The Early Augmentationfor Diabetes Diagnosis Using Data Mining Approaches

[1]Baratam Yasaswi, [2]Bodapati Prajna
[1,2]Dept. of Computer Science and Systems Engg., Andhra University, Visakhapatnam, AP, India

Abstract
In medicinal field specialists need accessible data for decision making. Presently day's data mining strategy is connected in therapeutic exploration so as to break down substantial volume of restorative data. This study endeavors to utilize data mining technique to investigate the database of diabetes and utilized for diagnosing for documenting better result. So this paper concentrate on examination of diabetes data by different data mining system. Cutting edge medication creates a lot of data which is betrayed into the restorative database. An appropriate examination of such data may uncover some fascinating truths, which may some way or another be covered up or go disperse. Data mining is one such field which tries to concentrate some fascinating actualities from immense data set. In this paper an endeavor is made to break down the diabetic data set and get some fascinating truths from it which can be utilized to build up the expectation model. Illness conclusion is one of the applications where data mining instruments are demonstrating effective results. Diabetes malady is the main source of death everywhere throughout the world in the previous years. A few scientists are utilizing factual data. The accessibility of enormous measures of therapeutic data prompts the requirement for intense mining instruments to help medicinal services experts in the analysis of diabetes infection. Utilizing data mining method as a part of the determination of diabetes malady has been completely explored, demonstrating the satisfactory levels of exactness. As of late specialists have been examining the impact of hybridizing more than one procedure demonstrating improved results in the finding of diabetes ailment.

Keywords
Data Mining, Diabetes, C4.5 Classifiers.

I. Introduction

Data mining is the investigation of expansive datasets to previously unknown patterns, relationships and knowledge that are hard to identify with conventional factual strategies. Data mining is quickly becoming effective in an extensive variety of uses, for example, examination of natural mixes, budgetary determining, medicinal services and climate anticipating Data mining in social insurance is a developing field of high significance for giving guess and a more profound comprehension of therapeutic data. Data mining applications in medicinal services incorporate investigation of social insurance communities for better wellbeing arrangement making and avoidance of clinic mistakes, early location, counteractive action of sicknesses and preventable healing facility passing's, more esteem for cash and cost reserve funds, and discovery of fake protection claims. Analysts are utilizing data mining systems as a part of the finding of a few infections, for example, diabetes, stroke, tumor, and coronary illness. Diabetes ailment is the main source of death on the planet in the course of recent years. Propelled by the overall expanding mortality of Diabetes infection patients every year and the accessibility of tremendous measure of patients' data from which to remove valuable learning, scientists have been utilizing data mining strategies to help human services experts in the conclusion of Diabetes sickness. Building up a device to be inserted in the doctor's facilities administration framework to encourage and offer exhortation to the social insurance experts in diagnosing and giving reasonable treatment to Diabetes malady patients is imperative. A few data mining systems are utilized as a part of the conclusion of Diabetes malady, for example, Naïve Bayes, Decision Tree, neural system, piece thickness, naturally characterized bunches, sacking calculation, and bolster vector machine indicating distinctive levels of correctness's. In spite of the fact that applying data mining in illness finding and treatment is advantageous, less research has been done in distinguishing treatment anticipates patients and particularly for Diabetes ailment patients. Specialists have demonstrated that healing centers don't give the same nature of administration despite the fact that they give having Diabetes ailment. Diabetes malady experts store noteworthy measures of patient's data. It is essential to break down these datasets to separate valuable learning.[2]Data mining is a powerful instrument for examining data to separate helpful learning. A tremendous measure of data gets amassed in the healing facilities, the greater part of them simply get put away in some type of documents which are never touched back, if these data are broke down legitimately they help in determining some intriguing realities. A little touch of data mining will help in producing intriguing realities which stayed unrevealed something else, subsequently thinking about the diabetes mellitus a point by point examination of diabetic data set is performed utilizing data mining procedure [4].

II. Related Work

Diabetes is an especially perfect malady for data digging innovation for various reasons. To begin with, in light of the fact that the heap of data is there and second, diabetes is a typical ailment that costs a lot of cash, thus has pulled in administrators and payers in the ceaseless mission for sparing cash and cost productivity. Third, diabetes is an ailment that can deliver loathsome difficulties of visual impairment, kidney disappointment, removal, and untimely cardiovascular passing, so doctors and controllers might want to know how to enhance results however much as could reasonably be expected. Data mining may demonstrate a perfect match in these circumstances. Keeping the ailment of diabetes is a progressing region important to the social insurance group. Taking into account the data from the 2011 National Diabetes Fact Sheet, diabetes influences an assessment of 25.8 million individuals in the US, which is around 8.3% of the populace. Also, around 79 million individuals have been determined to have prediabetes [7]. Pre-diabetes alludes to a gathering of individuals with higher blood glucose levels than ordinary however not sufficiently high for a finding of diabetes. Expanded mindfulness and treatment of diabetes ought to start with anticipation. A significant part of the attention has been on the effect and significance of preventive measures on sickness event and particularly cost reserve funds came about because of such measures. Numerous studies in regards to diabetes expectation have been directed for quite a long while. The primary goals are to foresee what variables are the causes, at

high hazard, for diabetes and to give a preventive activity toward individual at expanded danger for the sickness. A few variables have been accounted for in writing, which are clarified in the following heading.[3][24] As indicated by WHO 2011 report: • 347 million individuals worldwide have Diabetes Mellitus. • In 2004, an expected 3.4 million individuals kicked the bucket from outcomes of high glucose. • More than 80% of Diabetes Mellitus passings happen in low-and center pay nations. • WHO extends those Diabetes Mellitus passings will twofold somewhere around 2005 and 2030. Sound eating regimen, customary physical action, keeping up a typical body weight and staying away from utilization of tobacco can avoid or defer the onset of sort 2 Diabetes Mellitus [8]. Data mining is the procedure of recovering data from an extensive data distribution center where data is recovered taking into account expectation. Data mining is additionally called as learning discovery in database (KDD) [9] [10]. The forecasts that are utilized to discover the data from the distribution center are resolved with the assistance of different spaces like counterfeit consciousness, machine learning, measurements, business insight and database framework. Data mining has sway on different fields which incorporates amusements, business, science and building, human rights, restorative data mining, spatial data mining, sensor data mining, visual data mining, music data mining, observation, design mining, subject based data mining, learning lattice [10]. Our medicinal social insurance frameworks are rich in data however poor in learning so there is a colossal need of having a strategies and devices to concentrate data from the enormous data set so that restorative finding should be possible [11]. SantiWulanPurnami et al. [12][23], in their examination work utilized bolster vector machine for highlight determination and grouping of bosom malignancy furthermore stresses how 1-standard SVM can be utilized as a part of highlight choice and smooth SVM (SSVM) for characterization. Two issues tended to here are, the first is to distinguish the significance of the parameters on the bosom tumor. The second research issue is to analyze bosom malignancy in light of nine characteristics of Wisconsin bosom disease dataset. To recognize the significance of the parameters, the 1-standard SVM of the first data was finished. The more grounded parameters are as per the following: parameter 1 (Clump thickness), parameter 3 (Uniformity of Cell shape), parameter 6 (Bare Nuclei), parameter 7 (Bland Chromatin), and parameter 9 (Mitoses), while parameter 2 (Uniformity of Belsize), parameter 4 (Marginal Adhesion), parameter 5 (Single Epithelial Cell Size) and parameter 8 (Normal Nucleoli) are weaker. The got preparing and testing order exactness utilizing 10 fold cross approval were 97.52% and 97.01% individually. When one of the feeble parameters was expelled both preparing and testing demonstrates a little diminishing in precision [5].

III. Methodology

Utilized as a part of Data Mining Different data mining methods have been utilized to help medicinal services experts in the conclusion of Diabetes infection. Those most every now and again utilized spotlight on order: native bayes choice tree, and neural system. Other data mining strategies are likewise utilized including part thickness, consequently characterized bunches, stowing calculation, and bolster vector machine. In spite of the fact that applying data mining is valuable to human services, illness finding, and treatment, few inquires about have explored creating treatment anticipates patients [13].

A. Pre-processing

Before data mining calculations can be utilized, an objective data set must be gathered. As data mining can just reveal designs really introduce in the data, the objective data set must be sufficiently extensive to contain these examples while staying sufficiently brief to be mined inside an adequate time limit. A typical hotspot for data is an data bazaar or data stockroom. Pre-handling is key to break down the multivariate data sets before data mining. The objective set is then cleaned. Data cleaning expels the perceptions containing clamor and those with missing data [14].

B. Extraction of Useful Knowledge

- Brings an arrangement of instruments and procedures that can be connected to this handled data to find shrouded designs
- That give social insurance experts an extra wellspring of learning for deciding
- The choices comes about then coordinated with social insurance experts sentiments [6].

Essentially expressed, "Data mining alludes to removing or "mining" learning from a lot of data". There are some different terms which convey a comparable or somewhat distinctive intending to data mining, for example, learning mining from data, data extraction, data or example investigation, and data paleohistory. Data mining functionalities are utilized to determine the sort of examples to be found in data mining undertakings. When all is said in done, arranged data mining assignments into two classes: unmistakable and prescient. Expressive mining undertakings describe the general properties of the data in the database where as prescient mining assignments perform deduction on the present data with a specific end goal to make expectations. Data mining, likewise prominently known as Knowledge Discovery in Database alludes to the procedure of finding intriguing learning from a lot of data put away in databases, data distribution centers, or other data storehouses. Data mining procedures are utilized to work on vast volumes of data to find concealed examples and connections accommodating in basic leadership. While data mining and learning disclosure in database are much of the time regarded as equivalent words, data mining is entirely of the data discovery process [15].

IV. Data Mining

Data Mining represents a process developed to examine large amounts of data routinely collected. The term also refers to a collection of tools used to perform the process. Data collected from various areas such as marketing, health, communication, etc., are used in data mining. Data Mining is the extraction of hidden predictive data from large databases; it is a powerful technology with great potential to help organisations focus on the most important data in their data warehouse. Questions those traditionally were too time consuming to resolve can be answered by the data mining tools in an effective manner. This helps to find the hidden patterns, predictive data that facilitates the experts with solution outside their expectations [16]. The goal of data mining is to extract knowledge from dataset in humanunderstandable structures. In recent years data mining has been used widely in the areas of science and engineering, such as bioinformatics, genetics, medicine, education and engineering.

A. Techniques used in Data Mining

Clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more

similar (in some sense or another) to each other than to those in other clusters. Clustering is a main task of explorative data mining, and a common technique for statistical data analysis used in many fields, including machine learning, pattern recognition, image analysis, data retrieval and bioinformatics.

Classification trees are used to predict the classes of a categorical dependent variable from their measurements on one or more predictor or independent variables. Decision Trees have emerged as a powerful technique for modeling general input / output relationships [19-20]. They are tree – shaped structures that represents a series of roles that lead to sets of decisions. They generate rules for the classification of a dataset and a logical model represented as a binary (two – way split) tree that shows how the value of a target variable can be predicted by using the values of a set predictor variables. Decision trees, which are considered in a regression analysis problem, are called regression trees. Thus, the decision tree represents a logic model of regularities of the researched phenomenon.

Part is a rule based algorithm and produces a set of if then rules that can be used to classify data. It is a modification of C4.5 and RIPPER algorithms and draws strategies from both. PART adopts the divide-and-conquer strategy of RIPPER and combines it with the decision tree approach of C4.5. PART generates a set of rules according to the divide-and conquer strategy, removes all instances from the training collection that are covered by this rule and proceeds recursively until no instance remains. To generate a single rule, PART builds a partial decision tree for the current set of instances and chooses the leaf with the largest coverage as the new rule. It is different from C4.5 because the trees built for each rules are partial, based on the remaining set of examples and not complete as in case of C4.5. 2.2.4. ID3 Algorithm starts with all the training samples at the root node of the tree. An attribute is selected to partition these samples. For each value of the attribute a branch is created, and the corresponding subset of samples that have the attribute value specified by the branch is moved to the newly created child node. The algorithm is applied recursively to each child node until all samples at a node are of one class. Every path to the leaf in the decision tree represents a classification rule [17-18].

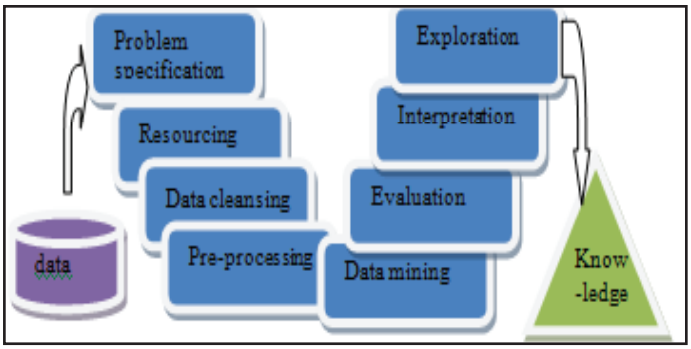

Fig. 1: Data Mining Process

V. Diabetes

Most of the food we eat is converted to glucose, or sugar which is used for energy. The pancreas secretes insulin which carries glucose into the cells of our bodies, which in turn produces energy for the perfect functioning of the body. When you have diabetes, your body either doesn't make enough insulin or cannot use its own insulin as well as it should. This causes sugar to build up in your blood leading to complications like heart disease, stroke, and neuropathy, poor circulation leading to loss of limbs, blindness, kidney failure, nerve damage, and death [21].

General Symptoms of Diabetes
- Increased thirst
- Increased urination - Weight loss
- Increased appetite - Fatigue
- Nausea and/or vomiting - Blurred vision
- Slow-healing infections - Impotence in men Types of Diabetes

Type 1 - Diabetes also called as Insulin Dependent Diabetes Mellitus (IDDM), or Juvenile Onset Diabetes Mellitus is commonly seen in children and young adults however, older patients do present with this form of diabetes on occasion. In type 1 diabetes, the pancreas undergoes an autoimmune attack by the body itselftherefore; pancreas does not produce the hormone insulin. The body does not properly metabolize foodresulting in high blood sugar (glucose) and the patient must rely on insulin shots. Type I disorder appears in people younger than 35, usually from the ages 10 to 16.

Type II - Diabetes is also called as Non-Insulin Dependent Diabetes Mellitus (NIDDM), or Adult Onset Diabetes Mellitus. Patients produce adequate insulin but the body cannot make use of it as there is a lack of sensitivity to insulin by the cells of the body. Type II disorder occurs mostly after the 40.India has the dubious distinction of being the diabetic capital of the world. Home to around 33 million people with diabetes, 19% of the world's diabetic population is from India. Nearly 12.5% of Indian's urban populations have diabetes. The number is expected to escalate to an alarming 80 million by the year 2030. Amongst the chronic diabetic complications, diabetic foot is the most devastating result. Diabetes in India. Diabetes patients can often experience loss of sensation in their feet. Even the smallest injury can cause infection that can be various serious. 15% of patients with diabetes will develop foot ulcers due to nerve damage and reduced blood flow. Diabetes slowly steals the person's vision. It is the cause for common blindness and cataracts. Over 50,000 leg amputations take place every year due to diabetes in India. Diabetes patients can often experience loss of sensation in their feet. Even the smallest injury can cause infection that can be various serious. 15% of patients with diabetes will develop foot ulcers due to nerve damage and reduced blood flow. Diabetes slowly steals the person's vision. It is the cause for common blindness and cataracts.

VI. Tools Used For Diabetes Prediction

There are different soft computing techniques and tools are applied for the prediction and data analysis. In this section, some of the techniques are discussed.

A. Artificial Neural Network

The artificial neural network is much similar as natural neural network of a brain. Artificial Neural Network (ANN) basically has three layers, they are;

B. Input Layer:

Input neurons define all the input attribute values for the data mining model, and their probabilities.

C. Hidden Layer:

Hidden neurons receive inputs from input neurons and provide outputs to output neurons. The hidden layer is where the various probabilities of the inputs are assigned weights. A weight describes the relevance or importance of a particular input to the hidden neuron. The neuron with greater weight is assigned to an input.

The value of that input is more important weights can be negative, which means that the input can inhibit, rather than favour, a specific result.

D. Output layer:

Output neurons represent predictable attribute values for the data mining model.

F. C4.5 Algorithm

Accepted decision tree algorithms consist of C4.5, the equivalent time as the name imply. The performance of C4.5 is recursively separate inspection in branches to build tree for the purpose of improving the calculation accuracy. Systemsthat construct classifiers are one of the commonly used tools in data mining. Such systems take as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs [22].

VII. Proposed System Architecture

The diabetic data set is given as input to the system, which comprises of hive and R. The raw data is just a file consisting of comma separated values, for the first time when we look into it, it just looks like a junk of data. But a proper analysis of this data set will reveal some interesting facts. The raw input is given as input to hive, the data set is analysed and partitioned based on different attribute the output which is obtained from hive is well formatted data, and then this output is given as input to R. It is one of the best languages which is used for statistical computing as well as for generating graphs. As we all know that pictures speak more than words, after analysing the data using hive the graphs are generated for each data set using R.

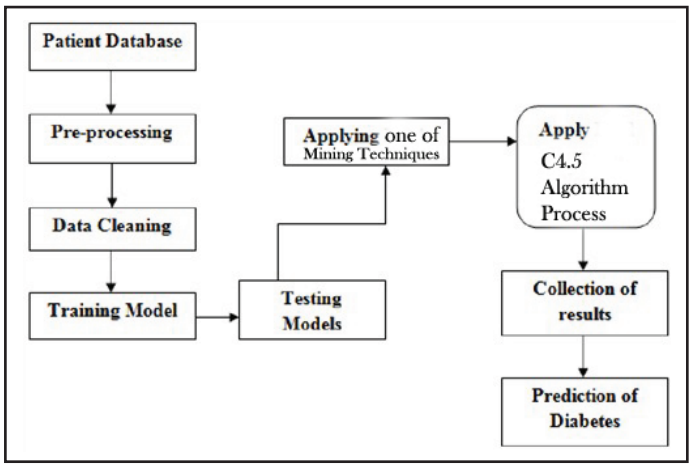

Fig. 2: Architecture of the Proposed System

A. The Data

There are many risk factors that become a cause of diabetes. It is very difficult to diagnose these factors easily. Most of the time the disease is diagnosed at the last stage of disease. With the help of risk factors, it is easy to diagnose disease possibilities in advance. The dataset used in this research composed of 9 risk factors are blood cholesterol, plasma glucose, diastolic blood pressure, triceps (SFT), Insulin, BMI, DPF, age class. On the basis of these risk factors the results are computed to know whether the patient has risk of Diabetes or not. The dataset contains 768 people data collected from the UCI repository. The dataset consists of 9 attributes as shown in Table 1.

Table 1: Attributes of Diabetes Dataset

S. No.	Name of the Attributes	Description
1.	Blood cholesterol	Blood cholesterol is measured in units to measure the cholesterol levels (mmol/l)
2.	Plasma glucose	Plasma glucose concentration measured using two hours oral glucose tolerance test (mm Hg)
3.	Diastolic BP	Diastolic blood pressure
4.	Insulin	2 hours serum insulin (mu U/ml)
5.	Triceps (SFT)	Triceps skin fold thickness (mm)
6.	BMI	Body Mass Index (weight kg/ height in $(mm)^2$
7.	DPF	Diabetes Pedigree function
8.	Age	Age of Patient
9.	class	Diabetes on set within 5 years

VIII. Result Validations and discussions

Data mining can unintentionally be misused, and can then produce results which appear to be significant; but which do not actually predict future behavior and cannot be reproduced on a new sample of data and bear little use. Often this results from investigating too many hypotheses and not performing proper statistical hypothesis testing. A simple version of this problem in machine learning is known as over fitting, but the same problem can arise at different phases of the process and thus a train/test split - when applicable at all - may not be sufficient to prevent this from happening The final step of knowledge discovery from data is to verify that the patterns produced by the data mining algorithms occur in the wider data set. Not all patterns found by the data mining algorithms are necessarily valid. It is common for the data mining algorithms to find patterns in the training set which are not present in the general data set. This is called over fitting. To overcome this, the evaluation uses a test set of data on which the data mining algorithm was not trained. The learned patterns are applied to this test set, and the resulting output is compared to the desired output. For example, a data mining algorithm trying to distinguish "spam" from "legitimate" emails would be trained on a training set of sample e-mails. Once trained, the learned patterns would be applied to the test set of e-mails on which it had not been trained. The accuracy of the patterns can then be measured from how many e-mails they correctly classify.

VIII. Conclusion

Data mining and C4.5 algorithms in the medical field extracts distinctive concealed patterns from the medical data. They can be utilized for the examination of vital clinical parameters, expectation of different diseases, estimating assignments in pharmaceutical, extraction of medical knowledge, treatment planning support and patient administration. Various algorithms are present in literature for the prediction and finding of diabetes. These techniques give more precision than the accessible conventional frameworks.The proposed approach has shown that mining helps to retrieve useful correlation even from attributes which are not direct indicators of the class we are trying to predict. Moreover these data analysis results can be used for further research in enhancing the accuracy of the prediction system in future.

References

[1] UCI Machine Learning Repository- Center for Machine Learning and Intelligent System, http://archive.ics.uci.edu.

[2] PardhaRepalli, "Prediction on Diabetes Using Data mining Approach".

[3] Joseph L. Breault., "Data Mining Diabetic Databases:Are Rough Sets aUseful Addition".

[4] G. Parthiban, A. Rajesh, S.K.Srivatsa, "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method ", International Journal ofComputer Applications (0975 – 8887) Volume 24– No.3, June 2011.

[5] P. Padmaja, "Characteristic evaluation of diabetes data using clustering techniques", IJCSNS International Journal of Computer Science and NetworkSecurity, VOL.8 No.11, November 2008.

[6] Lin, C., Lee, C., "Neural Fuzzy Systems," PrenticeHall, NJ, 1996.

[7] Tsoukalas, L., Uhrig, R., "Fuzzy and Neural Approaches in Engineering," John Wiley & Sons, Inc., NY, 1997.

[8] De Oliveira, J. V., &Pedrycz, W. (2007).Advances in fuzzy clustering and its applications. (1 ed., pp. 4-69). London: Wiley.

[9] Everitt, S., Landau, S., Leese, M. (2011). Cluster Analysis. (5 ed., pp. 76- 80). London: Wiley.

[10] T. J. Ross, Fuzzy Logic with Engineering Applications, Third Edition, John Wiley & Sons, 2010

[11] Abdullah A. Aljumah, Mohammed GulamAhamad, Mohammad Khubeb Siddiqui," Application of data mining: Diabetes health care in young and old patients",2012,http://www.sciencedirect.com/science/article/pii/S1319157812000390

[12] JyotiSoni, Ujma Ansari, Dipesh Sharma, SunitaSoni," Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", 2011, http://citeseerx.ist.psu.edu/viewdoc/download?rep=rep1&type=pdf&doi=10.1.1.206.3899

[13] HianChyeKoh ,Gerald Tan," Data Mining Applications in Healthcare", http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.92.31 84&rep=rep1&type=pdf

[14] Joseph L. Breault, MD, MPH, MS, "Data Mining Diabetic Databases: Are Rough Sets a Useful Addition?",http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.815&rep=rep1&type=pdf

[15] Mancia G., DeBacker G., Dominiczak A., Cifkova R., Fagard R., and Germano G., "ESHESC Practice Guidelines for the Management of Arterial Hypertension: ESH-ESC Task Force on the Management of Arterial Hypertension," Journal Hypertens, vol. 25, no. 9, pp. 1751-1762, 2007.

[16] Rajesh K. and Sangeetha V., "Application of Data Mining Methods and Techniques for Diabetes Diagnosis," the International Journal of Engineering and Innovative Technology, vol. 2, no. 3, pp. 224-229, 2012.

[17] Richards G., Rayward-Smith V., Sonksen P., Carey S., and Weng C., "Data Mining For Indicators of Early Mortality in A Database of Clinical Records," Artificial Intelligence in Medicine, vol. 22, no. 3, pp. 215-231, 2001.

[18] T. Santhanam a, M.S Padmavathi b, "Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis", Procedia Computer Science 47 (2015) 76 – 83, Elsevier.

[19] Syed Umar Amin, Kavita Agarwal, Dr. Rizwan Beg, "Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors", Proceedings of 2013 IEEE Conference on Information and Communication Technologies.

[20] Pinky Bajaj, KavitaChoudhary, Renu Chauhan, "Prediction of Occurrence of Heart Disease and Its Dependability on RCT Using Data Mining Techniques", 2015 Advances in Intelligent Systems and Computing 340, Springer.

[21] Sujata Joshi, Mydhili K. Nair, "Prediction of Heart Disease Using Classification Based Data Mining Techniques", 2015 Computational Intelligence in Data Mining - Volume 2, Springer.

[22] V Krishnaiah, M Srinivas, Dr.GNarsimha, Dr.NSubhash Chandra, "Diagnosis of Heart Disease Patients Using Fuzzy Classification Technique", IEEE, 2014.

[23] BakshiRohit Prasad, Sonali Agarwal, "Modeling Risk Prediction of Diabetes – A Preventive Measure", IEEE, 2014.

[24] Sulaimon Ibrahim, Pradeep Chowriappa, SumeetDua, U. Rajendra Acharya, Kevin Noronha, SulathaBhandary, HatwibMugasa,"Classification of diabetes maculopathy images using data-adaptive neuro-fuzzy inference classifier", International Federation for Medical and Biological Engineering 2015 Springer.

Baratam Yasaswi is pursuing 5-year integrated dual degree (B.tech+M.tech) in the Dept. of Computer Science and Systems Engineering, Andhra University College of Engineering(A), Visakhapatnam, Andhra Pradesh, India.

Prof. Bodapati Prajna is working as professor in Dept. of Computer Science and Systems Engineering, Andhra University College of Engineering (A), Visakhapatnam, Andhra Pradesh, India.