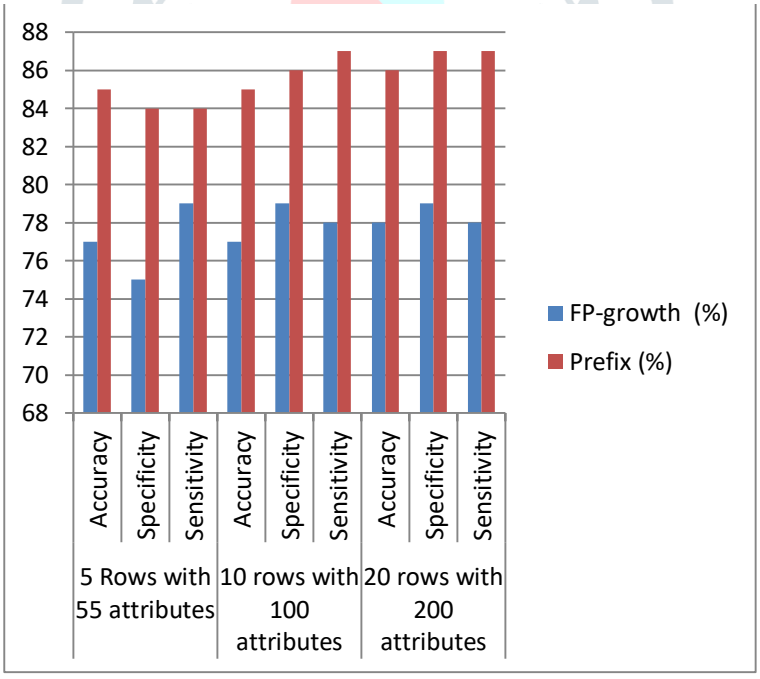

Figure 4.1: Confusion matrix

Results and performance analysis as indicated through the plot shows that prefix span algorithm along with MPV algorithm yield better result.

V. CONCLUSION AND FUTURE SCOPE

In this paper an automated system that utilizes MPV along with Prefix span algorithm for detecting diabetes proposed. Pre-processing phase is critical and is well defined using noise handling and resizing operation. Obtained images are fed into the trained network for feature extraction using prefix span algorithm and classification is performed using MPV. Hybrid approach followed gives better results.

The main objective of the proposed literature is creating optimized detection using prefix span for better accuracy. Higher accuracy is achieved by the use of said literature. In future, proposed strategy can be examined against the real time datasets for better evaluation of accuracy.

REFERENCES

[1] Abbasghorbani, S., and Tavoli, R. 2015, "Survey on Sequential Pattern Mining Algorithms," *2015 2nd Int. Conf. Knowledge-Based Eng. Innov.*, pp. 1153–1164, 2015.

[2] Ahmed, F.2017,"A Simple Acute Myocardial Infarction (Heart Attack) Prediction System Using Clinical Data and Data Mining Techniques," pp. 22–24, 2017.

[3] Alamanda, S., Pabboju, S. and Gugulothu, N. 2017, "An Approach to Mine Time Interval Based Weighted Sequential Patterns in Sequence Databases," *2017 13th Int. Conf. Signal-Image Technol. Internet-Based Syst.*, pp. 29–34, 2017.

[4] Alzahrani, MY. 2016,"Discovering Sequential Patterns from Medical Datasets," 2016.

[5] Anusha,C., Vinay, SK., Pooja, HJ. and Ranganatha, S. 2013,"Medical data mining and analysis for heart abnormal patterns dataset using classification techniques," *Natl. Conf. Challenges Res. Technol. Coming Decad. (CRT 2013)*, p. 1.09-1.09, 2013.

[6] Béchet,N., Cellier, P., Charnois, T., Cremilleux, B. and Jaulent,MC. 2012,"Sequential pattern mining to discover relations between genes and rare abnormal pattern," *Proc. - IEEE Symp. Comput. Med. Syst.*, 2012.

[7] Chang, HY., Lin, JC., Cheng, ML. and Huang, SC. 2016,"*A novel incremental data mining algorithm based on FP-growth for big data". Proc-2016 Int Conf Netw Appl NaNA 2016 375-378.doi:10.1109/NaNA.2016.77.*

[8] Chen, CJ., Pai, TW., Lin, SS.,Yeh, CC., Liu, MH. and Wang, CH. 2017,"Application of PrefixSpan Algorithms for Abnormal patterns Pattern Analysis," *Proc. - 2016 Int. Comput. Symp. ICS 2016*, pp. 274–278, 2017.

[9] Cheng, Y., Lin, Y-F., Chiang, K-H. and Tseng, V. 2017,"Mining Sequential Risk Patterns from Large-Scale Clinical Databases for Early Assessment of Chronic Abnormal pattern: A Case Study on Chronic Obstructive Pulmonary Abnormal patterns," *IEEE J. Biomed. Heal. Informatics*, pp. 1–1, 2017.

[10] Eenan,BRM. 2009, "Non-homogeneous Markov models for sequential pattern mining of healthcare data," pp. 327–344, 2009.

[11] Farooq, MA., Azhar, MA. and Raza, RH. 2016, "Automatic Lesion Detection System (ALDS) for Skin Cancer Classification Using SVM and Neural Classifiers," *2016 IEEE 16th Int. Conf. Bioinforma. Bioeng.*, pp. 301–308, 2016.

[12] Ghosh, S., Feng, M., Nguyen, H. and Li, J. 2015, "Hypotension Risk Prediction via Sequential Contrast Patterns of ICU Blood Pressure," *IEEE J. Biomed. Heal. Informatics*, vol. 20, no. 5, pp. 1416–1426, 2015.

[13] Huang, F., Wang, S. and Chan, C. 2012,"Predicting Abnormal patterns By Using Data Mining Based on Healthcare Information System," *2012 IEEE Int. Conf. Granul. Comput. Predict.*, pp. 12–15, 2012.

[14] Jain, S. and Pise, N. 2015,"Computer aided Melanoma skin cancer detection using Image Processing," *Procedia - Procedia Comput. Sci.*, vol. 48, no. Iccc, pp. 735–740, 2015.

[15] Jothi, N., Rashid, NA. and Husain, W. 2015,"Data Mining in Healthcare - A Review," *Procedia Comput. Sci.*, vol. 72, pp. 306–313, 2015.

[16] Kunjir,A., H. Sawant., and Shaikh NF. 2017,"Data mining and visualization for prediction of multiple abnormal

pattern in healthcare," *Proc. 2017 Int. Conf. Big Data Anal. Comput. Intell. ICBDACI 2017*, pp. 329–334, 2017.

[17] Leland, WE. and Wilson, DV. 1992, "On the Self-Similar Nature of Ethernet Traf c," *Comput. Commun. Rev.*, vol. 2, no. August 1989, pp. 203–213, 1992.

[18] Naraei, P. and Street, V. 2016,"Application of Multilayer Perceptron Neural Networks and Support Vector Machines in Classification of Healthcare Data," no. December, pp. 848–852, 2016.

[19] Palaniappan, S. and Awang, R. 2008,"Intelligent heart abnormal patterns prediction system using data mining techniques," *2008 IEEE/ACS Int. Conf. Comput. Syst. Appl.*, pp. 108–115, 2008.

[20] Patel, R. and Chaudhari, T. 2016, "A Review on Sequential Pattern Mining using Pattern Growth Approach," pp. 4–7, 2016.

[21] Pinheiro, E., Weber, W. and Barroso, L. 2007, "Failure trends in a large disk drive population," *Proc. 5th USENIX Conf. File Storage Technol. (FAST 2007)*, no. February, pp. 17–29, 2007.

[22] Sarac, F. and Seker, H. 2017,"An instance selection framework for mining data streams to predict antibody-feature function relationships on RV144 HIV vaccine recipients," *2016 IEEE Int. Conf. Syst. Man, Cybern. SMC 2016 - Conf. Proc.*, pp. 3356–3361, 2017.

[23] Sharma A. and Mansotra V. 2014, "Emerging applications of data mining for healthcare management - A critical review," *2014 Int. Conf. Comput. Sustain. Glob. Dev.*, pp. 377–382, 2014.

[24] Sivagowry, S., Durairaj, M. and Persia, A. 2013, "An empirical study on applying data mining techniques for the analysis and prediction of heart abnormal patterns," *2013 Int. Conf. Inf. Commun. Embed. Syst.*, pp. 265–270, 2013.

[25] Student, ME. and Nadu, CT. 2016,"Heart abnormal patterns classification and its co-morbid condition detection using WPCA genetic algorithm," pp. 287–291, 2016.

[26] Ţăranu, I. 2015,"Data mining in healthcare: decision making and precision," *Database Syst. J.*, vol. 5, no. 4, pp. 33–40, 2015.

[27] Uragaki, K., Hosaka, T., Arahori, Y., Kushima, M., Yamazaki, T., Araki, K. and Yokota, H. 2016, "Sequential pattern mining on electronic medical records with handling time intervals and the efficacy of medicines," *Proc. - IEEE Symp. Comput. Commun.*, vol. 2016–Augus, pp. 20–25, 2016.

[28] Yan, K., You, X., Ji, X., Yin, G. and Yang, F. 2016 ,"A Hybrid Outlier Detection Method for Health Care Big Data," *2016 IEEE Int. Conf. Big Data Cloud Comput. (BDCloud), Soc. Comput. Netw. (SocialCom), Sustain. Comput. Commun.*, pp. 157–162, 2016.

[29] Zhang, J., Wang, Y., Zhang, C. and Shi, Y. 2016, "Mining contiguous sequential generators in biological sequences," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 13, no. 5, pp. 855–867, 2016.

[30] Zihayat, M., Davoudi, H. and An, A. 2017, "Top-k utility-based gene regulation sequential pattern discovery," *Proc. - 2016 IEEE Int. Conf. Bioinforma. Biomed. BIBM 2016*, pp. 266–273, 2017.