Data Exploration of chronic Disease through Hybrid Web Mining Algorithms

Manju Saini[1], Dharmender Kumar[2]
[1,2]Department of Computer Science and Application
Mewar University, Chittogarh Rajasthan[1], GJUS&T, Hisar, Haryana[2]

Abstract - This article discusses on the highly contributory factor of diagnosis and prediction. The survey is presented for three different types of data for different diseases, datasets and percentage errors. It requires the selection of diseases which have recently affected humans and their main characteristics responsible for growth and development. Kidney diseases were further explored for preliminary investigations to understand the nature of the specific effects. Here, we are discussing the important factors lead to maximization of the disease. The biggest effect may be that it is difficult to look for a background with large factors, so the basic needs are the basis for rating the most likely features of the hood for a particular disease. The synergy techniques performed in excellence when incorporating those mathematical terms to their relationship. The survey data included largepatient's information to understand how each person interacted with others. This analysis will provide a set of each optional attribute. In addition, the results of each attribute need to be considered in the ultimate automated learning algorithms such as A-Priori, swarm optimization, etc. The research survey began some of the main tasks to achieve the goal of preparing the automated control unit.

Keywords - A-Priori, Web computing, Cluster Particle Suggestions (C-PSO), Prediction of disease, Data Mining

I. INTRODUCTION

Expert systems are increasingly used in medical diagnosis. There is no doubt that assessing patient data and making decisions from experts is the most important factor in diagnosing the disease. However, experts and various artificial intelligence techniques can also help in classifying to related researcher [1] [2] [3]. To this end, many methods of network computing have been proposed. Rating systems help to find with errors which may occur due to fatigue or reduce expert experiences provide medical data that is cable to examine in more detail in less time. Critical Diabetes or simple diabetes is a metabolic disease where a group of people faces with having into advanced blood sugar. The maximum sugar in the blood leads to the urine increases the thirst and increases appetite. Disordered diabetes can cause many complications. Serious complications may result long-term disease, kidney failure and eye loss. Diabetes is caused by the fact that pancreas does not produce enough insulin, or so the kidney cells do not respond accurately to the insulin product. When the coronary artery is congested, blood flow to the body muscle is diminished. Heart recordings have been analyzed to reveal abnormalities in arrhythmia caused by cardiovascular disease. The medical industry gathers large amounts of medical records that must be used to discover clues to make effective judgments.Often there are patterns and hidden relationships. Doctors and patients need reliable information about the risks of various diseases in the individual. Perfectly, they will have very accurate data and can use the ideal risk model.These models can classify people with the diseases and disorders. In fact, the ideal model can predict when the disease will occur.

A. Disease Diagnosis Using Web-computing - The common people need the low-cost solution to their disease related problems due to their low condition of economic and social development. The experienced doctors provide a good judgment on the behalf of the diagnosis report. Since early systems which construction for the automatic prediction through web computing technique has been used for the prediction of disease. So, Web mining algorithms help to predict the disease with their network algorithms which can embedded on the graphical user interface.

B. Importance of web computing - In the past decade, they have seen explosive growth of information available on the World Wide Web (WWW). Today, browsers have easy access to various sources of text and multimedia data. Search engines index pages over 1,000,000,000 pages and finding the information you need is not easy. This wealth of resources encourages the need to create automated mining techniques on the web to create the term "network mining". In order to achieve network intelligence and eliminate the need for human intervention, artificial intelligence needs to be embedded with their network tools. The need to build server and client intelligent systems has attracted the attention of researchers in information, knowledge discovery, machine learning and artificial intelligence, and these systems have an effective impact on the Internet and web too etc. Conversely, the problem of developing automated tools for searching, extracting, filtering and evaluating information need to be more development or it is not in the mature stage. Softcomputing seems to be a novel choice for solving these properties and overcoming some existing method limits. The field of research that combines these two areas can be called "soft mesh mining".

C. A-priori - An important idea of the A-priori algorithm is to pass the database multiple times. It uses an iterative method called breadth first search (horizontal search) through the search space. Here, a set of k items (k + 1)-is used to explore the project. First, find a collection of frequent one-item sets. The set of items containing the support threshold is represented. Each subsequent process begins with a large set of project seeds in the previous process. This set of seeds is used to generate a new set of possible large items, called candidate sets, and to calculate the actual support for these candidate sets as they deliver data. At the end of the deliver, we determine which candidate sets are large (frequently) and they become seeds for the next step.

Therefore, it is used to find until no more frequent set of k items is found and is used to find a set of two frequent items.

D. Clustered Particle Swarm Optimization(C-PSO) - C-PSO has been successfully implemented in the field of database extraction so that the classification system can be eliminated based on governance. C-PSO is very useful for a specific set of rules to diagnose different types of diseases. Develop a debugging algorithm to extract a base for diagnosis of mental illness. The C- PSO is very fine technique that is the integration of clustering techniques with the optimization techniques.

```
Initialize: Cluster of N-data set of disease
Step 1: Define the cluster range
Step 2: Provide the range R: (ini= 1; i<N, i++)
Step 3: The data sets of clusters find the accuracy
Step 4: Find the best accuracy through PSO
Step 5: Fetch the threshold value with local minima and maxima with global maxima and minima.
```

E. Web Mining for Healthcare Administration - Healthcare systems is a kind of organizations where it can provide healthcare services to people, institutions, and resources that enabled to meet the health requirements of the targeted population.The development of Internet technology supports to latest advances in medicine, engineering research, communications and information technology development. Online technology gives us effective and improved medical information about patients and their health. In the field of healthcare, a direct interview between the patient and the doctor, the doctor and the doctor is essential. In the absence of these meetings, thedesigned model plays a very essential role in getting better treatment and care. It also covers all forms of communication between users. Patients and healthcare workers will able to deliver these electronic devices from remote places and areas.

II. RESEARCH BACKGROUND

[10] Conversing the slow progress of chronic kidney disease, early detection and effective cure are the only ways to reduce mortality. Automated learning technology is one of the great importance in medical diagnosis due to its high rating ability. The goal of the rating algorithm relies upon the utilization of the right element choice calculation to lessen the dimensions of the informational collection. In this examination, the vector support appraisal calculation was utilized to analyze constant kidney sickness. So, to analyze interminable kidney infection, two essential sorts were chosen, packing and filtration methods to diminish the span of constant kidney sickness gatherings. In the assembly method, a subset of the workbook is used with a greedy step-by-step search engine and an aggregate group evaluator with a higher-level search engine. In the filtering method, the subset evaluator is used to identify related properties through by greedy gradually search engine and filter subset with the best search engine.Consequently, itshows that the diagnosis of chronic kidney disease through sub-group filter evaluation is the highest search engine or in other ways to select the accuracy of vector meter (98.5%) and the best advantage. Specific. [11] It has been reported that diabetes has become a global epidemic of chronic diseases. The motivation behind this investigation was to group diabetes by creating robotized frameworks utilizing mechanized learning methods. Our methodology was created through assembly, noise cancellation and classification methods. Therefore, we make the desired maximization, analysis of key components, and support for bus machine assembly functions, noise cancellation and classification. We also develop incremental contexts by applying incremental component analysis and vector support tools to incremental data learning. Pima Indian Diabetes data shows that this method improves prediction accuracy and greatly reduces computation time compared to non-incremental methods. Hybrid intelligence systems can help as a decision support system to healthcare professionals for any healthcare [12]. Cardiovascular illness is the important reason of tediousness and death in current lifestyles. Evidence of the distinction between cardiovascular diseases is an urgent and complex task that needs to be implemented in an accurate and skilled manner, and appropriate robots will be particularly attractive. So, as an expert, not everyone has the same skills. It is impossible for all experts to have the same talent in every sub-family. In many places, we do not have any talent or power without any effort. The mechanical framework in therapeutic analysis will enhance medical considerations and reduce costs. The key involvement of this research is to support non-professional doctors make the right decisions about heart disease risk. The proposed rules created by the system take precedence over the original rules, pruning rules and rules, no redundancy, classification rules, classification rules and Polish. The implementation of the assessment framework is closely related to the arrangement and the results indicate that the framework has the extraordinary potential to more accurately predict the risk of coronary heart disease.

III. OBJECTIVE AND CO-RELATED IMPACTED VALUES

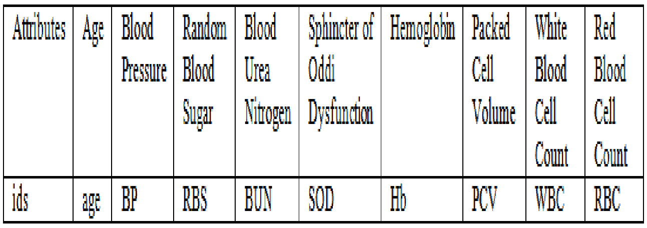
This paper mainly focuses on the four stages as follows.

- Clustering/Defragmentation of Data Sets
- Providing the threshold value
- Provide the local maxima and minima
- Defining the Global threshold local maxima andminima.
- Implanting the value to the C-PSO to get it optimization of the sets of data.

IV. MATERIALS AND METHODS

The data has been collected for analysis from https://www.kaggle.com/ []. Nine attributes have been selected for analysis from about 400 patients. The database is open access and non-proprietary format. The data sets are available in most of supporting files like as CSVs, JSON, SQLite, Archives, Big Query etc. Further, the correlation coefficients amongst selected all attribute has been calculated using the data analytic pack tools in excel work sheet of MS-Office.

Table 1: The ids of the eight attribute of patient which are in the data sets



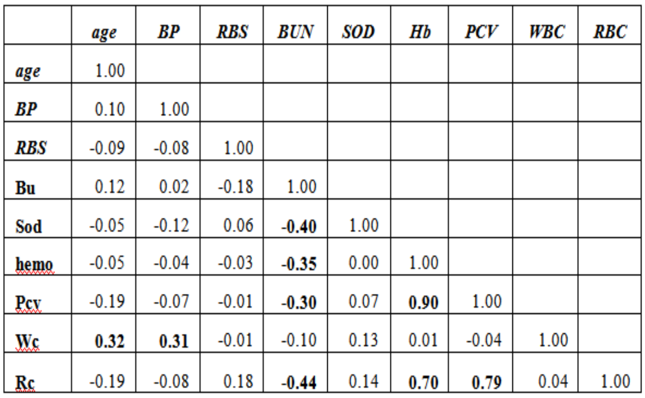
Attributes	Age	Blood Pressure	Random Blood Sugar	Blood Urea Nitrogen	Sphincter of Oddi Dysfunction	Hemoglobin	Packed Cell Volume	White Blood Cell Count	Red Blood Cell Count
ids	age	BP	RBS	BUN	SOD	Hb	PCV	WBC	RBC

V. ANALYSIS AND INTERPRETATION

The correlation coefficient has been calculated for eight attribute of the patients and It can be seen in Table 2. The table is in lower triangular form and each entries of the table is showing a correlation coefficient between two corresponding attributes according to their corresponding row and column. It had been observed that PCV and Hb, RBC and Hb, PCV and RBC are highly positive correlated, whereas BUN is negatively correlated with SOD, Hb, PCV and RBC. Thus, the few patients attribute information will be simultaneously change like as PCV and Hb, RBC and Hb, PCV and RBC. Others like SOD, Hb, PCV and RBC value of attribute decrease as increase the value of attribute BUN or vise-versa. Hence, conclusively the only knowledge of few attribute can be getting the enough information about the disease.

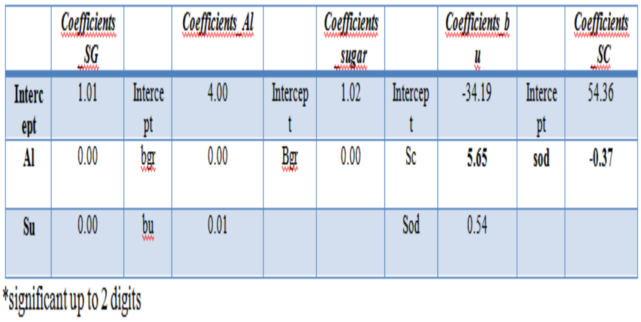
Further, the regression analysis has been executed for categorical variables which are represented in Table 3.The outcome finds from the table for Al, Bu based on regression analysis basically depend on the respective attributes.

Table 2: The correlation coefficient amongst selected all nine attributes of patient



	age	*BP*	*RBS*	*BUN*	*SOD*	*Hb*	*PCV*	*WBC*	*RBC*
age	1.00								
BP	0.10	1.00							
RBS	-0.09	-0.08	1.00						
Bu	0.12	0.02	-0.18	1.00					
Sod	-0.05	-0.12	0.06	**-0.40**	1.00				
hemo	-0.05	-0.04	-0.03	**-0.35**	0.00	1.00			
Pcv	-0.19	-0.07	-0.01	**-0.30**	0.07	**0.90**	1.00		
Wc	**0.32**	**0.31**	-0.01	-0.10	0.13	0.01	-0.04	1.00	
Rc	-0.19	-0.08	0.18	**-0.44**	0.14	**0.70**	**0.79**	0.04	1.00

Table 3: Regression Analysis and their coefficients



	Coefficients SG		*Coefficients Al*		*Coefficients sugar*		*Coefficients bu*		*Coefficients SC*
Intercept	1.01	Intercept	4.00	Intercept	1.02	Intercept	-34.19	Intercept	54.36
Al	0.00	bgr	0.00	Bgr	0.00	Sc	**5.65**	**sod**	**-0.37**
Su	0.00	bu	0.01			Sod	0.54		

*significant up to 2 digits

VI. CONCLUSIONS

With the help of technological advancements, network computing will be extended the limit of accessibility of proper medicinal services, which will make it the most practical and perfect complement to practitioners' accreditation. Diagnostic systems have been made to make it easier for us to use for more sensitive, powerful and non-professional knowledge. With the revision of literature, we found that the proposed diagnostic system has some problems so far, which was mentioned in the flaws. There is an automation of unusual symptoms present in the current diagnostics system, which is not able to indicate liver disease. The data training is not a lack of data tools in the preprocessing of data or symptoms, which is very difficult for a non-specialist doctor. Repression Analysis provides contact matrix in a form of structure. This exploration of the data sets of disease using A-priory and C-PSO with data analysis provide the good and accurate prediction of disease. The above exploration toward finding the great impact attributes among the other contributory factors.

VII. REFERENCES

[1]. Vijayarani, S., &Dhayanand, S. (2015). Data mining classification algorithms for kidney disease prediction. *International Journal on Cybernetics & Informatics (IJCI)*, *4*(4), 13-25

[2]. Nair, V., Komorowsky, C. V., Weil, E. J., Yee, B.,Hodgin, J., Harder, J. L., & Lemley, K. V. (2018). A molecular morphometric approach to diabetic kidney disease can link structure to function and outcome. *Kidney international*, *93*(2), 439-449.

[3]. Sweety, M.E., & Jiji, G.W. (2014). Detection of Alzheimer disease in brain images using PSO and Decision Tree Approach. *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*, 1305-1309.

[4]. Taie, Shereen & Ghonaim, Wafaa. (2017). Title CSO-based algorithm with support vector machine for brain tumor's disease diagnosis. 183-187. 10.1109/PERCOMW.2017.7917554.

[5]. Dey, Rajeeb& Bajpai, Vaibhav & Gandhi, Gagan& Dey, Barnali. (2009). Application of Artificial Neural Network (ANN) technique for Diagnosing Diabetes Mellitus. 1 - 4. 10.1109/ICIINFS.2008.4798367.

[6]. Uma Rani, K & Holi, Mallikarjun. (2013). Automatic detection of neurological disordered voices using mel cepstral coefficients and neural networks. 76-79. 10.1109/PHT.2013.6461288.

[7]. S. N. Kumar, D. Dinesh, T. Siddharth, S. Ramkumar, S. Nikhill and R. Lavanya, "Selection of features using Particle Swarm Optimization for microaneurysm detection in fundus images," *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, Chennai, 2017, pp. 140-144. doi: 10.1109/WiSPNET.2017.8299735.

[8]. Padol, P.B., & Sawant, S.D. (2016). Fusion classification technique used to detect downy and Powdery Mildew grape leaf diseases. *2016 International Conference on Global Trends in Signal Processing, Information Computing and Communication (ICGTSPICC)*, 298-301.

[9]. Undre, Poonam & Kaur, Harjeet& Patil, Prakash. (2015). Improvement in prediction rate and accuracy of diabetic diagnosis system using fuzzy logic hybrid combination. 1-4. 10.1109/Pervasive.2015.7087029.

[10]. P. Ravivarma, B. Ramasubramanian, G. Arunmani and B. Babumohan, "An efficient system for the detection of exudates in colour fundus images using image processing technique," *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*, Ramanathapuram, 2014, PP. 1551-1553.doi: 10.1109/ICACCCT.2014.7019366.

[11]. Sweety, M.E., & Jiji, G.W. (2014). Detection of Alzheimer disease in brain images using PSO and Decision Tree Approach. *2014 IEEE International Conference on Advanced Communications, Control and Computing Technologies*, 1305-1309.

[12]. Uma Rani, K & Holi, Mallikarjun. (2013). Automatic detection of neurological disordered voices using mel cepstral coefficients and neural networks. 76-79. 10.1109/PHT.2013.6461288.

[13]. TilvaVidita, Patel Jignesh, & Bhatt Chetan (2013). Weather based Plant Diseases Forecasting System using Fuzzy Logic. NUiCONE-13. Ahmedabad, India.

[14]. N. G. Hedeshi and M. S. Abadeh, "An ensemble PSO-based approach for diagnosis of coronary artery disease," *2011 International Symposium on Artificial Intelligence and Signal Processing (AISP)*, Tehran, 2011, pp. 77-82.doi: 10.1109/AISP.2011.5960975.

[15]. Dey, Rajeeb& Bajpai, Vaibhav & Gandhi, Gagan& Dey, Barnali. (2009). Application of Artificial Neural Network (ANN) technique for Diagnosing Diabetes Mellitus. 1 - 4. 10.1109/ICIINFS.2008.4798367.

[16]. Ayodele, Taiwo. (2010). Types of Machine Learning Algorithms. 10.5772/9385.

[17]. Ballini, Rosangela&Gomide, F. (2002). A recurrent fuzzy neural network: Learning and application. 153. 10.1109/SBRN.2002.1181460.

[18]. Ravani, P., Tripepi, G., Malberti, F., Testa, S., Mallamaci, F., & Zoccali, C. (2005). Asymmetrical dimethylarginine predicts progression to dialysis and death in patients with chronic kidney disease: a competing risks modeling approach. *Journal of the American Society of Nephrology*, *16*(8), 2449-2455.

[19]. R. Brachman, T. Khabaza, W.Kloesgan, G.Piatetsky Shapiro & E. Simoudis (1996), “Mining Business Databases”, *Comm. ACM*, Vol. 39, No. 11, Pp. 42–48.

[20]. T. Mitchell (1997), “Machine Learning”, *McGraw Hill*.

[21]. J. Han & M. Kamber (2000), “Data Mining; Concepts and Techniques”, *Morgan Kaufmann Publishers*.

[22]. World Health Organization. (2007). Everybody's business--strengthening health systems to improve health outcomes: WHO's framework for action.

[23]. Xavier, B., &Dahikar, P. (2013). A perspective study on patient monitoring systems based on wireless sensor network, its development and future challenges. *International Journal of Computer Applications*, *975*, 8887.

[24]. Pandian, P. S., Safeer, K. P., Gupta, P., Shakunthala, D. T. I., Sundersheshu, B. S., &Padaki, V. C. (2008). Wireless sensor network for wearable physiological monitoring. *JNW*, *3*(5), 21-29.

[25]. Zadeh, L. A. (1994). Fuzzy logic, neural networks, and web computing. *Communications of the ACM*, *37*(3), 77-85.

[26]. Thomas, J., &Princy, R. T. (2016, March). Human heart disease prediction system using data mining techniques. In *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)* (pp. 1-5). IEEE

[27]. Dewan, A., & Sharma, M. (2015, March). Prediction of heart disease using a hybrid technique in data mining classification. In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)* (pp. 704-706). IEEE

[28]. Rani, K. U. (2011). Analysis of heart diseases dataset using neural network approach. *arXiv preprint arXiv:1110.2626*.

[29]. Durairaj, M., &Kalaiselvi, G. (2015). Prediction of diabetes using web computing techniques-A survey. *International journal of scientific & technology research*, *4*(3), 190-192

[30]. Kumari, V. A., & Chitra, R. (2013). Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications*, *3*(2), 1797-1801

[31]. Masethe, H. D., &Masethe, M. A. (2014, October). Prediction of heart disease using classification algorithms. In *Proceedings of the world Congress on Engineering and computer Science* (Vol. 2, pp. 22-24).

[32]. Masetic, Z., & Subasi, A. (2016). Congestive heart failure detection using random forest classifier. *Computer methods and programs in biomedicine*, *130*, 54-64

[33]. Subasi, A., Alickovic, E., &Kevric, J. (2017). Diagnosis of chronic kidney disease by using random forest. In *CMBEBIH 2017* (pp. 589-594). Springer, Singapore.

[34]. Neves, J., Martins, M. R., Vilhena, J., Neves, J., Gomes, S., Abelha, A., & Vicente, H. (2015). A web computing approach to kidney diseases evaluation. *Journal of medical systems*, *39*(10), 131.

[35]. Polat, H., Mehr, H. D., & Cetin, A. (2017). Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. *Journal of medical systems*, *41*(4), 55

[36]. Nilashi, M., Bin Ibrahim, O., Mardani, A., Ahani, A., &Jusoh, A. (2018). A web computing approach for diabetes disease classification. *Health Informatics Journal*, *24*(4), 379-393.

[37]. Saxena, K., & Sharma, R. (2016). Efficient heart disease prediction system. *Procedia Computer Science*, *85*, 962-969.