Journal of Engineering, Mechanics and Modern Architecture
Vol. 3, No. 06, 2024 ISSN: 2181-4384

THE USE OF ARTIFICIAL INTELLIGENCE IN THE DETECTION OF MORAL RISK PATTERNS IN REGISTERED MEDICAL HEALTH DATA

Abstract: *Today, artificial intelligence has permeated all aspects of life. Human health is an important issue that has also received the attention of artificial intelligence. Electronic health records (EHRs) can be used to discover ethical risk patterns in artificial intelligence in EHRs. However, much less has been said about the readiness of EHR data for such data mining projects. The main goal of this article is to use artificial intelligence based on machine learning in the analysis of frequent patterns to detect moral risk patterns. In the proposed method, using the FP-Growth-based method, identification of frequent moral risk patterns has been done. These patterns are used as attributes for the input of categories. The best result for identifying risk patterns has been obtained in classification based on neural network. The evaluations show the superiority of the proposed method in accuracy 97.63, accuracy 99.70, recall rate 99.89, and prediction error 0.06, and the execution time is less than 0.6 seconds, showing the superiority of the proposed method compared to There are other studies.*

Keywords: *AI Healthcare, identification of frequent patterns, risk patterns, potential futures.*

Information about the authors

Ali A. Alaidany
Fuel and Energy Techniques Engineering Department, Shatt Al-Arab University College, Basra ,Iraq

Zahraa Abbas Hasan Tayyeh
Department of Computer Engineering Technical College AL-Ayen Iraqi University

Marwah M. Mahdi
PhD Candidate, Department of Computer Eng., Isfahan (Khorasgan) Branch, Islamic Azad University, Isfahan, Iran

1. INTRODUCTION

The ever-increasing costs of the health care system in the world have attracted considerable attention. This issue has received more attention after the corona epidemic in the world. Central to the ideas that aim to curb this trend is the use of technology in the form of electronic health records (EHRs) . EHR contains patient information such as demographic information, medications, laboratory test results, diagnosis codes, and procedures. A comprehensive review of EHRs can lead to improved patient health management, as EHRs contain detailed prognostic information for large patient populations [1].

Smart health care can be considered a combination of traditional health care, biosensors, wearable devices, information and communication technology (ICT) and emergency response systems, which consist of various functional mechanisms such as information and communication technology, program

Smartphones and advanced data mining techniques are used to record and analyze medical information and finally make medical decisions (especially in emergency situations) [2-4].

Generally, the stored digital data of people from smart devices is called electronic health record (EHR). EHR contains a wide range of information about individuals such as demographics, observations, tests, diagnostic reports, treatments, therapies, prescriptions, and allergies [5]. By keeping track of people's EHRs over a long period of time, changes in their health over time can be tracked [6]. Therefore, these data over time can be a good source for data analysis and forecasting models. However, it is difficult to provide an effective model for the health domain because the real data has noise and heterogeneity [7].

One area that preventive healthcare focuses on is risk assessment. Usually, risk assessment is done by experts based on their experience and is therefore limited to the ability of experts to process them in a short period of time[8].

Many ethical risk assessment systems have been used in the medical field to help professionals make decisions, some of which are APACHE, SAPS, and MPM for intensive care patients [9]. Based on their general application, these methods are defined based on factors provided by experts' experience and then validated based on studies [10]. Based on the advances in computing technology and EHR availability, frameworks in machine learning and data mining have been presented that can be used in decision making. In recent years, methods for risk assessment have been presented [11], most of them focusing on EHR. GHE data recordings and the challenges they face have not been comprehensively investigated [12]. Various researches have been done in order to evaluate the risk challenge. In [13], they proposed a method called FLR-Mining, in which a distributed method with load balancing was proposed to discover frequent patterns in distributed processing systems. In [14], they have presented a knowledge-based diagnosis method that allows content-based systems to adapt their behavior over time using the analysis of mass data stored in the systems. In [15], they have presented a model that uses data related to smart homes to identify human behaviors in order to use them to predict and monitor people's health. In [16], they used a method by exploring repeated items to discover moral risk patterns in heart patients. In [17], recent works have reviewed the use of deep learning technologies to advance the health care field. In [18], deep learning models have been proposed for predicting and diagnosing hip fracture using health and care variables of patients and people under observation. In [11], they investigated the complexity of analyzing massive and diverse medical data. In this regard, the effect of massive data in the health sector and the various tools available in the Hadoop ecosystem to deal with them have been investigated. In [12], a new framework called HealthFog is proposed to integrate group deep learning in edge computing devices in cloud computing and it is used for a real application in automatic heart disease analysis. [13] have presented a COVID-19 Early Warning System (CovEWS), a risk scoring system to assess the risk of mortality due to COVID-19 using data totaling more than 2863 years of observation time from a cohort. Consists of 66,430 patients. In [19] using two complementary methods including machine learning techniques and real unsupervised methods, presented in electronic health records (EHRs) of 3,009,048 individuals in England using primary care data from the Clinical Research Data Linkage (CPRD).) Used. [20] discussed that the availability of electronic health records represents an excellent research opportunity on several diseases, one of the most relevant public health problems today. In [21], they proposed a method to set the best training sample with limited sample size for a classification model, which minimized its phenotypic/classification mean square error (MSE) in the proposed method. In this study, the research method is that the stored medical data (medical data set) is entered into the system as an input and after pre-processing, which includes removing some unnecessary data and preparing the data. It is for the main processing, first the classification algorithm and then the frequent pattern exploration and sequential (sequential) pattern exploration algorithms are executed on it. These explorations can be implemented by various methods, such as Apriori and FPGrowth, which is performed on the FP tree, and GSP, SPAM, and PrefixSpan sequential pattern exploration. In the first case, it is assumed that the method of FPGrowth algorithm is used to discover frequent pattern on Hadoop distributed environment, and any method can be used for sequential pattern.

In the following, this article is divided as follows. In the second part, the background of the research will be presented. In the third part, the proposed method is proposed. Finally, the results are discussed in the fourth section. In the last part, the conclusion will be drawn.

2. CONCEPT OF PAPER:

Electronic health data

Broadly defined, EHRs are longitudinal data (in electronic format) collected during the delivery of health care. EHRs generally contain demographic, vital, administrative, claims (medical and pharmaceutical), clinical, and patient-centered information (eg, from health-related quality of life instruments, home monitoring devices, and frailty or care assessment). Watchful) . The type of EHR varies greatly around the world [17].

EHRs are designed to optimize diagnosis and clinical care, which increases their relevance for clinical research. An EHR may include individual components of care (eg, primary care, emergency department, and intensive care unit) or reflect data from an integrated hospital-wide or inter-hospital system. EHRs also reflect evolving technology capabilities or external influences (eg, changes in the type of data collected related to coding or reimbursement practices) and may change over time. EHRs emerged mainly as a means to improve the quality of health care and collect health data. EHRs may potentially be used to assess study feasibility, facilitate patient registration, simplify data collection, or conduct fully EHR-based observational studies, randomized clinical or postmarketing registration studies, or comparative effectiveness studies. be used. Various applications of EHR for observational studies, safety monitoring, clinical research and regulatory purposes are shown in Table 1 [18].

Table (1) different applications of EHR

Type	Example	Condition
Field studies	Health utilization, drug use, epidemiology (incidence/prevalence), natural history, risk factors	It is widely used and accepted
Safety monitoring	Traditional post-marketing safety surveillance	It is widely used and accepted
	Active monitoring (eg, Sentinel a)	Emerging
Clinical research	Hypothesized	accepted
	Feasibility assessments	accepted
	Improve performance, follow instructions	accepted
	Patient recruitment	Emerging
	Comparative Effectiveness, Health Technology Assessment	Emerging
	Practical experiments (such as PROBE design)	Emerging
	Point-of-care randomization	Emerging
	Register randomized trials to test new interventions	Emerging
	Data source to populate eCRF (eliminate or minimize the need for data extraction/data entry)	Emerging / Potential
	Determination of end point or SAE	

Frequent patterns

Exploring association rules and identifying frequent patterns is aimed at discovering interesting and important relationships between information items in large databases and transaction stores. Associative data mining and extracting associative rules from the dataset was first presented by Agrawal to discover the knowledge and purchasing patterns of users of a store.

An interesting method that obtains the set of frequent items without generating a set of candidate items is the FP-Growth algorithm, which uses a divide and solve strategy. This method divides the database into a set of databases, each of which has a frequent item, and explores each database separately. In the first scan of the database, as a priori, the set of single-member items and their support is determined. The set of frequent items are arranged in descending order of their support.

Then, a tree is created as follows: First, the root of the tree is created with the tag null. After that, the database is scanned for the second time. The items of each transaction are processed in L order and a branch is created for each transaction. In order to facilitate tree navigation, a table is created in which each item refers to its own place in the tree. . The tree is complete after scanning all transactions.

This algorithm is one of the most well-known data mining algorithms for exploring frequent items on huge databases. By traversing the database twice, it extracts repeating items as well as association rules.

In the first navigation, it first extracts the repeated items.

- The first scan first extracts repeated items from the first order and then stores them in a list called L in descending order based on their support.
- Sorting the items in a table called the header table in the same order. The header table not only contains the items and their support value, but also contains a pointer to the same items in the FP-Growth tree.
- In the second navigation, the items of each transaction are placed in the order in the L list as a path in the FP-Growth tree. Separate paths that have the same prefix are merged when added to the tree to reduce memory consumption.

In general, three major advantages can be mentioned for FP-growth:

- **First**, FP-growth compresses the entire database into a smaller data structure (FP-Growth tree), and as a result, browsing the entire database is done in only two steps.
- **Second**, it prevents random generation of duplicate candidate data items.
- **Third**, using the conditional pattern tree to explore frequent data items and the result is to reduce the search space.

3. PROPOSED METHOD:

The main purpose of this research is to provide a suitable and optimal method for data mining in the community domain in order to classify information related to EHR data and other valid methods. Since the method of maximum repeating patterns makes the information more compact than the method of repeating closed patterns, it is expected that by using the FP-Growth method for maximum repeating patterns, suitable results can be obtained for extracting association rules. Therefore, in general, the purpose of this research can be in the form of reducing memory consumption by extracting valuable and important rules using the FP-Growth technique and removing unimportant features, increasing the speed of information processing by extracting the maximum features and rules, and finally Improving the accuracy of EHR data classification by applying classification techniques such as decision tree and so on. The problem of finding a suitable class item using the association category can be divided into five main steps

First step: extracting frequent patterns

Second step: find all the frequent rules

Third step: generation of all category association rules from among the repeated rules extracted in step 2, whose confidence level is higher than the minconf threshold.

fourth step: choosing a subset of the group association rules to form the class group from the items obtained from the second step.

fifth step: measuring the quality of the obtained class items on the test data.

Block diagram of the proposed method

In the proposed method to reduce the number of provided patterns, the time and memory cost will be greatly reduced. The proposed research method is based on the methods of condensing frequent patterns such as the maximum frequent pattern and combining them with vertical data presentation algorithms such as node list and FP-Growth. Node list and FP-Growth methods, due to their tree and hierarchical nature, reduce the number of extracted patterns and thus reduce pattern retrieval time. Since the method of maximum repeated patterns makes the information package more compact than the method of repeated patterns, it is expected that by using the FP-Growth method for maximum repeated patterns, suitable results can be obtained for the extraction of association rules. After applying the FP-Growth method on EHR data, a subset of features that have the greatest impact on the output can be extracted.

After extracting a subset of features that have a greater impact on the corresponding output, a large amount of relatively worthless data is not considered in the classification operation. The training of models based on KNN and decision tree algorithms is based on only the data related to the features extracted in the previous step. Therefore, after applying the output related to the FP-Growth algorithm section, the models related to the KNN algorithms and the trained decision tree and the new samples are evaluated and the corresponding class is assigned to it. Figure 1 shows the block diagram of this method.

Data preprocessing and preparation

Data pre-processing is one of the main pillars of data mining, which ignores useless samples or features and actually removes these data from the original data to improve the accuracy and speed of model execution. be made Considering that the mentioned examples are known as useless records or have missing values, they should be removed from the corresponding data source or replaced with default values. It is possible that during data collection, some columns are empty or have unreasonable data. At this stage, it is necessary to identify such data. Therefore, according to the needs of the problem and also to increase the accuracy of the relevant model, samples that have useless values will be removed from the data. Up to this stage, the operation of cleaning and removing outliers from the data, which is part of the data pre-processing, has been applied to the data source.

Frequent rule extraction system

As shown in the block diagram of the proposed method, the proposed method, which is an improvement of the FP-Growth algorithm, has two general phases, the first existing phase is extracting frequent and optimal rules from the data set. In this phase, the main goal is to extract a subset of features that have a great impact on determining the output of the problem. In other words, the main goal is to extract a set of features that are very effective in determining EHR and have a higher score than other subsets of features

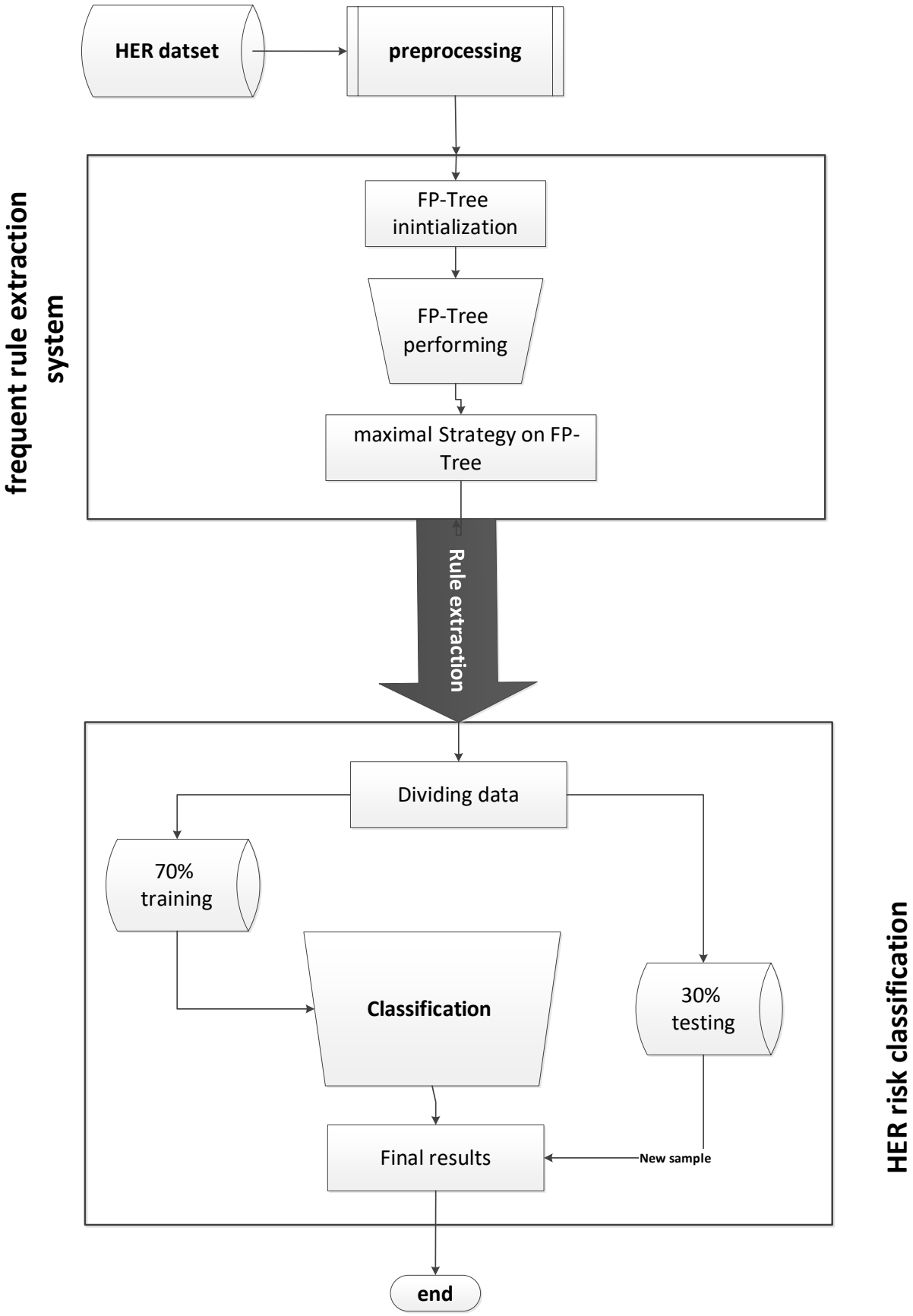

Figure (1) block diagram of the proposed method in identifying the risk of frequent patterns

Initialization of FP-Growth Algorithm

After applying pre-processing on the original EHR data, by this point the data is available which is standardized and does not contain any useless values or incomplete records. In the proposed method, variables are defined for the FP-Growth algorithm, and by default, these variables should be set first and then the main algorithm should be applied to the data set. In this article, 4 basic parameters are defined for the FP-Growth algorithm, which are:

- **Min Support**: This parameter shows the minimum value that any subset of features can be included among the optimal and superior rules at least with this value. In other words, examples that have a probability greater than or equal to this parameter are considered to be repeated rules.
- **Max-Support**: This parameter also shows the maximum probability. In other words, a 100 sample can be among the most repeated rules.
- **Min Length**: shows the minimum number of variables (features) that can participate in a rule.
- **Max Length**: shows the maximum number of variables (features) that can participate in a rule. In fact, the higher this value is, the combination of more features will have an effect on the output and more complete rules will be extracted.

Extracting the set of repeated items using the FP-Growth algorithm

After initializing the parameters related to the FP-Growth algorithm, the data that is pre-processed and also prepared for execution is applied to the FP-Growth algorithm. This step was done using C# programming language and Tanagra data mining tool. It should be noted that the default values of Min Support=0.8, Max Support=1, Min Length=4, and Max Length=14 are considered. After applying the FP-Growth algorithm, the repeated elements and samples are extracted normally without applying the maximum strategy. Considering that the number of features in the EHR dataset is a maximum of 14, in applying the FP-Growth algorithm, the maximum number of extracted subsets (Max Length) is 14. Therefore, by applying the Maximal strategy to the output of the FP-Growth algorithm, more complete outputs are provided, and as a result, rules and a subset of features that have maximum and optimal values are extracted.

EHR classification system

In the block diagram of the proposed method, which is shown in Figure 1, the features that were extracted in the previous step by the maximum algorithm were separated from the data, and in order to test the model, they were classified into two parts, training data and data. Tests are divided in order to train the relevant model and evaluate new samples. Then the training data, which constitutes 70 of the total data, was entered into the decision tree algorithm and led to the production of a model. After generating the model which is in the form of a tree, the training data which is 30 of the data is entered into the decision tree algorithm to evaluate the accuracy of the proposed model. Finally, one by one, the data is categorized by the tree and the accuracy is calculated.

4. RESULTS

Electronic health records (EHRs) provide opportunities to enhance patient care, incorporate performance measures into clinical practice, and improve the identification and recruitment of eligible patients and health care providers into clinical research. On a macroeconomic scale, EHRs (by enabling pragmatic clinical trials) may help evaluate whether new treatments or innovations in health care delivery lead to improved outcomes or savings in health care. In the following, the proposed method is examined.

Research database

All the data of this research is quantitative, which was extracted from the records of patients visiting more than 130 clinics and hospitals in the United States of America from 1999 to the end of 2008 for 10

years. The data of this statistical population of this research includes 53 characteristics presented in patients' referrals to hospitals and treatment clinics.

All the data in this research is based on the database presented in the research "The impact of HbA1c index measurement on the readmission rate of patients in the hospital based on the analysis of 70,000 clinical data for patients by Mr. Beta Sterk et al in 2014". It has been prepared and arranged.

All the statistics, information and data that have been used in this research were extracted and analyzed based on the data available in the aforementioned research available at https://www.openml.org/d/4541. The diagram of table 2 shows the comparison of the accuracy of the proposed method in this research compared to other methods for classifying EHR data. As can be seen in the graph, the accuracy of the proposed method is better than other methods. Therefore, the accuracy of the proposed method has improved by 1.005.005 compared to the k-nearest neighbor algorithm, 1.0013.000 compared to the neural network, and 1.0031.000 compared to the Neobiz algorithm.

Table (2) Comparison chart of EHR classification accuracy of the proposed method compared to other methods

Proposed method	KNN accuracy	ANN accuracy	NB accuracy
99.94	99.89	99.81	99.63

Table 3 shows the accuracy comparison of the proposed method compared to other methods. Table 4 also shows the comparison of the calling criteria of the proposed method compared to other methods.

Table (3) comparing the precison of EHR classification of the proposed method compared to other methods

Proposed method	KNN	ANN	NB
99.70	96.37	95.31	99.00

Fi Table (4) comparison chart of EHR classification Recall

Proposed method	KNN	ANN	NB
99.86	98.15	96.50	97.11

Table 5 shows the error comparison of the proposed method in this research compared to other methods for classifying EHR data. As can be seen in the above diagram, the error of the proposed method is better than other methods. Therefore, the error of the proposed method has improved 1.83 compared to the nearest neighbor algorithm, 3.16 compared to the neural network, and 6.6 compared to the Neobiz algorithm.

Table (5) comparing the EHR classification error of the proposed method compared to other methods

Proposed method	KNN	ANN	NB
0.06%	0.11%	0.19%	0.37%

One of the most important parameters in these problems and the extraction of repetitive rules based on the FP-Growth method is the discussion of time. The proposed method is implemented in two steps. The processing time of extracting the subset of features that has the most impact on the output and the stage of classification of new samples has a time based on milliseconds. Table 6 also shows the comparison of the proposed method with other methods. Table 7 also shows the comparison of the average execution time of the proposed method with other methods. As it turns out, the proposed method has improved on average compared to the DCI_PLUS method by about 2.5 times, compared to the dCHARM method by about 11.46 times, and finally by about 66.5 times compared to the NAFCP method.

Table (6) Comparison of the execution time of the proposed method with other methods (milliseconds)

	MinSupport			
methods	0.6	0.5	0.4	0.2
Proposed	53	95	150	500
DCI_PLUS	100	200	300	1400
dCHARM	600	700	850	7000
NAFCP	200	300	420	3600

5. CONCLUSION

Medical institutions use EHRs to record a series of medical events, including diagnostic information (diagnosis codes), procedures performed (procedure codes), and admission details. Many data mining technologies are used in EHR datasets to discover knowledge that is valuable to medical practice. The knowledge found is useful for developing treatment programs, improving health care and reducing medical costs, in addition, it can be of further help to predict and control the outbreak of epidemics. An ideal EHR management system requires capabilities to acquire, store, organize, and analyze health-related data. In addition to flexibility, interpretation, reporting and display of such a system should also take into account the statistical and spatial aspects of the problem. Therefore, the use of new technologies such as data mining and pattern finding can be effective in health studies. After simulating the proposed method, it was observed that the proposed method is better than other methods in terms of memory consumption, execution time, and error.

REFERENCE

1. O. P. Singh, A. Anand, A. K. Agrawal, and A. K. Singh, “Electronic Health Data Security in the Internet of Things through Watermarking: An Introduction,” *IEEE Internet of Things Magazine,* vol. 5, no. 2, pp. 55-58, 2022.
2. S. S. Ghasemi, A. A. K. Zarchi, Y. Alimohamadi, M. Raei, and M. Sepandi, “Survival analysis and its related factors among patients with breast cancer referred to a military hospital in Tehran,” *EBNESINA,* vol. 25, no. 1, pp. 4-12, 2023.
3. A. Rehman, M. Harouni, M. Karimi, T. Saba, S. A. Bahaj, and M. J. Awan, “Microscopic retinal blood vessels detection and segmentation using support vector machine and K-nearest neighbors,” *Microscopy research and technique,* vol. 85, no. 5, pp. 1899-1914, 2022.
4. M. Karimi, M. Harouni, and S. Rafieipour, "Automated medical image analysis in digital mammography," *Artificial intelligence and internet of things*, pp. 85-116: CRC Press, 2021.
5. S. S. Seyed Abolghasemi, M. Emadi, and M. Karimi, “Accuracy Improvement of Breast Tumor Detection based on Dimension Reduction in the Spatial and Edge Features and Edge Structure in the Image,” *Majlesi Journal of Electrical Engineering*, 2023.
6. M. Emadi, M. Karimi, and F. Davoudi, “A Review on Examination Methods of Types of Working Memory and Cerebral Cortex in EEG Signals,” *Majlesi Journal of Telecommunication Devices,* vol. 12, no. 3, 2023.
7. N. Mahmodi, M. Sepandi, A. S. Mohammadi, and H. Masoumbeigi, “Epidemiological aspects of occupational exposure to sharp tools among nurses in a military hospital in Tehran, Iran,” *Iranian Journal of Health, Safety and Environment,* vol. 2, no. 4, pp. 374-379, 2015.
8. R. Ramezanian, A. Peymanfar, and S. B. Ebrahimi, “An integrated framework of genetic network programming and multi-layer perceptron neural network for prediction of daily stock return: An application in Tehran stock exchange market,” *Applied soft computing,* vol. 82, pp. 105551, 2019.

9. M. Ko, M. Shim, S.-M. Lee, Y. Kim, and S. Yoon, “Performance of APACHE IV in medical intensive care unit patients: comparisons with APACHE II, SAPS 3, and MPM0 III,” *Acute and critical care*, vol. 33, no. 4, pp. 216, 2018.
10. D. Shikh-hasani, M. Alifarri, and B. Karimi, “Measuring efficiency score by cross-efficiency method in data envelopment analysis and its relation to profitability and risk in banks admitted to Tehran stock exchange,” *Management Accounting*, vol. 13, no. 46, pp. 103-119, 2020.
11. R. Yarahmadi, R. A. Dizaji, A. Hossieni, A. Farshad, and S. Bakand, “The Prevalence of Needle sticks injuries among health care workers at a hospital in Tehran,” *Iranian Journal of Health, Safety and Environment*, vol. 1, no. 1, pp. 23-29, 2014.
12. M. Harouni, M. Karimi, A. Nasr, H. Mahmoudi, and Z. Arab Najafabadi, "Health Monitoring Methods in Heart Diseases Based on Data Mining Approach: A Directional Review," *Prognostic Models in Healthcare: AI and Statistical Approaches*, pp. 115-159: Springer, 2022.
13. K. W. Lin, and S.-H. Chung, “A fast and resource efficient mining algorithm for discovering frequent patterns in distributed computing environments,” *Future generation computer systems*, vol. 52, pp. 49-58, 2015.
14. A. R. M. Forkan, I. Khalil, A. Ibaida, and Z. Tari, “BDCaM: Big data for context-aware monitoring—A personalized knowledge discovery framework for assisted healthcare,” *IEEE transactions on cloud computing*, vol. 5, no. 4, pp. 628-641, 2015.
15. A. Yassine, S. Singh, and A. Alamri, “Mining human activity patterns from smart home big data for health care applications,” *IEEE Access*, vol. 5, pp. 13131-13141, 2017.
16. R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, “Deep learning for healthcare: review, opportunities and challenges,” *Briefings in bioinformatics*, vol. 19, no. 6, pp. 1236-1246, 2018.
17. P. Schwab, A. Mehrjou, S. Parbhoo, L. A. Celi, J. Hetzel, M. Hofer, B. Schölkopf, and S. Bauer, “Real-time prediction of COVID-19 related mortality using electronic health records,” *Nature communications*, vol. 12, no. 1, pp. 1058, 2021.
18. V. Kuan, H. C. Fraser, M. Hingorani, S. Denaxas, A. Gonzalez-Izquierdo, K. Direk, D. Nitsch, R. Mathur, C. A. Parisinos, and R. T. Lumbers, “Data-driven identification of ageing-related diseases from electronic health records,” *Scientific reports*, vol. 11, no. 1, pp. 2938, 2021.
19. C. A. Nelson, R. Bove, A. J. Butte, and S. E. Baranzini, “Embedding electronic health records onto a knowledge network recognizes prodromal features of multiple sclerosis and predicts diagnosis,” *Journal of the American Medical Informatics Association*, vol. 29, no. 3, pp. 424-434, 2022.
20. J. Carmona-Pírez, B. Poblador-Plou, A. Poncel-Falcó, J. Rochat, C. Alvarez-Romero, A. Martínez-García, C. Angioletti, M. Almada, M. Gencturk, and A. A. Sinaci, “Applying the FAIR4Health solution to identify multimorbidity patterns and their association with mortality through a frequent pattern Growth association algorithm,” *International journal of environmental research and public health*, vol. 19, no. 4, pp. 2040, 2022.
21. S.-H. Lee, Y. Ma, Y. Wei, and J. Chen, “Optimal sampling for positive only electronic health record data,” *Biometrics*, vol. 79, no. 4, pp. 2974-2986, 2023.