International Journal of Engineering & Technology, 6 (4) (2017) 98-102

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v6i4.8214
Research paper

Big data management with machine learning inscribed by domain knowledge for health care

E.P.Ephzibah *, R. Sujatha

School of Information Technology & Engineering VIT University, Vellore-632014
**Corresponding author E-mail: ep.ephzibah@vit.ac.in*

Abstract

In this work, a framework that helps in the disease diagnosis process with big-data management and machine learning using rule based, instance based, statistical, neural network and support vector method is given. Concerning this, big-data that contains the details of various diseases are collected, preprocessed and managed for classification. Diagnosis is a day-to-day activity for the medical practitioners and is also a decision-making task that requires domain knowledge and expertise in the specific field. This framework suggests different machine learning methods to aid the practitioner to diagnose disease based on the best classifier that is identified in the health care system. The framework has three main segments like big-data management, machine learning and input/output details of the patient. It has been already proved in the literature that the computing methods do help in disease diagnosis, provided the data about that particular disease is available in the data center. Thus this framework will provide a source of confidence and satisfaction to the doctors, as the model generated is based on the accuracy of the classifier compared to other classifiers.

Keywords: *Big Data; Classification; Disease Diagnosis; Domain Knowledge; Machine Learning.*

1. Introduction

In the fast phase world, the health care is facing challenges in spotting the disease of the patients. The complexities prevail can be sorted greatly by making use of the historic or prevailing data source that acts as the great support for correlating the happening with the current diagnosis. The abundant of data is available, and the process of pre-processing helps to get the more accurate results at the quicker time. The impact of data and lag of getting the interesting decisions is well said in the following two statements.

We are drowning in information, but starving for knowledge - John Naisbett.

Information by itself is a pretty thin meal, if not mixed with other ingredients - Internet quote

The trending technique of big data with machine learning and domain expert based approach will help in perfect diagnosis. The classifying the disease based on the attributes is the significant area of research in the diagnosis perspective. The efficient classification can be achieved with the aid of the appropriate machine learning strategy that yields the better accuracy [1].

Big data is the trending area in the information technology with large data set at rest. The data possess the attributes like volume, velocity, variety, veracity and value. The analysis over the data falls under the broad categories of descriptive, diagnostic, predictive, prescriptive and decisive. The analysis applied on the happened data to make a review is done with descriptive and diagnostic analytics. The present with the historical data providing the preview about the future was achieved with the predictive, prescriptive and decisive analytics. In the transition from the typical health to smart health, the part of the big data is inevitable. The smart city is the ultimate destination of the developing country such that increased utilization of information and communication technology. [2-6]

The machine learning is the zenith area that could be applied in all the application to derive the impactful decisions. The data preprocessing followed by learning and evaluation stage is the complete one cycle of machine learning. The preprocessing is the critical phase where the work on the raw data commences. The learning feedback based nature paves to the supervised, unsupervised and reinforcement techniques. The machine learning another dimension based on availability is categorized as batch and online. The evaluation phase helps in knowing the strength of the utilized algorithms and selected datasets. It is measured with the certain metrics and measures. A lot of medical oriented research is being made in combination with machine learning strategies. The pure data and image are being taken as the background for analytical purpose, detection and in turn for diagnosis purpose [7], [8].

The Domain based approach with the machine learning capability helps in the better decision making. The stated ultimate objective of the domain driven data mining is to build the next-generation approaches, techniques and tools for a conceivable paradigm change from data-centers hidden pattern mining to domain-driven actionable knowledge discovery for both technical and business viewpoint. The area requires involvement and integration of intelligence from human, domain, data, network, organizational and social, and the meta-synthesis of the ubiquitous. The system with machine learning and domain based perspective will deliver business-friendly and decision-making rules and actions that have higher significance [9].

2. Literature survey

A wide variety of literatures is available, especially on big data and medical management. This work focuses both the sectors equally. This survey will give a prospective outlook on how big data has been effectively used for medical field, especially in dis-

ease diagnosis. In the paper by Pramanik et.al, the authors have defined a framework for smart health care that combines big data and health care system. Their manuscript provides the major components of smart city and smart health as smart home, office, health care, transport education and security with respect to its applications and hospital-based health care, digital classical health care, pervasive health care with respect to health care systems. The frame work combines the smart city, healthcare and big data. Their focus is to provide smart healthcare service with the help of research and knowledge discovery, smart service based infra structure, architecture and big data analytics with big data platform and tools [2].

In the paper by Sohail et al., the authors have designed a model that access big data for healthcare using heterogeneous devices. Based on the details collected from the sensors of IOT devices the model identifies the drug with the side effects and issues it to the patients. It provides a link between the patients and the doctors through devices. The authors have taken into account the big data to be available in the cloud and the services are also provided from the cloud. Semantic interoperability is combined with the cloud services which in turn are tied with the big data analytics. Their proposed model contains three components like sensor network, cloud and big-data semantic interoperability [10].

The researcher has developed a machine learning health care prediction model that takes the input form the tweets and sends it to the machine learning component of the system. The data available in the cloud is processed and analyzed. The authors have taken the input from the tweeter communication channel. The big data processing engine spark is capable of receiving the streaming data across various resources. Here it has been used to map the tweet to a data for implementation of a decision tree. The OAuth authentication has been used for security reasons, so that the data being sent is safe and secured. The data is then analyzed using the machine learning algorithm for heart disease prediction [11].

Machine learning is an effective computing technique that helps in prediction and classification. There are several types of machine learning techniques available for health care systems. The authors have worked on many techniques that are suitable for specific diseases. It is true that not all the ML algorithms produce good classification accuracy for all the diseases. Therefore, the authors have identified a few techniques appropriate for each disease. The algorithms include artificial intelligence; rule based algorithms, statistical methods and instance based techniques [12].

They have designed computer-aided diagnosis for Alzheimer's disease using some computing techniques like lattice computing and KNN classifier. The voxel based morphometry process has been used to extract important features that would help in efficient and faster classification. The correlation values among the features are also taken into consideration for feature selection. The important features alone are used in the next step of classification using lattice theory. The proposed method has produced acceptable results when compared with other similar works in the literature [13].

Huang has proposed a scheme for severity prediction of dementia disease using computing methods. This approach has been performed on Arterial Spin Labelling (ASL) image data. The proposed scheme has two stages namely testing and training. During the training phase, the ASL data is analyzed, and the important features are extracted from it thereby reducing the irrelevant and redundant features. The ranking function learns the training data and ranks the collection of items. The testing data based on the extracted features produce better results [14].

The authors have developed a cloud based framework that helps a common man to understand his disease and get immediate diagnosis and suggestions as a treatment for it. The prototype system makes use of the clusters for efficient and faster data access from the cloud database. The Hadoop software has been used for storing distributed data and retrieving reduced one. An online distributed search cluster performs tasks like load balancing, data analysis, user query and access control. The data that comes from various sources are in different formats. Security of data and patient information also is taken care with a node selection algorithm [15].

Costa has projected the various institutions and companies that provide health care solutions using computing techniques. The author says that the combination of advanced information technology and biomedicine field has brought an evolution in healthcare solutions. The patients can easily find out solutions to their health issues thereby preventing them from long waiting time, unnecessary tests, expensive treatment charges and many more discomforts they face. The researchers from numerous institutions and industries have given many frameworks that could help anybody to use and get benefitted. The author has given the pictorial view on "Big data in Biomedicine" comprising the components like data generation, hardware and software. The data generation segment takes care of the DNA sequencing and data transfer. The hardware segment focuses on the data storage and security. The software segment encapsulates the tasks like data analytics, visualization and translation. This segment provides the link between the "omnics" and clinical data [16].

The authors have given a decision support system for disease diagnosis from the medical datasets that will assist the practitioners to accurately perform the healthcare practise. The system has been developed with the combination of Classification And Regression Tree algorithm (CART) and fuzzy logic. The CART algorithm helps in generating and identifying the fuzzy rules for accurate prediction. Mean absolute error estimator and Coefficient of determination R^2 were the metrics used for evaluating the performance of the system and are found to produce better results [17].

3. Proposed framework

The process of the framework starts with the big data source collected from various destinations, and the data transformation takes place to ensure the suitability of data to work for analytics purpose. The big data analytics applications are querying, reports, data mining and so on. The machine learning part of data mining comprises of a versatile algorithm that applied over the data set provides the way of processing further. The big data comprises of data from web and social media, biometric data, machine-to-machine data, big transaction data and human generated data. Since the data is digitized and network is excellent over the data set the health care system is being benefited greater by earlier disease detection and quality of care is optimized [18].

In the health care system, the electronic health record is the source of patient's treatment history that provides information to authorized users. Digital version helps practitioners to know about the patient's health condition over the period in single click Follow-up activities are visualized very effectively. It is a widespread report of the patient's whole health. The health care is the huge, complex system that required to be handled with great care and concern. The stakeholders are patient, clinicians, lab technicians, public health practitioners, pharmacist, and care takers. The heterogeneous stored data helps the practitioners in the great manner for the efficient diagnosis [19], [20].

The life of an individual patient is closely monitored in this integrated system. It Furthermore, helps in the government during the crisis that may arise due to epidemic diseases. The life of the individual needs to be cared with great care and interpretation of the domain expert will make the process still more confident and ethical in nature. The experience of the practitioners when clubbed with the machine learning and huge data at the background yields the best output in the diagnosis process [21]. Figure 1 (Fig.1) gives the proposed framework.

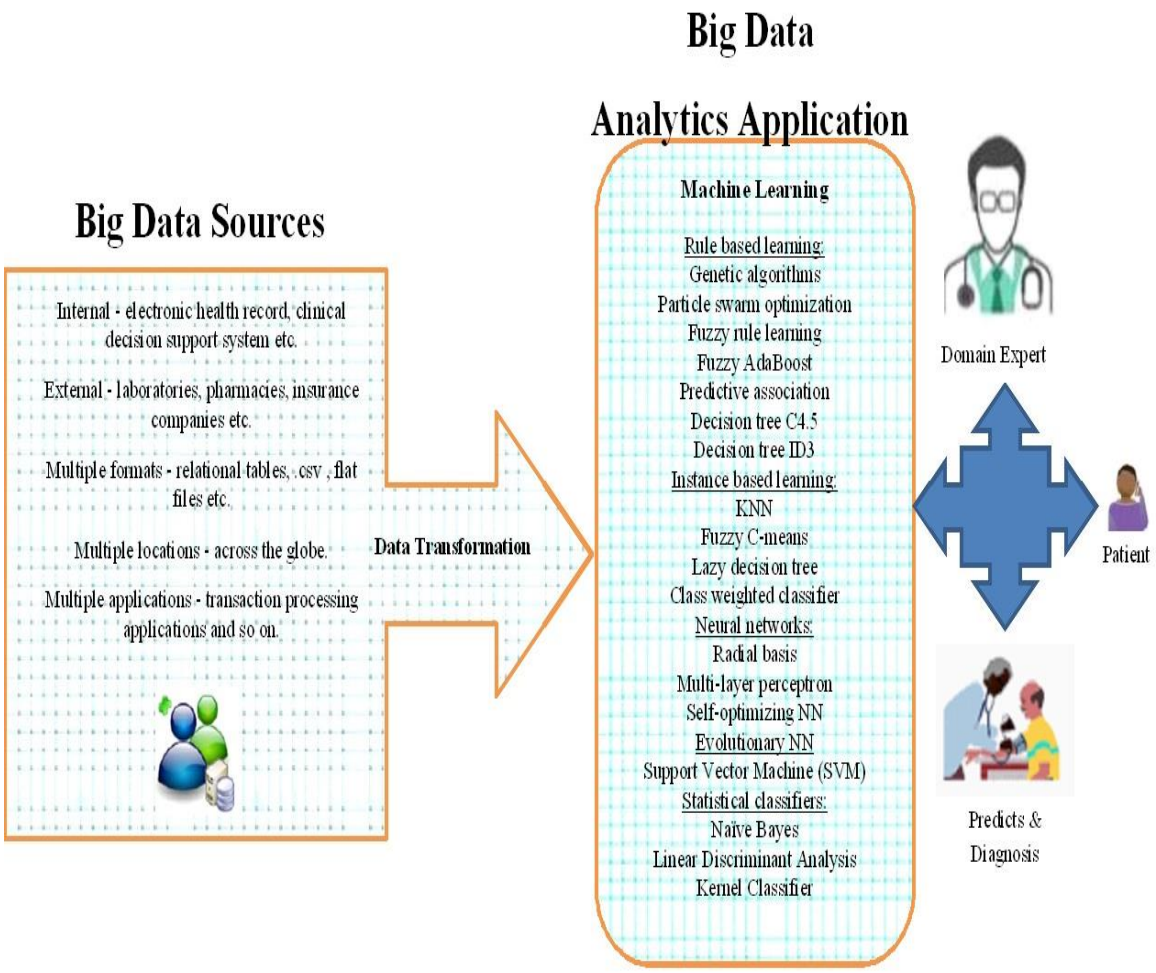

Fig. 1: A Framework of Domain Knowledge Enabled Big Data Analytics in Health Care System.

The repository data and new diagnosis data from a patient for classifying are utilized in the calculative manner. With the help of algorithm's appropriate prediction and diagnoses are achieved. The attributes for diagnosis or fed in the various learning algorithms to make over the classification. The accuracy parameter helps in finding the suitable machine learning for the disease prediction. The data used for the diagnosis differ from disease to disease.

In the field of medicine, information technology plays a vital role. In the literature, it is found that lot of researchers has developed new models, frameworks, support systems, especially for healthcare and disease diagnosis. This paper has been designed with a focus on health care to the patients. It may not be directly available to the patients, as it may lead to unnecessary confusions, misguidance and psychological problems. Hence the proposed architecture gives suggestions to the domain experts- the medical practitioners. The final decision is to take by the expert with his/her or her knowledge in the medical field. The architectures comprises of three components namely,

1) Big data sources
2) Big data Analytics Application
3) Domain Expert knowledge with patient interaction.

3.1. Big data sources

About big data sources, we mean the collection of large volumes of data from different formats. The internal data collected from electronic health records, clinical decision support system data, relational tables, XML data, flat files and so on. The data is collected from various resources, transformed to a particular format and stored in the repository. The stored data may contain redundant entries, missing values and noisy. Hence the data is preprocessed to clean and keep it ready for further classification and prediction.

3.2. Big data analytics application

In this phase the machine learning techniques come into play for developing the models. The principle behind this segment is to attract the data related to the input being sent by the doctor and to develop a model which is more appropriate to the input in terms of its higher classification accuracy and minimum prediction error. The different models that are generated are categorized based on their mechanisms like, rule based learning, instance based learning, neural networks, evolutionary neural network and statistical classifiers.

Rule based learning comprises of the genetic algorithms, particle swarm optimization, fuzzy rule learning, fuzzy AdaBoost, predictive association, decision tree (C4.5&ID3). This type of learning methodology involves association rule mining. This is the machine learning process that identifies the relationship between the individual attributes available in the data [22]. The support and confidence are the measures that reveal the correlation between the attributes according to the well-known Apriori algorithm [23]. The rule-based learning helps to find the patterns hidden in the existing data, which will be useful to generate the rules. It has been proved that evolutionary rule based learning produces better results in disease diagnosis [24].

Instance based learning comprises of the nearest neighbor classifier, fuzzy c-means classifier, lazy decision tree; class weighted classifiers. This type of classifiers helps in disease diagnosis based on the instances available in the data center. The distance measures are used to match the incoming new instance with the existing one. One such distance function is the Euclidean distance. Each training instance is associated with a weight that determines its characteristics with the class labels. Especially in disease diagnosis the class labels can be presence or absence of the disease or the severity in different ranks or scores [25]. Class labels represent

the type of output or the resultant class. This type of learning is widely available in bio-medical fields for classification and prediction [26].

Neural networks based learning involves radial basis function networks, multi-layer perceptron networks and self-optimizing neural networks. These networks are found to better in learning and training when compared to other types. Neural networks contain basic elements like neurons or cells that carry meaningful information in terms of weights. The different layers in the network help in propagating the data from input to output through hidden layers [27]. The network is trained, and the weights are adjusted so as to minimise the error and improve the prediction accuracy.

Support Vector Machine (SVM) classifiers are included in the framework to diagnose the disease as they are found to give promising results, especially in biomedicine [28]. This type of classifier emphasizes on supervised learning method for classification and regression that depends on the support vectors that decide the boundary between the classifiers thereby making the prediction task much easier and faster. It is flexible and also can handle large solution space. SVM can be combined with evolutionary methods and algorithms for classification and prediction [29].

Statistical method for classification is an early method that attracted many researchers and has occurred in various research works and produced better results with good accuracy. In the proposed framework we have used Naïve Bayes approach, Linear Discriminant Analysis and Kernel classifiers.

3.3. Domain expert

Machine learning is complete and successful only with the help of domain expert. In the medical field, the practitioners play a vital role in disease diagnosis. As accuracy plays a vital role in disease prediction and further treatment, the domain expert understands and analyses the solution obtained from the computing methods. Whether the diagnosis is using data in numerical or in image form, the expert has to select the appropriate prediction method based on his experience and domain knowledge. The reason for including the domain expert in the frame work is to make the task user-friendly and more professional.

4. Conclusion

The digital world relies on the technology in all the walk of the life. The vital thing in human's life is the good health, and proper care is needed during the crucial illness time. The advent of the information and communication technology made this process very transparent and effective. The fuzzy algorithms used in the machine learning of our framework provide a way to the best output even for the finer or granular inputs. In addition, the expert is appended to ensure the correctness. The enhancement can be done by instilling in the cloud platform along with the concept of deep learning. The smart city concepts' main idea is to make the effective case management in the door step by providing preliminary treatment and with the Internet of things that could be achieved. The extension can be done in various segments that will provide the quality treatment for the diagnosed diseases.

References

[1] K. Polat and S. Güne, 2007. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digital Signal Processing, 17, 702–710. https://doi.org/10.1016/j.dsp.2006.09.005.

[2] Pramanik, M.I., Lau, R.Y., Demirkan, H. and Azad, M.A.K., 2017. Smart Health: Big Data Enabled Health Paradigm within Smart Cities. Expert Systems with Applications. https://doi.org/10.1016/j.eswa.2017.06.027.

[3] S. Sakr and A. Elgammal, 2016. Towards a Comprehensive Data Analytics Framework for Smart Healthcare Services. Big Data Research, 4, 44–58. https://doi.org/10.1016/j.bdr.2016.05.002.

[4] L. Wang and C. A. Alexander, 2015. Big Data in Medical Applications and Health Care. American Medical Journal, 6 (1),1- 8 https://doi.org/10.3844/amjsp.2015.1.8.

[5] W. Raghupathi and V. Raghupathi, 2014. Big data analytics in healthcare : promise and potential. Health Information and Science Systems, 2(1), 1–10. https://doi.org/10.1186/2047-2501-2-3.

[6] D. Haluza and D. Jungwirth, 2015. ICT and the future of health care : aspects of health promotion. International Journal of Medical Informatics, 84(1), 48–57. https://doi.org/10.1016/j.ijmedinf.2014.09.005.

[7] M. De Bruijne, 2016. Machine learning approaches in medical image analysis : From detection to diagnosis. Medical Image Analysis, 33, 94–97. https://doi.org/10.1016/j.media.2016.06.032.

[8] Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, 2017. Machine Learning and Data Mining Methods in Diabetes Research. Computational and Structural Biotechnology Journal, 15, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005.

[9] K. Greyson et al., 2013. Formulation Process of Knowledge for an Expert Healthcare System Unit. AASRI Procedia, 4, 190–195. https://doi.org/10.1016/j.aasri.2013.10.029.

[10] Sohail Jabbar, Farhan Ullah, Shehzad Khalid, Murad Khan, and Kijun Han, 2017.Semantic Interoperability in Heterogeneous IoT Infrastructure for Healthcare. Wireless Communications and Mobile Computing, 2017, https://doi.org/10.1155/2017/9731806.

[11] Nair, L.R., Shetty, S.D. and Shetty, S.D., 2017. Applying spark based machine learning model on streaming big data for health status prediction. Computers & Electrical Engineering. In Press. https://doi.org/10.1016/j.compeleceng.2017.03.009.

[12] Fatima, M. and Pasha, M., 2017, Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications, 9(01), 1-16. https://doi.org/10.4236/jilsa.2017.91001.

[13] Papakostas, G.A., Savio, A., Graña, M. and Kaburlasos, V.G., 2015, A lattice computing approach to Alzheimer's disease computer assisted diagnosis based on MRI data. Neurocomputing, 150, 37-42. https://doi.org/10.1016/j.neucom.2014.02.076.

[14] Huang, W., 2016. A novel disease severity prediction scheme via big pair-wise ranking and learning techniques using image-based personal clinical data. Signal Processing, 124, 233-245. https://doi.org/10.1016/j.sigpro.2015.08.004.

[15] Lin, W., Dou, W., Zhou, Z. and Liu, C., 2015. A cloud-based framework for Home-diagnosis service over big medical data. Journal of Systems and Software, 102, 192-206. https://doi.org/10.1016/j.jss.2014.05.068.

[16] Costa, F.F., 2014. Big data in biomedicine. Drug discovery today, 19(4), 433-440. https://doi.org/10.1016/j.drudis.2013.10.012.

[17] Nilashi, M., bin Ibrahim, O., Ahmadi, H. and Shahmoradi, L., 2017.An Analytical Method for Diseases Prediction Using Machine Learning Techniques. Computers & Chemical Engineering, 106, 212-223. https://doi.org/10.1016/j.compchemeng.2017.06.011.

[18] Cottle, M., Hoover, W., Kanwal, S., Kohn, M., Strome, T. and Treister, N., 2013. Transforming Health Care through Big Data Strategies for leveraging big data in the health care industry. Institute for Health Technology Transformation, http://ihealthtran.com/big-data-in-healthcare.

[19] J. Yang et al., 2015. Computers in Industry Emerging information technologies for enhanced healthcare. Computers in Industry, 69, 3–11. https://doi.org/10.1016/j.compind.2015.01.012.

[20] L. Ericson, T. Hammar, N. Schönström, and G. Petersson, 2017. Stakeholder consensus on the purpose of clinical evaluation of electronic health records is required. Health Policy and Technology, 6(2), 152–160. https://doi.org/10.1016/j.hlpt.2017.02.005.

[21] C. Muriana, T. Piazza, and G. Vizzini, 2016. An expert system for financial performance assessment of health care structures based on fuzzy sets and KPIs. Knowledge-Based Systems, 97, 1–10. https://doi.org/10.1016/j.knosys.2016.01.026.

[22] Chaves, R., Ramírez, J., Gorriz, J.M. and Alzheimer's Disease Neuroimaging Initiative, 2013. Integrating discretization and association rule-based classification for Alzheimer's disease diagnosis. Expert Systems with Applications, 40(5), 1571-1578. https://doi.org/10.1016/j.eswa.2012.09.003.

[23] He, R., Xiong, N., Yang, L.T. and Park, J.H., 2011. Using multimodal semantic association rules to fuse keywords and visual features automatically for web image retrieval. Information Fusion, 12 (3), 223-230. https://doi.org/10.1016/j.inffus.2010.02.001.

[24] Cheruku, R., Edla, D.R., Kuppili, V. and Dharavath, R., 2017. RST-BatMiner: A Fuzzy Rule Miner Integrating Rough Set Feature Selection and Bat Optimization for Detection of Diabetes Disease.

Applied Soft Computing, In Press. https://doi.org/10.1016/j.asoc.2017.06.032.

[25] Tong, T., Wolz, R., Gao, Q., Guerrero, R., Hajnal, J.V., Rueckert, D. and Alzheimer's Disease Neuroimaging Initiative, 2014. Multiple instance learning for classification of dementia in brain MRI. Medical image analysis, 18(5), 808-818. https://doi.org/10.1016/j.media.2014.04.006.

[26] Gagliardi, F., 2011.Instance-based classifiers applied to medical databases: diagnosis and knowledge extraction. Artificial intelligence in medicine, 52(3), 123-139. https://doi.org/10.1016/j.artmed.2011.04.002.

[27] Arabasadi, Z., Alizadehsani, R., Roshanzamir, M., Moosaei, H. and Yarifard, A.A., 2017. Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm. Computer Methods and Programs in Biomedicine, 141, pp.19-26. https://doi.org/10.1016/j.cmpb.2017.01.004.

[28] Dolatabadi, A.D., Khadem, S.E.Z. and Asl, B.M., 2017. Automated diagnosis of coronary artery disease (CAD) patients using optimized SVM. Computer methods and programs in biomedicine, 138, pp.117-126. https://doi.org/10.1016/j.cmpb.2016.10.011.

[29] Sartakhti, J.S., Zangooei, M.H. and Mozafari, K., 2012. Hepatitis disease diagnosis using a novel hybrid method based on support vector machine and simulated annealing (SVM-SA). Computer methods and programs in biomedicine, 108(2), pp.570-579. https://doi.org/10.1016/j.cmpb.2011.08.003.