PREDICTION OF DIABETES USING MACHINE LEARNING

1) Introduction:

Diabetes mellitus has a direct signal of high blood sugar, together with some symptoms including frequent urination, increased thirst, increased hunger and weight loss. Patient of diabetes usually need constant treatment, otherwise, it will possibly lead to many dangerous life-threatening complications. The diabetes is diagnosed with the 2-hour post-load plasma glucose being at least 200mg/dL ,and the necessity of identifying diabetes timely calls in various studies about diabetes recognition.

Many previous research studies have been done about machine learning in diabetes identification. Research has been done focused on diabetes identification through GDA (Generalized Discriminant Analysis) and SVM (Support Vector Machine) and they obtained some inspiring results. Another research was to do the same thing by GRNN (General Regression Neural Network), which also had a very high accuracy. Comparing to the previous work, we make a more comprehensive study containing a number of common techniques used to diabetes identification, intending to compare their performance and find the best one among them.

Through this experiment, we compare several common and data preprocessors for each of the classifiers we use, and find the best preprocessor respectively. Then we compare these classifiers after we modify the parameters of them to reach their approximate maximum accuracy, and we particularly analyze how to modify the parameters in DNN (Deep Neural Network). At last, we also analyze the relevance of each feature with the classification result, and this will help to modify the data set in future studies.

2) Motivation:

Diabetes mellitus is a common disease of human body caused by a group of metabolic disorders where the sugar levels over a prolonged period is very high. It affects different organs of the human body which thus harm a large number of the body's system, in particular the blood veins and nerves. Early prediction in such disease can be controlled and save human life. One of the biggest causes of death worldwide are diabetes. The detection of diabetes is of great importance, concerning its severe complications. Detection of diabetes in early stages and faster plays vital role in curing diabetes. The early identification of this disease can be achieved by developing machine learning models.

1. CANCER DIAGNOSIS USING DATA MINING TECHNOLOGY

Desc: Cancer is a set of diseases in which some cells of the body grow abnormally. These cells then destroy other surrounding cells and their normal functions. Cancer can spread throughout the human body. Since it is a very treacherous disease its diagnosis is very important. In some forms it spreads within days. So the diagnosis of cancer at early stages is very important. The challenge is to first diagnose the main type and then its subtypes. This research uses data mining classification tools to make a decision support system to identify different types of cancer on the Genes dataset. Data mining technology helps in classifying cancer patients and this technique helps to identify potential cancer patients by simply analyzing the data.

Drawbacks

- This system suitable for cancer prediction, doesn’t suitable for diabetes identification.
- Generates less accurate results.

- Some techniques used here takes more time for prediction.

2. Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases

Desc: This paper presents the overview of machine learning techniques in classification of diabetes and cardiovascular diseases (CVD) using Artificial Neural Networks (ANNs) and Bayesian Networks (BNs). The comparative analysis was performed on selected papers that are published in the period from 2008 to 2017. The most commonly used type of ANN in selected papers is multilayer feedforward neural network with Levenberg-Marquardt learning algorithm. On the other hand, the most commonly used type of BN is Naive Bayesian network which shown the highest accuracy values for classification of diabetes and CVD, 99.51% and 97.92% retrospectively. Moreover, the calculation of mean accuracy of observed networks has shown better results using ANN, which indicates that higher possibility to obtain more accurate results in diabetes and/or CVD classification is when it is applied to ANN.

Drawbacks

- System used for classification of diabetes and cardiovascular diseases.
- ANN techniques are used.
- Less efficient

3. Association Rule Extraction from Medical Transcripts of Diabetic Patients

Desc: Medical databases serve as rich knowledge sources for effective medical diagnosis. Recent advances in medical technology and extensive usage of electronic medical record systems, helps in massive production of medical text data in hospitals and other health institutions. Most of this text data that contain valuable information are just filed and not utilized to the full extent. Proper usage of medical information can bring about tremendous changes in medical field. This paper present a new method of uncovering valid association rules from medical transcripts. The extracted rules describes association of disease with other diseases, symptoms of a particular disease, medications used for treating diseases, the most prominent age group of patients for developing a particular disease. NLP (Natural Language Processing) tools were combined with data mining

algorithms (Aprior algorithm and FP-Growth algorithm) for the extraction of rules. Interesting rules were selected using the correlation measure, lift.

Drawbacks

- Used to predict the association among different diabetes parameters.
- Not suitable for diabetes disease prediction.

Existing System:

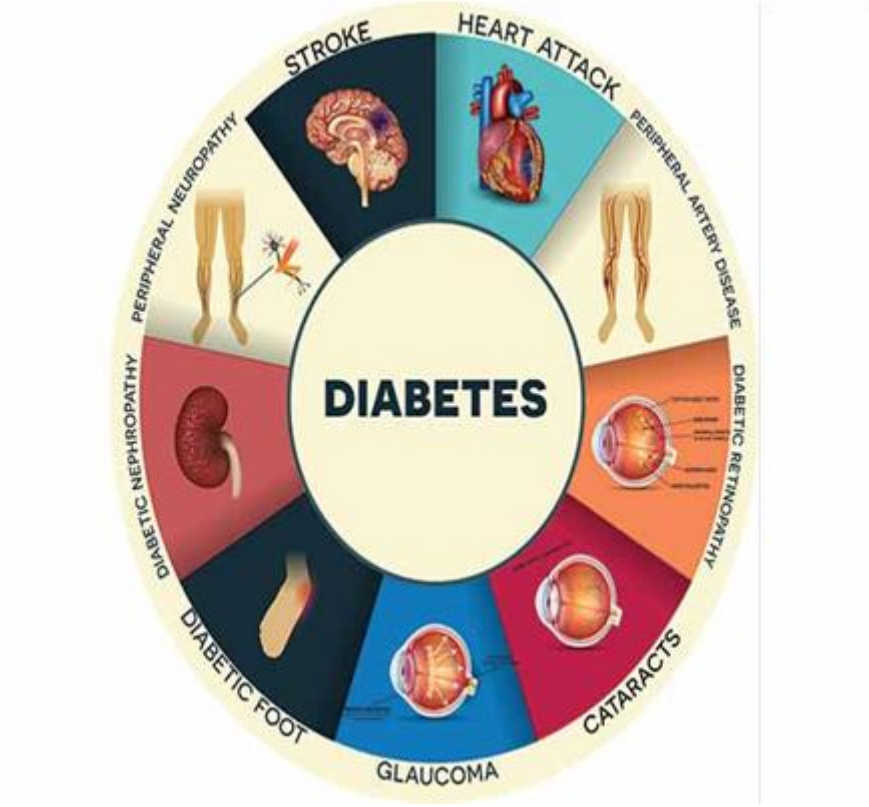
The detection of diabetes is of great importance, concerning its severe complications. Current system is a manual process where the concerned doctor analyzes patients reports manually and it requires more time for the diabetes identification. Home glucose monitoring also done by the patients using the glucose monitoring device (mentioned in the below picture) to detect the diabetes. There have been plenty of research studies about diabetes identification, many of which are based on the Pima Indian diabetes data set. Most of the research studies done before mainly focused on one or two particular complex technique to test the data, while a comprehensive research over many common techniques is missing.

Limitations of Existing System

- Manual process (done by doctor).
- Manual analysis of test reports.
- Glucose monitoring device such as onetouch.
- Few techniques used for diabetes prediction.
- Less accurate results.
- Time consuming.

Proposed Work:

The detection of diabetes is of great importance, concerning its severe complications. The diabetes complications showed in the below picture. Detection of diabetes in early stages and faster plays vital role in curing diabetes.



Diabetes mellitus has a direct signal of high blood sugar, together with some symptoms including frequent urination, increased thirst, increased hunger and weight loss. Patient of diabetes usually need constant treatment, otherwise, it will possibly lead to many dangerous life-threatening complications. Detection of diabetes in early stages and faster plays vital role in curing diabetes. Proposed system is an automation for diabetes

identification using the old diabetes patients data. Proposed system is a medical sector application which is useful to physicians (diabetic doctors) in identifying the disease. Proposed system uses machine learning techniques for diabetes identification.

Input and Output

Input: Old diabetes Patients data such as blood test reports, age, DOB, weight, BP, Height etc

Output: Diabetes identification for new patient based on the medical attributes.

Dataset:

We choose dataset of pima Indian women about their diabetes recognition It is great dataset to study due to high onset rate in this area which means data here will reveal more features or symptoms of diabetes than other data set. The issue has been long time study since 1965 by National Institute of Diabetes and Digestive of Kidney Disease

With 8 particular Features choosen to be studied .Eight features are listed below:

- Number of times Pregnant
- Plasma Glucose Concentration a 2 hours in a oral Glucose Tollerance Test
- Diastolic Blood Preasure
- Triceps Skin Fold Thickness
- 2 hour serum Insulin(mu U/ml)
- Body Mass Index (Weight in kg)
- Diabetes Pedigree function
- Age(Years)

Objectives:

- Proposed system is a medical application used by diabetes doctors (physicians).
- Proposed system is a medical software for diabetes identification.
- Proposed system uses machine learning techniques for diabetes identification.
- Proposed system uses old diabetes patients data for diabetes identification of the new patient.

- Proposed aims at diabetes identification based on the attributes such as blood test reports, age, DOB, weight, BP, Height etc…
- Proposed system is a real time application which uses ASP.NET as front technology and SQL Server as backend technology.

Literature Survey:

1) Safavian S. R. and Landgrebe D., "A survey of decision tree classifier methodology," IEEE transactions on systems, man, and cybernetics, vol. 21. 3, 1991, pp. 660-674.

2) Kayaer K and Yildirim T, "Medical diagnosis on Pima Indian diabetes using general regression neural networks," Proceedings of the international conference on artificial neural networks and neural information processing, 2003, pp. 181-184.

3) Principles of Internal medicine by Harrison (Vol 1 and Vol2)

4) Lee H., "Tutorial on deep learning and applications," NIPS 2010 Workshop on Deep Learning and Unsupervised Feature Learning, 2010.

5) Lin Y., "Support vector machines and the Bayes rule in classification," Data Mining and Knowledge Discovery, vol. 6. 3, 2002, pp. 259-275

6) Kemal Polat, Salih Gunes, and Ahmet Arslan, "A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine," Expert Systems with Applications, vol. 34. 1, January. 2008, pp. 482-487.

7) Carpenter G. A. and Markuzon N., "ARTMAP-IC and medical diagnosis: Instance counting and inconsistent cases," Neural Networks, vol. 11. 2, 1998, pp. 323-336.

8) Kemal Polat, Salih Gunes, and Ahmet Arslan, "A cascade learning system for classification of diabetes disease: Generalized Discriminant Analysis and Least Square Support Vector Machine," Expert Systems with Applications, vol. 34. 1, January. 2008, pp. 482-487.