Nile Journals

Nile Journal of Communication & Computer Science

Volume 4, Number 1, Dec 2022

/Journal Webpage: https://njccs.journals.ekb.eg

Print ISSN: 2805-2366
Online ISSN: 2805-2374

The Role of data mining in healthcare Sector

Nehal A.Mansour[1] and Hesham.A. Sakr[2]

1: Assistant lecturer at Nile Higher Institute for Engineering and Technology, Artificial intelligence Lab., Mansoura, Egypt

E-mail:nehal.anees.mansour@nilehi.edu.eg

2: Assistant professor-ECE-Department- Nile higher institute of engineering- Egypt

Email: heshamsakr535@nilehi.edu.eg

Abstract

Data mining is a valuable technique for identifying new patterns in healthcare organisations. Data mining is important in predicting and diagnosing various disorders in the health care business. It employs several ways to uncover hidden patterns, and these patterns can be utilised by physicians to assess the efficacy of medical treatments. the IoT medical sensors produce large amount of data at an amazing speed. Access to medical data is difficult due to confidentiality of patients' records and the importance of critical individual information. However, the huge amount of hospital and clinical data has a great potential for discovering relations and hidden patterns. The current paper investigates the utility of several data-mining methods in the healthcare sector and emphasises various applications in healthcare. The primary goal of this research will be to conduct a survey of existing data mining techniques utilised in the healthcare sector.

1. Introduction

Data mining is a strong method for extracting previously unknown patterns and important information from large datasets related to human disorders or diseases. Information technologies are progressively being deployed in healthcare organisations nowadays in order to meet the needs of doctors in their daily decision-making activities. Data mining techniques can be found in a variety of medical-related fields, including the medical device sector, the pharmaceutical industry, and hospital management. The goal of data mining applications is to extract useful and hidden knowledge from databases. Data mining techniques used in the healthcare business play an important role in disease prediction and diagnosis. Data mining is a rapidly evolving technique that is widely employed in biomedical sciences and research. . Modern medicine creates a large amount of data, which is maintained in a medical database. Medical data, for example, may include ECG, MRI, blood pressure, blood sugar, cholesterol levels, and so on, as well as physician interpretation. Extracting valuable knowledge and delivering scientific decision-making for disease treatment, diagnosis, and prediction from databases is becoming increasingly important, and are better positioned to satisfy their long-term needs. [1] [2]. Data mining, as described by Han et al., is the act of discovering interesting knowledge from

enormous amounts of complicated data held in data warehouses, databases, or other information repositories [3]. Data mining, according to Witten et al., is the process of collecting implicit, previously unknown, and possibly beneficial information from data [4]. Hand et al. defined data mining as the examination of observational data sets in order to discover patterns. Today's healthcare industry generates massive amounts of complex data about patients, hospital resources, disease diagnosis and prediction, electronic patient records, medical devices, and so on. This large complex data set is a critical resource to be processed and analysed for knowledge extraction, which allows for decision support and cost savings. Data mining provides a collection of tools and techniques. Techniques that can be applied to processed data to find previously unknown patterns, providing healthcare professionals with an extra source of knowledge for decision making.

As with the use of data mining tools, massive amounts of complex or voluminous healthcare data are being collected and made available for a variety of purposes, including doctors who use patterns by measuring clinical indicators, quality indicators, customer satisfaction, and economic indicators, and physician performance. viewpoints on optimising resource usage, cost efficiency, and evidence-based decision making, identifying high-risk patients and intervening proactively, optimising healthcare, and so on [6].

2. Problem definition

Healthcare system is a dynamic and changing environment so that the healthcare providers continually face new challenges every day such as ; the emergence of epidemics like coronavirus that need to be diagnosed without direct contact between patients and doctors, the emergence of many diseases that need early diagnosis like cancer, doctors' available time is usually limited, and medical errors hinder accurate diagnostic process . With the rapid development of computer software/hardware and internet of things (IoT) technology, the IoT medical sensors produce large amount of data at an amazing speed. Access to medical data is difficult due to confidentiality of patients' records and the importance of critical individual information. However, the huge amount of hospital and clinical data has a great potential for discovering relations and hidden patterns. Knowledge extraction from the generated data is strongly needed.Using data mining methods reduces time and cost in prognosis and diagnosis of diseases. It also has a special role in the treatment plan and medical.

3.Big data

Big data as an abstract concept currently affects all walks of life, and although its importance has been recognized, its definition varies slightly from field to field. In the field of computer science, big data refers to a dataset that cannot be perceived, acquired, managed, processed, or served within a given time by using traditional IT, software and hardware tools.

Across the medical industry, various types of medical data are generated at high speed. The sources of big data in healthcare are such as; mobile data and wearable devices, academic researches, electronic hospital records, medical claims, pharmacy claims, and government medical claims. Trends indicate that applying big data in the medical field helps to improve the quality of medical care and optimizes medical processes and management strategies [7-10].

4. Data mining

Computer scientists have made outstanding contributions to the application of big data and introduced the concept of data mining to solve difficulties associated with applications. Data mining (also known as knowledge discovery in databases) is the process of analyzing a massive amount of data to identify meaningful patterns and detect relations, which can lead to future trend prediction and appropriate decision making.

Data mining is a step in Knowledge Discovery in Database (KDD) which consists of data selection, data-pre-processing, data transformation, data mining, interpretation or evaluation of the model and using the discovered knowledge [11].

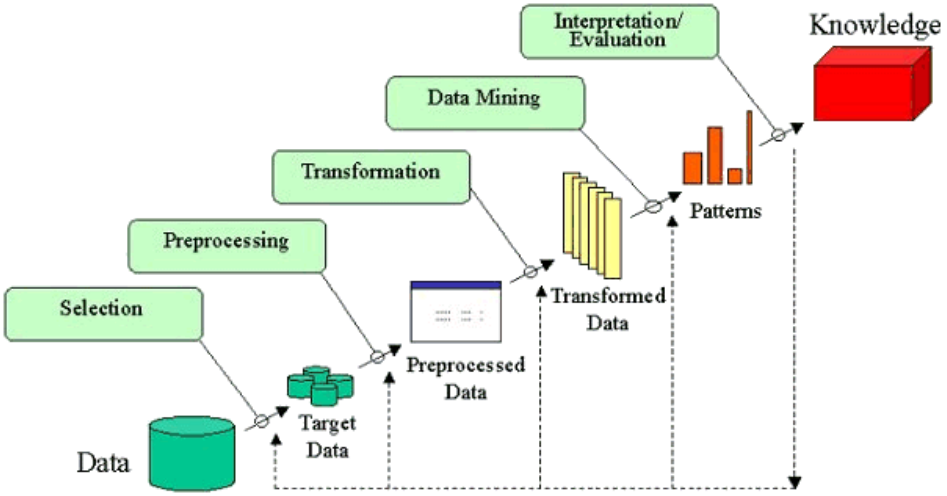

Fig.1, Steps of Knowledge discovery in database (KDD)

Data-mining technology does not aim to replace traditional statistical analysis techniques, but it does seek to extend and expand statistical analysis methodologies. Data mining technologies have been applied widely in various fields such as; marketing, telecommunication, disease detection, fraud detection, financial data analysis, intrusion detection, recommender systems, medical systems etc.

4.1.Role of data mining in healthcare

Recently, Data Mining is becoming popular and widely used in the healthcare sector because there is a demand for powerful and intelligent analytical methodology that can handle and analyse complex health data to make certain decisions regarding patient health. Data mining provides several benefits such as;

(i) Recognize and discover chronic diseases, their symptoms, possible reasons and identify medical treatment methods.

(ii) Grouping the patients having the same type of health issues or diseases so that healthcare providers can give them effective treatments [12].

(iii) It can be used for predicting the duration of the survival of patients in the hospital for medical diagnosis.

(iv) It can help the healthcare providers to determine effective treatments and best practices as well as to develop guidelines and care protocols.

(v) The analysis can be further extended for tracking of high-risk areas which are vulnerable to spread the diseases.

(vi) It can assist the healthcare researchers for making efficient healthcare policies, designing drug recommendation systems, identifying risk factors, complications, therapies, genetic effects, and environmental effects of diseases.

4.2.Data mining tasks

Data mining tasks can be divided into two categories which are descriptive and predictive .Descriptive mining tasks explore the general properties of data so that it can discover the relationship relating the data and results in a few clusters with the same or similar attributes .In contrast, predictive mining tasks perform deduction on the current data in order to make predictions about the variable value of a specific attribute based on the variable values of other attributes.

4.2.1.Predictive Models

- Classification derives a model to determine the class of an object based on its attributes. Binary and multiclass are the two methods of classification. In binary classification, only two possible values are assigned to the target class, while in multi-class classification, more than two values assigned to the target class.
- Prediction involves developing a model based on the available data and this model is used in predicting future values of a new data set of interest.
- Time series is a sequence of events where the next event is determined by one or more of the preceding events.
- Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables [13].

4.2.2.Descriptive Models

- Clustering means to find groups of data point (clusters) so that data points that belong to one cluster are more similar to each other than data points belonging to different clusters.
- Association rule mining is the process of finding if–then rules for discovering an association between variables in large data
- Summarization presenting the summary of generated data in an easily comprehensible and informative manner. In general, data can be summarized numerically as a table (tabular summarization), or visually as a graph (data visualization).

4.2.3.Data-mining methods

The data mining method depends on whether or not dependent variables (labels) are present in the analysis. Predictions with labels are generated through supervised learning, which can be performed by the use of linear regression, decision trees, random forest (RF) algorithm, and support vector machines (SVMs). In contrast, unsupervised learning involves no labels. Common

unsupervised learning methods include principal component analysis (PCA), association analysis, and clustering analysis.

5. Machine Learning

From a practical point of view, machine learning (ML) is the main analytical method in data mining. ML is defined as a field of study that gives computers the ability to learn without being explicitly programmed and it is a branch and subfield of AI. ML has a close relationship with statistics, applied mathematics, computer science, and software engineering disciplines. It uses data and algorithms to design and develop ML models that can learn to perform tasks by relying on patterns existing in the data. ML has been applied to solve several real time problems such as; medical diagnosis, image and speech recognition, self-driving cars, E-mail spam filtering , product recommendations , traffic prediction , automatic language translation , and virtual personal assistant. To classify and construct predictive models that are the main subject of most medical data mining articles, data is divided into two subsets. The large subset is called the training set used for constructing the classifier. The training set is made up of data set samples and their related class labels. The small subset is called the test set utilized for evaluating the classifier. Similar to the training set, the test set consists of data set records but the test set is independent of the training set, so it is not used for building the classifier. Generally, 70% of the data set is the training set and 30% is the test set.

There are several predictive models constructed by using different data mining methods. To evaluate the performance of these models, the statistical metrics are used which are comprised of accuracy, sensitivity, precision, and negative predictive value (NPV). These metrics are calculated as shown in table 2 [14-15].

Table 1, Confusion matrix which depicts how classification on

	Predicted true class	**Predicted negative class**
Actual True class	True Positive (TP)	False Negative (FN) Type 2 Error
Actual False class	False Positive (FP) Type 1 Error	True Negative (TN)

Table 2, Confusion matrix formulas.

Measure	Formula	Definition
Precision (PR)	TP / (TP + FP)	Metric indicates the number of the correct positive class.
Recall / Sensitivity (RE)	TP / (TP + FN)	Metric indicates the number of correct positive class made out of all positive class classification
Accuracy (AC)	(TP + TN) / (TP + TN + FP + FN)	The classifier's ability to classify the class label correctly.
Error (ER)	1-Accuracy	**The percentage of classification which is inaccurate.**
NPV	TN / (TN + FN)	**Metric indicates the number of the correct negative class.**

Classification of machine learning algorithms

The machine Learning Algorithms can be classified into five categories which are :are Unsupervised Learning, Supervised Learning, Reinforcement Learning and Semi-Supervised Learning, and deep learning .

5.1.Unsupervised Learning

In unsupervised learning, the class label of each sample is unknown. Moreover, the number of classes to be learned may not be known in advance. Unsupervised learning includes clustering, summarization, and association rules.

5.2. Supervised Learning

In Supervised Learning, the data is split into features and the corresponding class label, the model is built based on the mapping of features with the class label. Examples of Supervised Learning includes Regression algorithms namely Linear Regression, Lasso Regression, and Classification algorithms namely K Nearest Neighbor (KNN), Support Vector Machine (SVM), Decision tree,etc.

5.3. Reinforcement Learning

Reinforcement learning is a feedback-based learning method, in which a learning agent gets a reward for each right action and gets a penalty for each wrong action .The agent learns automatically with these feedbacks and improves its performance. The goal of an agent is to get the most reward points, and hence, it improves its performance.

5.4. Semi-Supervised Learning

As the name itself indicates, it is combination of unsupervised learning and supervised learning. In a semi-supervised learning, we use the labeled training data to build a model. However, labeling data is expensive. Instead, we convert non-labeled data to labeled data using the model and combined all data to refit a better model [16].

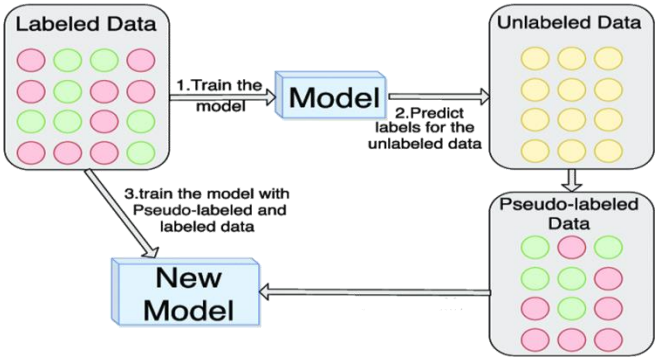

Fig.2, Semi-Supervised learning

5.5. Deep Learning

Deep Learning is a subfield of machine learning concerned with algorithms inspired by the structure and function of the brain called artificial neural networks. Machine learning and deep learning have similarities in training, testing data and finding an optimized model. The differences between deep learning and traditional artificial neural networks (ANNs) lie in the number of hidden layers, the connections of hidden layers and learning meaningful abstractions from the inputs. The designing of layers in deep learning is learned from data that is both unstructured and unlabeled not by developers. Deep learning applications include object detection in images, speech recognition and natural language understanding, medicine,etc.

5.6.Machine learning algorithms

The decision tree

The decision tree is one of the popular supervised machine learning methods used for classification, prediction and regression. Decision tree is a tree where each node shows an attribute, each branch shows a rule, and each leaf shows a categorical. This method can also handle both categorical and numerical data. It is built by using a training set of cases that are expressed in terms of a collection of features. A sequence of tests is carried out on the features in order to divide the training set into ever-smaller subsets. The process ends when each subset contains only cases belonging to a single category [17].

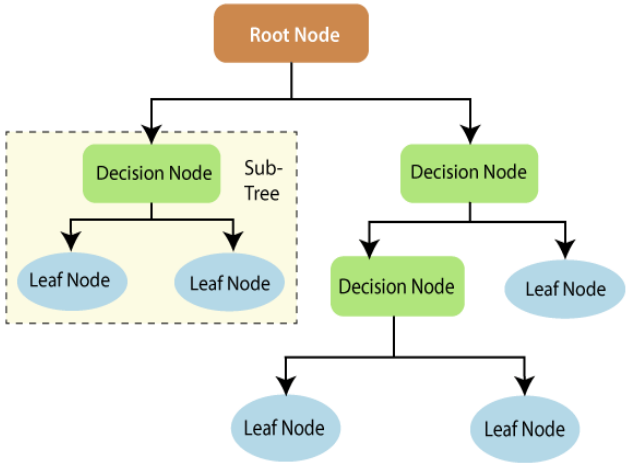

Fig.3, An example of a decision tree

Random forest

Random forest is an ensemble learning method for classification and regression. An ensemble method combines several base classifiers with the aim of creating a classification model. The ensemble declares a class prediction according to the votes of the base classifiers. It could be more accurate than a single classifier. In a random forest, each of the base classifiers is a decision tree, and the set of decision trees builds the forest.

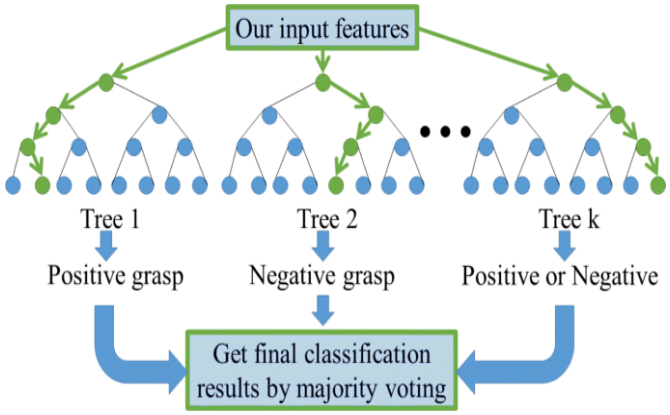

Fig.4, An example of a random forest

Support Vector Machine

A Support Vector Machine (SVM) is a supervised machine learning method for classification and regression analysis. The objective of the support vector machine algorithm for two-class problems is to find an optimal hyper plane with maximum distance to the closest samples of the two classes. A set of these closest samples to optimal hyper plane is called a support vector. This optimal hyper plane provides a linear classifier.

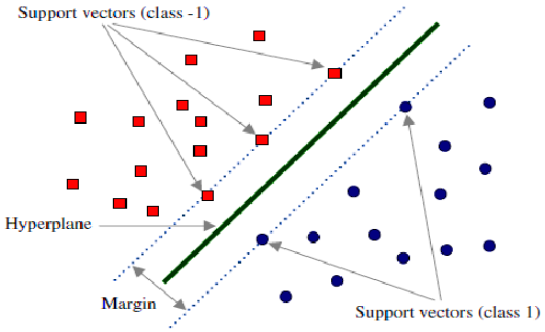

Fig.5, How SVM works

K-Nearest Neighbour

K-Nearest Neighbour (K-NN) classifier is a simple supervised data mining classifier that can be used in many applications in different areas such as health datasets, pattern recognition, cluster analysis, online marketing, image field, etc. The KNN model classifies a new data sample by searching the similar 'K' neighbours data samples in the entire training set and the data sample will be assigned to the class of highest samples [18].

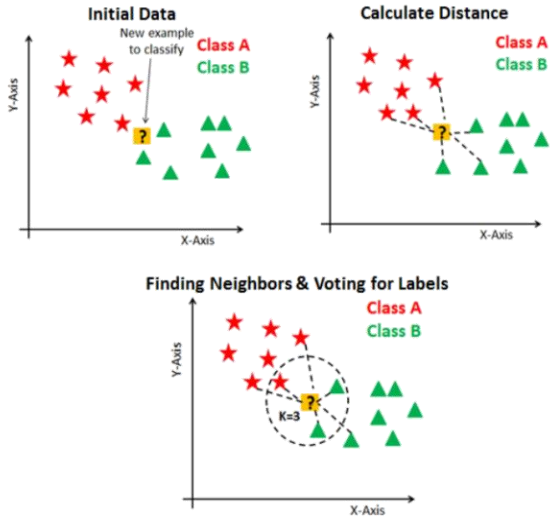

Fig.6, How KNN works

Naïve bayes

Naïve bayes is a simple, popular classifier and powerful machine learning technique. Naive Bayes is probabilistic classifier that handles the classification by See which class has higher probability, and then the higher probability class will be the class of the data input.

The steps of naïve bayes are shown as follow:-

Step 1:- Calculate prior probability for given class labels.
Step 2 :- Calculate conditional probability with each feature for each class.
Step 3 :- Multiply same class conditional probability.
Step 4 :- Multiply prior probability with step 3 probability .
Step 5 :- See which class has higher probability , higher probability class will be the class of the data input.

$$P(c_j|F) = \left[P(c_j) * \prod_{i=1}^{n} P(f_i|c_j) \right]$$

Where $P(c_j|F)$ is the conditional probability of class c_j given the feature vector F (also called posterior probability) , $P(F|c_j)$ is the conditional probability of F given the class c_j (also called likelihood), and $P(c_j)$ is the prior probability of class c_j

6. Improving model performance

In machine learning and data science, high-dimensional data processing is a challenging task for both researchers and application developers. Dimensionality reduction which is an unsupervised learning technique is important because it leads to better human interpretations, lower computational costs, and avoid redundancy by simplifying models. The process of feature selection and feature extraction can be used for dimensionality reduction. The primary distinction between the selection and extraction of features is that the “feature selection” keeps a subset of the original features, while “feature extraction” creates brand new ones.

6.1.Feature selection

Feature Selection is the process of reducing the number of input variables by selecting a subset of consistent, non-redundant, and relevant features when developing a predictive model. Hence, feature selection can reduce the modeling computational cost as well as improving the model performance.

The basic premise of using feature selection is that the input data can contain several features that are either irrelevant or redundant, and thus can be removed with no loss of information. The benefits of feature selection are; increased performance, making training faster, easier to understand, easier to debug, easier to build, and remove irrelevant features. Feature selection techniques can be roughly classified into number of methods which are; filter method, wrapper method, and hybrid method.

In filter methods, there are different ranking techniques used as the principle criteria for selecting the features regardless any machine learning algorithm. Instead, Ranking methods assign ranks to each feature based on various statistical tests and an appropriate ranking method is selected and used to rank the features; after this, these features are passed to the learning model. The examples of filter methods are Chi-squared test, Analysis of variance (ANOVA) test, and Pearson's correlation coefficient.

With the help of machine learning algorithms, the wrapper methods tune the model and select the features by evaluating the accuracy of classification and error rate. The aim of this method is to decrease the error of classification, and to increase the efficiency of classification. In general, it provides greater accuracy than the filter approach.

In the hybrid methods, the best properties of filters and wrappers are combined. Hybrid methods typically achieve high accuracy. Although almost any combination of filter and wrapper can be used to create a hybrid methodology, some interesting methodologies have recently been proposed such as; hybrid genetic algorithms, hybrid ant colony optimization, etc [19].

Feature extraction

In a machine learning-based model or system, feature extraction techniques usually provide a better understanding of the data, a way to improve prediction accuracy, and to reduce computational cost or training time. The aim of feature extraction is to reduce the number of features in a dataset by generating new ones from the existing ones and then discarding the original features. The majority of the information found in the original set of features can then be summarized using this new reduced set of features [20].

Conclusion

With the increasing rate of data generation in various areas of human life, knowledge extraction from the generated data is strongly needed. In healthcare sector, the huge amount of hospital and clinical data has a great potential for discovering relations and hidden patterns. Artificial intelligence (AI), particularly, machine learning (ML) have grown rapidly in recent years in the context of data analysis and computing that typically allows the applications to function in an intelligent manner. ML usually provides systems with the ability to learn and enhance from experience automatically without being specifically programmed. Various types of machine

learning algorithms such as supervised, unsupervised, semi-supervised, and reinforcement learning exist in the area. Besides, the deep learning, which is part of a broader family of machine learning methods, can intelligently analyze the data on a large scale. In healthcare field, data mining methods reduce time and cost in detection and diagnosis of diseases. It also has a special role in the treatment plan and medical decisions and helps construct accurate and reliable models.

Reference

[1] Sandhya Joshi, P. Deepa Shenoy, Venugopal K R, and L M Patnaik, "Data Analysis and Classification of Various Stages of Dementia Adopting Rough Sets Theory," International Journal on Information processing, vol. 4, no. 1, pp. 86-99, 2010.

[2] Benko A and Wilson B, "Online decision support gives plans an edge," Managed Healthcare Executive, vol. 13, no. 5, pp. 20- 25.

[3] Han J. and Kamber M., Data Mining: Concepts and Techniques. San Francisco, USA: Morgan Kaufmann, 2001.

[4] Ian H. Witten and Eibe Frank, Data mining: Practical machine learning tools and techniques with Java implementations. CA, San Francisco, USA: Morgan Kaufmann, 2000.

[5] Hand D. J., Mannila H., and Smyth P, Principles of data mining. Boston,MA, USA: MIT Press, 2001.

[6] Ishtake S. H and Sanap S.A., "Intelligent Heart Disease Prediction System Using Data Mining Techniques," International Journal of Healthcare & Biomedical Research, vol. 1, no. 3, pp. 94-101, 2013.

[7] Bellazzi R. and Zupan B., "Predictive data mining in clinical medicine: current issues and guidelines," International Journal Of Medical Informatics, vol. 77, no. 2, pp. 81-97, 2008.

[8] Übeyli E. D., "Comparison of different classification algorithms in clinical- decision making," Expert Systems, vol. 24, no. 1, pp. 17-31, february 2007.

[9] Harleen Kaur and Siri K.Wasan, "Empirical study on applications of data mining techniques in healthcare," Journal Of Computer Science, vol. 2, no. 2, pp. 194-200, 2006.

[10] Romeo M., Burden F., Quinn M., and Wood B. and McNaughton D., "Infrared microspectroscopy and artificial neural networks in the diagnosis of cervical cancer," Cellular and Molecular Biology(Noisy-le-Grand France), vol. 44, no. 1, pp. 179- 187, 1998.

[11] Ball G. et al., "An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers," Bioinformatics, vol. 18, no. 3, pp. 395–404, 2002.

[12] Sergey Aleynikov and Evangelia Micheli-Tzanakou, "Classification of retinal damage by a neural network based system," Journal of Medical Systems, vol. 22, no. 3, pp. 129-136, 1998.

[13] Potter R., "Comparison of classification algorithms applied to breast cancer diagnosis and prognosis advances in data mining," in 7th Industrial Conference ICDM 2007, Leipzig Germany, july 2007, pp. 40–49.

[14] Igor Kononenko, Ivan Bratko, and Matjaz Kukar, "Application of machine learning to medical diagnosis. Machine Learning and Data Mining: Methods and Applications," John Wiley and Sons Ltd, pp. 389-408, 1997.

[15] Sharma A. and Roy R.J., "Design of a recognition system to predict movement during anesthesia," IEEE Trans. Biomedical Engineering , vol. 44, no. 6, pp. 505–511, 1997.

[16] Einstein A. J., Wu H. S., Sanchez M., and Gil J., "Fractal characterization of chromatin appearance for diagnosis in breast cytology," Journal Of Pathology, vol. 185, no. 4, pp. 366-381, August 1998.

[17] Brickley M., Shepherd J. P., and Armstrong R. A., "Neural networks: a new technique for development of decision support systems in dentistry," Journal Of Dentistry, vol. 26, no. 4, pp. 305-309, May 1998.

[18] López-Vallverdú JA, Riano D, and Bohada J, Improving medical decision trees by combining relevant health-care criteria.: Expert Systems with Applications, 2012, vol. 39.

[19] J. Han and M. Kamber, Data Mining: Concepts and Techniques. San Francisco: Morgan Kauffmann Publishers, 2001.

[20] Apte and S.M. Weiss, "Data Mining with Decision Trees and Decision Rules," T.J. Watson Research Center,http://www.research.ibm.com/dar/papers/pdf/fgcsaptewe issue_with_cover.pdf, 1997.