International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE) Vol 1, Issue 4, Mar 2015

Prediction and Analysis ofDisease through Data Mining Techniques

[1]ArunaPavate, [2]Nazneen Ansari
Student, M.E. (Computer Engineering),Associate Professor (IT)
St. Francis Institute of Technology, Mumbai-400103, India.
gavkare@gmail.com,nazsfit@yahoo.com

Abstract: **—This Latest improvements in medical field and wide usage of electronic medical health records help in wide production of clinical data in hospitals. Most of the clinical data contains valuable information about patients. Proper usage of this information can be used for predicting different diseases. Diabetes is one of the serious disease for which the patient's personal participation is required to treat the disease. As diabetes is known to be a silent killer, an efficient controlling of the disease is essential. The aim of this research is to propose new approach using data mining techniques to identify the diabetic disease, its categories and further complications arises like heart disease, kidney disease, neuropathy, stroke, vision loss from the clinical database using KNN Algorithm and its versions, fuzzy logic ,genetic algorithm and its versions in an efficiently and an economically faster manner.**

Index Terms— **Fuzzy Logic; Hybrid Genetic; Diabetes disease Complications.**

I. INTRODUCTION

Current lifestyle and increasing intake of unhealthy diets cause obesity among adults and lead to diabetes mellitus. Diabetes disease has become a major health issue in all over world. Healthcare industries maintain huge amount of records about the patient disease and causes of diseases. These huge records serve as a source for the knowledge extraction and identification of diseases. Many methods have been proposed to detect and prevent diabetes. Patients with Diabetes Mellitus are at increased risk of developing heart disease, kidney disease, neuropathy, stroke, vision loss. Few studies have been devoted to the development of risk prediction models specific for diabetes population [1]. The number of people with type 2 diabetes is high and increasing rapidly: 366 million people worldwide were estimated to have type 2 diabetes in 2011, and rate is expected to rise by 51% by 2030 [2]. Global health expenditure to check and treat diabetes will cost around 490 billion dollars by 2030 [3]-[4]. Diabetic's complication goes undiagnosed due to lack of standardized diagnostic criteria and no clear definition or information [6]. The main aim of this research is to develop accurate, simple to use, cost effective methods to support medical practitioners to predict the diabetes disease and it complications in early stage. The early prediction of disease give a warning about the level of risk and complications that arise due to diabetes, where treatments and preventative action can facilitate the patient to increase the period of patient's healthy life.

A. Social Value Contribution:-

- To support in diagnosis and classification of disease with a large number of inputs and diagnose stressful situations in early stage.
- Based on modeled historical performance, one can select best medication course: e.g. best treatment plans.
- This system suitable in hospital management to optimize allocation of resources and assist in future planning for improved services like - Predicting length-of-stay for incoming patients.
- Computer-based training and evaluation : Disease models for the instruction and assessment of undergraduate medical and nursing students

B. Diabetes Mellitus

Diabetes describes a group of metabolic disease in which the person has high blood glucose because insulin production is insufficient, or because the body's cells do not react properly to insulin, or both [5]. Diabetes (diabetes mellitus) is classed as a metabolism disorder. Metabolism means the way our bodies use digested foodfor energy and growth. There are two names to diabetes 1) Diabetes mellitus (was derived from theGreek word meaning passing through and sweet as honey in which person has high

blood glucose). 2) Diabetes Insipidus – excess of fluid loss by the body.

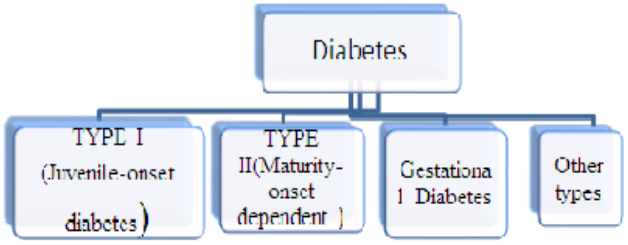

Figure 1: Types of diabetes

As shown in figure 1, Type I is insulin dependent diabetes. About 2500 children have this disease because pancreas produces little or no insulin and such patients are dependent on outside source of insulin [6]. Type I diabetes is usually found in youth, but can occur at any age. Such type of diabetes appears abruptly and progresses rapidly and not always present in other family members. Type I diabetes is not necessarily linked to obesity.[6]-[7]

Type II is non-insulin dependent diabetes. Additional insulin is not required to sustain life. Type II is much more common than type I. Usually such type of disease can be developed during adulthood and is more common in older people. This disease may go unnoticed for years. Type II diabetes was present in other family members. In Type II, insulin injections are not necessary. In such cases, 80 percent of all patients are overweight at time of diagnosis.[6]-[7]. Gestational diabetes refers to the type of diabetes that develops during pregnancy. Other types of diabetes are associated with certain conditions such as diabetes induced by drugs or chemicals. [9] Mostly all types of diabetes affect blood vessels, eye, kidneys and nerves. Other many complications like long-term injury, dysfunction, and functional abnormalities in eyes, kidneys, nerves, heart, and blood vessels arise [10]. Thus, it is considered as one of the most important health issues [11].Figure 2 shows main symptoms of diabetes in type 1 and type 2,type1 symptoms such as weight loss, nausea, vomiting abdominal pain etc. and type 2 symptoms such as blurred vision, glycosuria etc.[8]

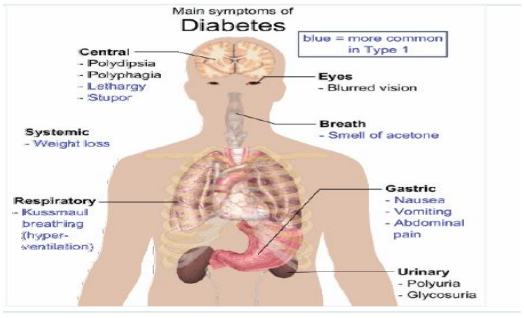

Figure 2: Symptoms of diabetes [8]

I. LITERATURE SURVEY

Most of the study includes predication of the diabetes disease. This work is an incremental improvement to the previous work which includes complications that arise due to diabetes so that old and young age group diabetes patients should be given an assessment and a treatment plan that is suited to their needs and lifestyles. Intake control, weight reduction, workout and smoking cessation are mutually beneficial to each other for the treatment of diabetes.

There are lots of literatures cited recently related with different disease diagnosis using data mining. Several algorithms of risk stratification and diagnostic models for Coronary Heart Disease [CHD] have created with different sets of risk factors [1]. Prediction models for stroke risk analysis using stacked topology of ANN model, support vector classification models, prediction through fuzzy inference system, neural prognostic models were developed by Sabibullah et al. [4]- [11]- [12]. Data mining is a multidisciplinary borrowing idea from data bases, Machine Learning (ML) and artificial intelligence, statistics, pattern recognitions etc. Genetic algorithm is one of the soft computing techniques that automatically solves problem. The proposed model using this technique organizes the data based on their relevant attributes and transforms it into human interpretable patterns or correlation. Feature selection algorithms detect the features that are relevant but not redundant to the solution. The main task is to rank the related features based on their fitness values. There are many algorithms that use a greedy search throughthe solution space. Decision tree algorithms such as Quinlan's ID3 [14] and C4.5 [15] ,CARTproposed in [15]aresome of the most successful supervised learning algorithms There are again many algorithms that are exhaustive, heuristic and random search..

P.C. Thirumal and N. Nagarajan proposed a method of averaging a K Nearest Neighbour Algorithm to detect Type-2 Diabetes.[17]. KNN is a non-parametric lazy algorithm. The proposed model Average K Nearest Neighbor Algorithm is similar to the KNN algorithm except that this reduces the time needed for classification in KNN by forming the super samples for each class. Other algorithms such as Smart Average KNN and Partial Average KNN can be used to obtain a higher accuracy.

II. PROPOSED APPROACH

The research has proposed a model which are compound of KNN and Genetic Algorithm, Fuzzy logic techniques. It is known as hybrid systems to identify and categorize diabetes mellitus, further complications arises because of diabetes and the risk associated with this.

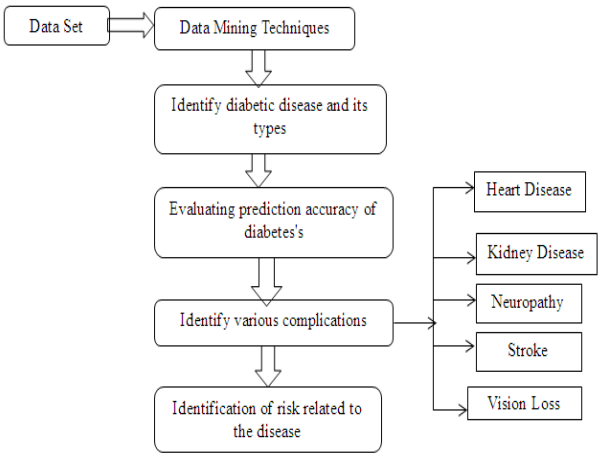

Fig: 4 Flow of the proposed system

Figure 4 shows the flow of proposed work, the input consisting of different attributes of disease. Different data mining techniques such K nearest neighbor, genetic algorithm and fuzzy logic are used for predicating diabetes disease. KNN and genetic algorithm are used to compute the best fitness value in evaluating the prediction accuracy of diabetes from clinical database. To identify the various complications and risk associated, fuzzy logic technique is used. Different complications related with the diabetes disease are long-term injury, dysfunction, and functional abnormalities in eyes, kidneys, nerves, heart, and blood vessels etc. and the risk associated with complication is considered as high, medium risk, low risk.

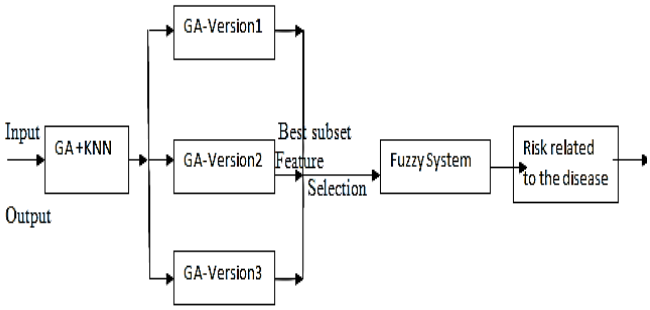

Fig 5: Proposed System

Figure 5 shows the proposed system. In the present study, the feature selection task is performed by a KNN and genetic algorithm, where the proposed subsets of features are ranked using different versions of a KNN classifier. According to the KNN algorithm, a query instance is assigned to the class represented by a majority of its k-nearest neighbor's in the training set. The WKNN Algorithm is considered to assign weights to the closer neighbors more heavily than the farther ones [10]. In WKNN, the distance-weighted function Wi to the i-th nearest neighbor is defined as,

$$Wi = \frac{k+1-i}{\sum_{m=1}^{k} m} \qquad (1)$$

Where

m➔ An integer in the interval (1,k) and

k➔ Total number of the neighbor's.

Thus, all weights are in the interval $\frac{1}{\sum_{m=1}^{k} m}$ $\frac{k}{\sum_{m=1}^{k} m}$ and a neighbor with smaller distance is weighted more heavily than one with greater distance.

In the DWKNN Algorithm, in order to address the effect of the number of neighbor's on the classification performance, a DWKNN algorithm has been proposed [12]. The DWKNN algorithm gives different weights to the k nearest neighbor's depending on distances between them and their ranking according to their distance from the query object.

The distance-weighted function of Wi the i-th nearest neighbor is computed as,

$$Wi = \begin{cases} \frac{dk^{NN} - di^{NN}}{dk^{NN} - d1^{NN}} \times \frac{1}{i}, & dk^{NN} \neq d1^{NN} \\ 1, & dk^{NN} = d1^{NN} \end{cases} \qquad (2)$$

Where

di^{NN} ➔ Distance of the i-th nearest neighbour from the query object

$d1^{NN}$ ➔ Distance of the nearest neighbour,

dk^{NN} ➔Distance of the k-furthest neighbour.

Thus, the weight of the nearest neighbor is 1, and the weight of the furthest k-th neighbor is 0, whereas other weights are distributed between 0 and 1.

The different version of the GA varies depending upon two criteria's:

1) On the use of the WKNN classifier or the use of the DWKNN classifier for the ranking of the candidate chromosomes in the FF.
2) The different versions of the GA vary on the number of criteria used for the ranking of the chromosomes in the FF. The value of the FF for each candidate chromosome is calculated according to,

$$FF = b*(Se) + (1-b)*(\frac{NIF}{L}) \qquad (3)$$

Where

b➔ One parameter of the FF that favors the importance of the classification performance of the candidate chromosome versus the importance of the number of not important features in the candidate chromosome.

Se ➔ Value of the sensitivity of the chosen classifier (WKNN or DWKNN) for the diabetes dataset.

NIF ➔ Number of not important features for the candidate chromosome

L ➔ Total number of features.

The difference between genetic algorithm and its different

versions is as follows:

GA-version 1

Steps included in GA-version 1 are as follows:

1) The value of the FF is calculated with the use of the sensitivity of the WKNN classifier.
2) The candidate chromosomes are ordered in ascending order with respect to their FF value.

GA-version 2

Steps included in GA-version 2 are as follows:

1) The value of the FF is calculated with the use of the sensitivity of the WKNN classifier.
2) The candidate chromosomes are ordered in ascending order with respect to their FF value.
3) Minimum and the maximum value of the FF are calculated and the interval between these two values is divided in M subintervals. In each subinterval, the candidate chromosomes are ordered in ascending order with respect to the accuracy of the classification of the WKNN classifier.

GA-version 3

Steps included in GA-version 3 are as follows:

1) The value of the FF is calculated with the use of the sensitivity of the DWKNN classifier.
2) In this step, the second level of ranking is done, the candidate chromosomes are ordered in ascending order with respect to the accuracy of the classification of the DWKNN classifier.

The different versions of the GA will be used in order to identify which one produces the best subsets of critical features related with the diabetes mellitus disease. The sensitivity, accuracy and specificity are achieved by the best subsets of features selected by GA-version 1, GA-version 2 and GA-version 3 for different number of the k nearest neighbors of the WKNN (GA-version 1 and GA-version 2) or DWKNN (GA-version 3) classifier.

The inputs for the fuzzy toolbox that consist of the selected features and their membership functions are framed as: high, normal, medium. The basic structure of fuzzy inference system consists of four components: fuzzification module, rule base, inference engine, and defuzzification module, rule base, inference engine. The fuzzy system conjointly depends on established domain knowledge and rule based reasoning. The fuzzy rule captures all the patient symptoms and it is useful to take the correct decision at the right time.

A questionnaire has been established with set of questions for sample data collection. It mainly covers symptoms including the people current health care condition(s), life style and body symptoms which are based on a general literature review on the use of technology on diabetes [20, 21]. All questions werereviewed by Health care experts. Sample data has been collected via questionnaire distribution with participants.

III. EXPECTED RESULT

A new prediction method is proposed in this research to identify the diabetic disease, complications arises and its types from the clinical database. The techniques will be helping as a

Supportive tool to assist medical experts, peoples to predict the disease and take precautions to prevent the further complications arises. It also will reduce the cost for various medical tests and facilitate patients to require preventive measures well beforehand.

IV. CONCLUSION AND FUTURE WORK

Diagnosis of diabetes disease is challenging and often puzzled with symptoms of other diseases. If diagnosis are predicted and the risk related to the disease then by advising proper prevention care can take the patients and can prevent the complication arises in early stages. Thus, such system has been proposed. Validation shall be done based on two parameters:

- Accuracy

The accuracy is the proportion of instances that will be correctly classified.

- Sensitivity:

Sensitivity is the proportion of positive instances that will be correctly classified as positive.

- Specificity

The specificity is the proportion of negative instances that will be correctly classified as negative.

- The number of features that each algorithm selects

This study only purposes a design sketch. In the future, diabetes disease diagnosis shall be designed based on symptoms like age, sex, obesity, food intake, exercise, etc. and the system shall be implemented.

REFERENCES

[1] S. Dieren et al., "Prediction models for the risk of cardiovascular disease in patients with type 2 diabetes: a systematic review," Heart, March 2012.

[2] International Diabetes Federation. IDF Diabetes Atlas, 5th edn. Brussels International Diabetes Federation, 2011.

[3] Pickup. J and William .G (1997), "Textbook of Diabetes", 2nd ed., London, Blackwel science, Vol-1, pp.3-10.

[4] ZarithaZainuddin, Ong Pauline and CemalArdil, "A Neural Network Approach in Predicting the Blood Glucose Level for Diabetic Patients", World Academy of Science, Engineering and Technology, 2009, Vol.6, No.50, pp 981-988.

[5] Kemal polat, salihgunes, ahmetArslan, "A cascade Learning System for classification of diabetes disease: Generalized Discriminant Analysis and Least square support vector Machine", Expert systems with Applications, 34, 482-487, 2008

[6] Pickup. J and William .G (1997), "Textbook of

Diabetes", 2nd ed., London, Blackwell science, Vol-1, pp.3-10
[7] Ada P. Khan m.p.h. (1994)"Diabetes Causes, Prevention & Treatment ", India
[8] http://www.medicalnewstoday.com/info/diabetes/ "What is Diabetes? What Causes Diabetes" 2014
[9] Sabibullah M and Kashmir Raja S V,"A study on cerebrovascular disease risk factor prediction through fuzzy inference system", Intl. Jr. of System Simulation, Vol.3, No.1, 15-23, 2009.
[10] Kaio Santos, Leonardo Feistauer, Karla Rezende, "An Extension of a system to Monitor Diabetic Patients", EATIS 2012, Valencia, Spain, pp.11-18.
[11] Mohammed khalilia, SounakChakraborty and MihailPopescu, "Predicting disease risks from highly imbalanced data using random forest", BMC Medical Informatics and Decision making, 2011.
[12] Brindle P, Besmick A, Fahey T, Ebrahim S, "Accuracy and impact of risk assessment in the primary prevention of cardiovascular disease: a systematic review", Heart, 2006; 92: 175.
[13]http://archive.indianexpress.com/news/-50-million-people-in-india-have-diabetes-/1030869
[14] Quinlan, J.R., (1986). Induction of decision trees, Machine Learning 1, 81–106.
[15] Quinlan, J.R., (1993). C4.5: Programs for Machine Learning, Morgan Kaufmann, San Francisco.
[16] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J., (1984). Classification and Regression Trees. Wadsworth, Belmont,CA.
[17] P.C. Thirumal and 2N.Nagarajan " Applying Average K Nearest Neighbour Algorithm to Detect Type-2 Diabetes", Australian Journal of Basic and Applied Sciences, 8(7) May 2014, Pages: 128-134
[18] Tom Owen, George Buchanan, Harold Thimbleby, "Understanding User Requirements in Take-Home Diabetes Management Technologies", Published by BISL, BCS HCI 2012.
19] Hudson, D. L. Medical Expert Systems. Encyclopedia of Biomedical Engineering, John Wiley and Sons, 2006.
[20] Bezdek, "Fuzzy Mathematics in Pattern Classification", Ph.D.,thesis, Applied Mathematics, Centre, Cornell University, Ithaca, 1973.
[21] DongSooS.kim, James P.Walsh, "Modelling Irregular Samples forAnalyzing the Risk of Complications of Diabetes Mellitus",ICUIMC'12, 2012.