Application of Data Mining and Knowledge Discovery in Medical Databases

Ahmed Mahdi Abdulkadium
Specialization Computer Science (Information System), Al-Qasim Green University, Babylon, Iraq.
E-mail: ahmed_mahdi@uoqasim.edu.iq

Raid Abd Alreda Shekan
Specialization Computer Science (Information System), College of Education for Pure Science, University of Babylon, Iraq.
E-mail: pure.raed.abd@uobabylon.edu.iq

Haitham Ali Hussain
Specialization Computer Science (Information System), Management Technical College of Basra, Southern Technical University, Basra, Iraq.
E-mail: Haitham.ali@stu.edu.iq

Received September 29, 2021; Accepted December 21, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V19I1/WEB19329

Abstract

While technical improvements in the form of computer-based healthcare information applications as well as hardware are enabling collecting of and access to healthcare data wieldier. In this context, there are tools to analyse and examine this medical data once it has been acquired and saved. Analysis of documented medical data records may help in the identification of hidden features and patterns that could significantly increase our understanding of disease onset and treatment therapies. Significantly, the progress in information and communications technologies (ICT) has outpaced our capacity to assess summarise, and extract insight from the data. Today, database management system has equipped us with the fundamental tools for the effective storage as well as lookup of massive data sets, but the topic of how to allow human beings to interpret and analyse huge data remains a challenging and unsolved challenge. So, sophisticated methods for automated data mining and knowledge discovery are required to deal with large data. In this study, an effort was made employing machine learning approach to acquire knowledge that will aid various personnel in taking decisions that will guarantee that the sustainability objectives on Health is achieved. Finally, the present data mining methodologies with data mining methods and also its deployment tools that are more helpful for healthcare services are addressed in depth.

Keywords

Medical Databases, Healthcare Data, Data Mining, Knowledge Discovery, Reliability, Big Data, Artificial Intelligence.

Introduction

The amount of data has increased at an incredible rate thanks to the rapid growth of computer software/hardware system and internet technology. The amount of knowledge on the planet is expected to increase every 20 months. Databases are growing in both size and number at an even faster rate. Because even simple transactions like making a phone call, using a credit card, or going to the doctor are typically registered on a computer, the automation of business operations creates an ever-increasing stream of data.

However, despite its importance, the definition of big data varies from field to field, with the result that it now affects all aspects of life and society (Herland M., 2014) For computer scientists, a dataset that cannot be viewed, gathered, managed, interpreted or delivered immediately enough using typical IT and software-and-hardware techniques is referred to as “big data”.

Data is being collected and accumulated at an alarming rate across a wide range of sectors. Large amounts of data are generated rapidly throughout the healthcare industry, as well as trends show that using big data in the field medical applications greatly improves the efficacy of medical healthcare, while, also optimising business operations. Medical databases have a wealth of data on patients and their health conditions that is constantly being added to.

It is possible that the connections as well as patterns found within this data will lead to new medical discoveries. Inopportunely, only a few techniques have been created and put to use in the search for this untapped wealth of data. Because of this, a new generation of machine learning tools and methodologies is desperately needed to help humans mine the massive amounts of digital data that are being generated at an accelerating pace. The growing field of knowledge discovery in databases (KDD) is based on these concepts and methods.

The most difficult part is figuring out how to make sense of all the information that has been collected. Large databases have hidden patterns that can be discovered using KDD. The sheer volume of data may have a major impact on the quality and effectiveness of most Intelligent Data Analysis (IDA) methodologies, especially when there are redundant or

irrelevant characteristics, missing data, or false positives (Azevedo A., 2019) (Sohail MN, 2019).

Knowledge Discovery in Databases

Knowledge has indeed been described in multiple ways and there are many different points of view on its disposition. There are numerous sorts of knowledge, each needing a particular management method.

According to Chen, Herrera and Hwang (Chen M, 2018), what distinguishes the current technological advance is not the prominence of information and knowledge, but the integration of such information and knowledge to information generating and data processing/communication technologies. Knowledge has become a critical aspect for growth, given its potential as a strategic determinant for developing new policies, planning new activities and stimulating creativity inside businesses.

Information discovery is a relatively recent discipline where techniques taken from artificial intelligence, algorithms, statistics and mathematics are used to anticipate and explain new knowledge buried in amounts of raw data generally kept in databases (Girardi D, 2016). Data mining is also taken as associated of KDD (Girardi D, 2016), but data mining has been one of the aspects of process of KDD that requires advanced methods as well as the readiness to look at the possibility of hidden patterns that resides with in data (Lee S, 2016) (Albahri AS,2020).

The conventional approach of converting data into knowledge depends on manual interpretation and analysis. For example, in the health-care business, it is typical for professionals to periodically review current trends and patterns in healthcare information, say, on a regular schedule (Arji G, 2019).

Data Mining

An algorithmic approach to finding new, relevant, and fascinating information in databases is known as data mining (DM). Mathematical statistics, probability distributions, statistical inference, as well as neural networks are all applied disciplines of study in DM algorithms. It is possible to utilise some of these areas' methodologies to uncover hidden relationships among data, which may then be used to build models that anticipate behaviour or define certain common characteristics of the things being examined

The knowledge pyramid is frequently used to explain the relationship between data, knowledge, and understanding (Figure 1). With context, a great amount of data may be turned into information, and then the finding of general patterns in data represents knowledge about the investigated topic through assessment and consolidation of data (Vázquez-Ingelmo A,2020). For the 2nd phase, amongst the most used terminology is Data Mining (DM).

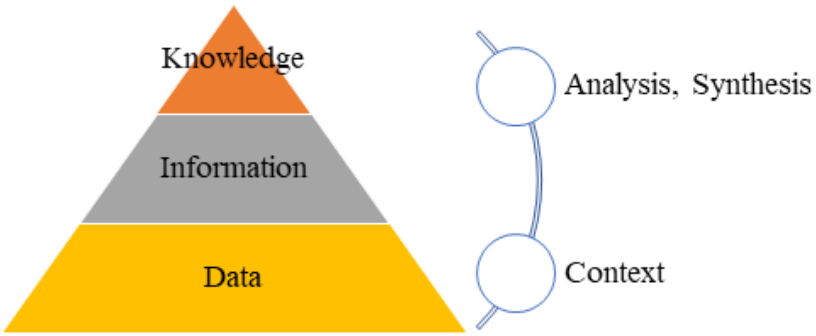

Figure 1 Typical knowledge pyramid

Research Objectives

Using structural modelling, we characterise the medical data mining and analysis process, that also begins with the transmission of the collected data from a computer-based patient record system (PRS) to a database server, followed by the formation of a medical database for pattern analysis.

Literature Review

Data mining and knowledge discovery in databases have recently received a lot of interest from researchers, business, and the media.

Healthcare organisations utilise data mining approaches like association, classification, as well as clustering to improve their capacity to draw relevant inferences about patient health from raw data (Lee S,2016).

Srimani and Koti (Srimani PK, 2014) provided a rough set approach for creating classification model from a collection of 360 observed breast cancer data samples, as well as the study demonstrated how rough set theory appears to be a valuable tool for identifying patterns in data.

Neto et al. (Neto C, 2019) used data pre-processing, RELIEF attribute selection, and Modest AdaBoost algorithms to retrieve knowledge features from breast cancer survival records in Thailand. They said that Mild AdaBoost outperforms Gentle AdaBoost, SVM, and C4.5, however, they did not reveal the size of the breast cancer databases or just use a separate set of data with a variable array of options and tuples to evaluate the algorithms' efficiency.

Diabetes is a serious public health issue in the United States, therefore diabetes registrations and archives with carefully gathered patient data have a long and storied data history. Perveen et al. (Perveen S, 2016) investigated one such diabetes data warehouse, demonstrating a means of using data mining tools as well as various analytical concerns, data challenges, and outcomes. They employed the decision tree classification technique with a binary target variable in Logistic regression and classification trees. They claimed that data mining may uncover unique relationships that would be extremely beneficial to physicians and administrators.

The goal of the study in (Pac M, 2021) was to evaluate the use of data mining methods in proteomics for malignancy detection/diagnosis and to investigate a unique analytic strategy using distinct feature extraction methods. In this study 3 serum SELDI-MS data were utilised to discover serum proteome patterns that separated ovarian cancer serum from non-cancer serum. Statistical testing and genetic algorithm-based approaches were employed for feature selection in a support vector machine-based approach.

The researchers at Mahoto et al. (Mahoto NA, 2021) employed data preprocessing, data transformations, and a data mining technique to discover how various observed characteristics correlate with patient survival. Decision rules were derived through the use of two separate data mining methods. In a decision-making algorithm, this set of principles predicted the survival of new patients who had not yet been observed by the researchers. Data mining was used to identify critical factors, which were then analysed for their potential medicinal importance. The theories developed in this study were put to the test on patients at four different dialysis centres. Patients on dialysis are predicted to live longer with the use of data mining, transformation of data, data segmentation, and decision-making algorithms, according to the study.

A unique bioinformatics technique presented by Krallinger, Leitner, and Valencia (Krallinger M, 2010) finds disease candidate genes based on their expression characteristics. They combined text mining of biomedical journal articles with data mining of publicly accessible human gene expression (HGE) profiles using the eVOC anatomical ontology. They used a data set of 417 candidate genes, including 17 recognized illness

genes, to show that their strategy worked and could be used broadly. For 15 of the 17 disorders, they were able to pick the disease gene and decrease the candidate gene dataset to 63.3 percent of its original size (18.8 percent). When it came to genomics and gene expression, they found that applying data mining methods made it easier to link the two together.

In another research data mining was demonstrated by Yang et al. (Yang J, 2020) by demonstrating how patterns in huge databases may be discovered and extracted for usage. This study shows how data mining may help the pharmaceutical business make better decisions by improve the effectiveness of the decision-making mechanism. A problem in the pharmaceutical sector is medication side effects (Wang F., 2014).

Materials and Method

Databases and Machine Learning

A database is a conceptually integrated collection of data stored in one or more files and arranged to assist the efficient storage, updating, as well as retrieval of relevant material in database administration. For example, in a relational model, data is structured into files or tabular forms of fixed-length entries. Each record consists of an ordered list of values, one for each field. A data dictionary is a distinct file that stores information about every field's name and probable entries. A database management system (DBMS) is a set of techniques for obtaining, storing, and altering data from databases (Mahoto NA, 2021).

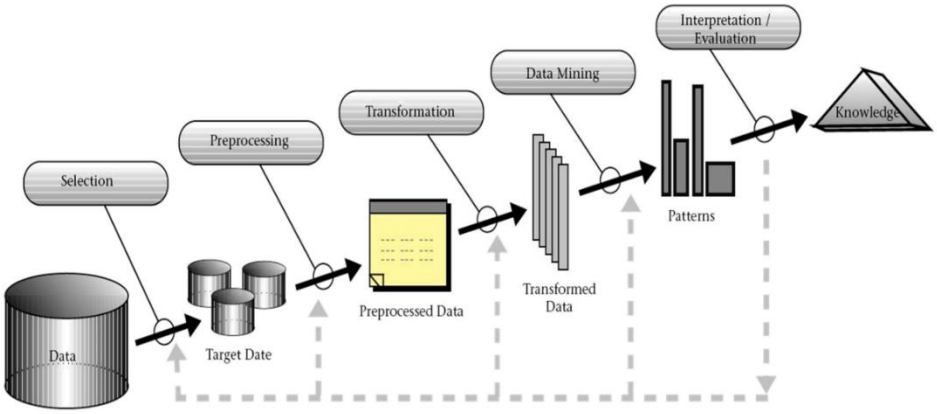

Figure 2 Steps in KDD process involving data mining (Arji G, 2019)

The data mining component of the KDD process is essential. When it comes to data mining, it involves selecting the appropriate data mining algorithm(s) and employing machine learning to generate previously unrecognized and potentially helpful and relevant data from the data stored in a database. This includes determining which models/algorithms as well

as variables are appropriate for a given situation, as well as matching a specific machine learning algorithm with the general standards of the KDD process. Classification, clustering, regression, summarization, and other data mining techniques are all available (Yang J ,2020).

Database

The computer-based patient record system (PRS) has been designated as the production database server for mining. Demographic, research results, issues, treatments, sensitivities, psychological and physical findings, as well as session reports are among the data gathered in PRS. The PRS database model employs a patented class-oriented approach that holds all patient records and information in a single reference record.

Knowledge Discovery and Data Mining

A knowledge discovery process, also known as KDD, is a very complicated non-linear process that includes not just data analysis but also data preparation, knowledge interpretation, and the application of newly obtained information. It is also known as knowledge discovery. Knowledge-based decision-making (KDD) is comprised of six critical processes, which are (Sohail MN, 2019) (Pac M, 2021): comprehending the issue domain; comprehending the data; preparing the data; data mining; evaluating the acquired knowledge; and applying that knowledge.

A non-linear process is one in which difficulties can be detected at any stage and the need to revert to some of the previous stages and restart the entire process from that point forward is necessitated by these problems. A KDD is not a one-pass procedure (Azevedo A. 2019), and each iteration results in a distinct perspective on the data being analysed. For example, we can detect many missing data points during the preprocessing step, and the prediction of these data points might be the aim of the first iteration of the KDD (Jatav S., 2018). In the second iteration, we can concentrate on the development of a model that might be used to forecast a patient's diagnosis based on the data (predictive data) or to evaluate patient who have symptoms and diagnoses that are like one another (descriptive data).

A. Understanding Medical Problem Domain

At the beginning of a KDD, we need to identify what type of knowledge should be found in the data. This requires understanding the problem domain. In case of medical data, the domain comes from medical area and, therefore, the main goals of this phase are (Islam MS, 2018):

a. Translation of medical goals into significant parameters for DM.
b. Determining success criteria from the medical point of view under DM.

B. Understanding Data

When we have a good understanding of the medical issue area, we can analyse the data that is accessible. This analysis determines which data will be utilised and which information will be required. The primary purpose of this stage is to generate a dataset for the subsequent KDD processes.

Data mining is a multifaceted topic that benefits from database management systems, statistics, machine learning, and pattern recognition (Islam MS, 2018). However, this method is not yet widely used in medical research, multiple studies have shown that data mining has great promise for developing disease-prediction models, estimating patient risk, and assisting clinicians in making therapeutic choices (Ghorbani R, 2019) (Kumar SR, 2019).

Data-mining Models

There are two types of models in data mining: explanatory models as well as predictive analytics. When it comes to other parameter estimates, where, predictive models are commonly used to anticipate unknown or future values, whilst descriptive methods are frequently employed to uncover patterns that explain data that can be comprehended by people (Rahimian F, 2018).

1. Data Mining Tasks

The most important function of these algorithms was to scan the information and transform alphabetic fields into quantitative variables, which allowed for statistical analysis to be performed (Lashari SA, 2018). After determining whether data values were obtained during or relating to the infant's preterm course, the script checked to make sure that no more than one value for about the same variable was present. If such values were found, the value that was recorded the closest to delivery or conceptions, dependent on assessed data quality for the parameter, was imported into the final dataset and used as the starting point for the analysis.

2. Data Mining Techniques

i. Rough Set

Rough set approach can be used to find structural correlations in imprecise or noisy data. It is applicable to characteristics with discrete values. Continuous-valued properties must thus

be discretized before they can be used. The construction of equivalence classes within the supplied training data is the foundation of rough set theory. All the data tuples in an equivalence class are indistinguishable, which means that the samples are equal in terms of the characteristics representing the data. It is typical in real-world data that certain classes cannot be identified based on the given attributes. Rough sets can be used to define such classes crudely or "roughly".

ii. Artificial Neural Network

An Artificial Neural Network (ANN) is a method of processing information. It functions similarly to how the human brain functions. ANN is made up of several interconnected processing units that work together to process data. They also provide significant results in various machine learning classification processes. We can use neural networks for more than just categorization. It may also be used for continuous target attribute regression. Neural networks have a wide range of applications in data mining, which is employed in a variety of industries. For instance, economics, forensics, and pattern recognition. After thorough training, it may also be utilised for data categorization in vast amounts of data.

3. Evaluation Metrics

While working with binary classification difficulties, we may always designate one class as a positive class the other as a negative class by reversing the labels. Examples of both good and bad outcomes are included in the test set. Several of the assignments made by a classifier are incorrect; nonetheless, some of the allocations were made incorrectly. The number of True Positives, True Negatives, False Positives, and False Negatives should then be established to evaluate the categorization findings.

To compare the performance of the algorithms, the following equations can be used:

$$Sensitivity = \frac{TruePositive}{TruePositive + FalseNegative}$$

$$Specificity = \frac{TrueNegative}{TrueNegative + FalsePositive}$$

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive}$$

$$Accuracy = \frac{TruePositive + TrueNegative}{TruePositive + TrueNegative + FalsePositive + FalseNegative}$$

Results and Discussion

Based on two years of cancer data, we provide early findings from a factor analysis, and we contrasted these findings with those from model research in the area. Finally, we cover a few challenges that should be taken into consideration while analysing medical data using data mining process.

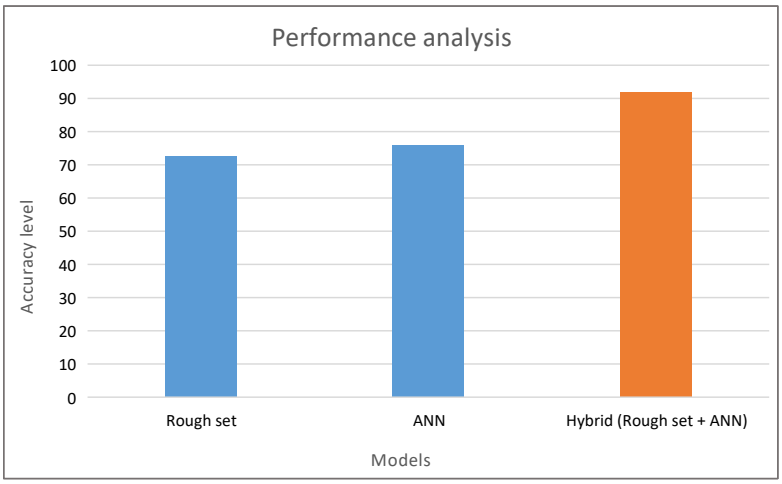

Figure 3 Performance of the modles

Table 1 Performance analysis of disease prediction

Performance	Unsuccessful (0)	Successful (1)
MSE	0.09278654	0.12054762
MAE	0.14351851	0.19346595
R	0.78895482	0.78549521
Min Abs. Error	0.00234516	0.00521514
Max Abs. Error	1.05135564	1.05512412
Correct (%age)	89.5154263	91.8835954

Table 2 Performance analysis of ANN model

Performance	Desired Output	Actual Output
MSE	0.20986331	0.21214753
MAE	0.22441862	0.25574359
R	0.49951121	0.49925144
Min Abs. Error	0.00242512	0.00251572
Max Abs. Error	1.02513145	0.99125895
Correct (%age)	76	75.521452

When compared individual methodologies, the use of a combination of Rough set model as well as Artificial Neural Networks produces superior results. When mining information from a database, it has been discovered that using a hybrid strategy that combines the usage

of two or more machine learning tools produces better results than using a single approach for mining data from the database.

Discussion

It can be observed from the findings that have been presented thus far that machine learning methods are excellent classifiers, despite the fact that bagging, which is a hybrid algorithm that was used, did not demonstrate any particularly noteworthy performance when measured against the evaluation metrics. In addition, we discovered that the majority of unusable data in a sampled dataset comprised lawful values that were eliminated because they were not acceptable by our data mining methodology.

Conclusion

This study evaluated data mining and knowledge discovery in healthcare database applications to find out which one was most effective at collecting meaningful data. It is difficult to anticipate diseases using data mining tools, yet doing so saves time and improves diagnostic precision. Creating effective data mining tools for a certain application might minimise the number of human resources and experience required, which would save up both money and time. Exploring medical data for information is a risky endeavour since the data that is gleaned are noisy, irrelevant, and vast. The use of data mining technologies is extremely intriguing in this case since it allows for the exploration of medical data knowledge.

References

Herland, M., Khoshgoftaar, T.M., & Wald, R. (2014). A review of data mining using big data in health informatics. *Journal of Big data*, *1*(1), 1-35.

Wang, F., Zhang, P., Wang, X., & Hu, J. (2014). Clinical risk prediction by exploring high-order feature correlations. *In AMIA Annual Symposium Proceedings*, 2014.

Xu, R., Li, L., & Wang, Q. (2014). dRiskKB: a large-scale disease-disease risk relationship knowledge base constructed from biomedical text. *BMC bioinformatics*, *15*(1), 1-13. https://doi.org/10.1186/1471-2105-15-105.

Idri, A., Benhar, H., Fernández-Alemán, J.L., & Kadi, I. (2018). A systematic map of medical data preprocessing in knowledge discovery. *Computer methods and programs in biomedicine*, *162*, 69-85.

Azevedo, A. (2019). Data mining and knowledge discovery in databases. In Advanced Methodologies and Technologies in Network Architecture. *Mobile Computing, and Data Analytics*, 502-514.

Sohail, M.N., Jiadong, R., Uba, M.M., & Irshad, M. (2019). A comprehensive looks at data mining techniques contributing to medical data growth: a survey of researcher reviews. *In Recent developments in intelligent computing, communication and devices, 21-26.*

Chen, M., Herrera, F., & Hwang, K. (2018). Cognitive computing: architecture, technologies and intelligent applications. *Ieee Access, 6,* 19774-19783.

Girardi, D., Küng, J., Kleiser, R., Sonnberger, M., Csillag, D., Trenkler, J., & Holzinger, A. (2016). Interactive knowledge discovery with the doctor-in-the-loop: a practical example of cerebral aneurysms research. *Brain informatics, 3*(3), 133-143.

Lee, S., & Holzinger, A. (2016). Knowledge discovery from complex high dimensional data. In Solving Large Scale Learning Tasks. *Challenges and Algorithms,* 148-167.

Albahri, A.S., Hamid, R.A., Alwan, J.K., Al-Qays, Z.T., Zaidan, A.A., Zaidan, B.B., & Madhloom, H.T. (2020). Role of biological data mining and machine learning techniques in detecting and diagnosing the novel coronavirus (COVID-19): a systematic review. *Journal of medical systems, 44,* 1-11.

Arji, G., Safdari, R., Rezaeizadeh, H., Abbassian, A., Mokhtaran, M., & Ayati, M.H. (2019). A systematic literature review and classification of knowledge discovery in traditional medicine. *Computer methods and programs in biomedicine, 168,* 39-57.

Layman, E.J. (2020). Ethical issues and the electronic health record. *The health care manager, 39*(4), 150-161.

Vázquez-Ingelmo, A., García-Holgado, A., García-Peñalvo, F.J., & Therón, R. (2020). A meta-model integration for supporting knowledge discovery in specific domains: a case study in healthcare. *Sensors, 20*(15).

Srimani, P. K., & Koti, M. S. (2014). Knowledge discovery in medical data by using rough set rule induction algorithms. *Indian Journal of Science and Technology, 7*(7).

Neto, C., Brito, M., Lopes, V., Peixoto, H., Abelha, A., & Machado, J. (2019). Application of data mining for the prediction of mortality and occurrence of complications for gastric cancer patients. *Entropy, 21*(12).

Perveen, S., Shahbaz, M., Guergachi, A., & Keshavjee, K. (2016). Performance analysis of data mining classification techniques to predict diabetes. *Procedia Computer Science, 82,* 115-121.

Pac, M., Mikutskaya, I., & Mulawka, J. (2021). Knowledge Discovery from Medical Data and Development of an Expert System in Immunology. *Entropy, 23*(6).

Mahoto, N.A., Shaikh, A., Reshan, A., Saleh, M., Memon, M.A., & Sulaiman, A. (2021). Knowledge Discovery from Healthcare Electronic Records for Sustainable Environment. *Sustainability, 13*(16).

Krallinger, M., Leitner, F., & Valencia, A. (2010). Analysis of biological processes and diseases using text mining approaches. *Bioinformatics Methods in Clinical Research,* 341-382.

Yang, J., Li, Y., Liu, Q., Li, L., Feng, A., Wang, T., & Lyu, J. (2020). Brief introduction of medical database and data mining technology in big data era. *Journal of Evidence-Based Medicine, 13*(1), 57-69.

Jatav, S. (2018). An algorithm for predictive data mining approach in medical diagnosis. *International Journal of Computer Science & Information Technology (IJCSIT), 10.*

Islam, M.S., Hasan, M.M., Wang, X., & Germack, H.D. (2018). A systematic review on healthcare analytics: application and theoretical perspective of data mining. *In Healthcare, 6*(2).

Ghorbani, R., & Ghousi, R. (2019). Predictive data mining approaches in medical diagnosis: A review of some diseases prediction. *International Journal of Data and Network Science, 3*(2), 47-70.

Kumar, S.R., Gayathri, N., Muthuramalingam, S., Balamurugan, B., Ramesh, C., & Nallakaruppan, M.K. (2019). Medical big data mining and processing in e-healthcare. *In Internet of Things in Biomedical Engineering,* 323-339.

Rahimian, F., Salimi-Khorshidi, G., Payberah, A. H., Tran, J., Ayala Solares, R., Raimondi, F., & Rahimi, K. (2018). Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. *PLoS medicine, 15*(11).

Lashari, S.A., Ibrahim, R., Senan, N., & Taujuddin, N.S.A.M. (2018). Application of data mining techniques for medical data classification: a review. *In MATEC Web of Conferences, 150*.