International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 5 May 2021, pp: 939-943 www.ijaem.net ISSN: 2395-5252

Diabetes Mellitus Prediction Using Back Propogation Neural Network Algorithm

S.R. Sowmiya. M.E.,[1] Assistant Professor, B. Parkavi,[2] S. Pavithra,[3] T.Pavithra,[4] R. Pooja[5]
Dhanalakshmi Srinivasan Engineering College (Autonomous), Perambalur.

Submitted: 05-05-2021 Revised: 17-05-2021 Accepted: 20-05-2021

ABSTRACT: Diabetes is a metabolic disease affecting a multitude of people worldwide. Its incidence rates are increasing alarmingly every year. If untreated, diabetes-related complications in many vital organs of the body may turn fatal. Early detection of diabetes is very important for timely treatment which can stop the disease progressing to such complications. An intelligent predictive model using deep learning is proposed to predict the patient risk factor and severity of diabetics using conditional data set. The model involves deep learning in the form of a deep neural network which helps to apply predictive analytics on the diabetes data set to obtain optimal results. The existing predictive models are used to predict the disease whether it is normal or not based on the data which is processed. In this project firstly, a feature selection algorithm is run for the selection process. Secondly, the deep learning model has a deep neural network which employs back propagation neural network as a basic unit to analyses the data by assigning weights to the each branch of the neural network. This deep neural network, coded on python, will help to obtain numeric results on the severity and the risk factor of the diabetics in the data set. At the end, can provide prescription based results for disease diagnosis based on Pima Indians diabetes. This will help to predict diabetes with much more precision as shown by the results obtained.

I. INTRODUCTION

BIG DATA

Big data is high-volume, high-velocity and/or high-variety information assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and process automation.

BIG DATA CHARACTERSTICS

Big data refers to massive complex structured and unstructured data sets that are rapidly generated and transmitted from a wide variety of sources. These attributes make up the three Vs of big data:

1. **Volume:** The huge amounts of data being stored.
2. **Velocity:** The lightning speed at which data streams must be processed and analyzed.
3. **Variety:** The different sources and forms from which data is collected, such as numbers, text, video, images, audio and text.

Big data is essentially the wrangling of the three Vs to gain insights and make predictions, so it's useful to take a closer look at each attribute.

Volume: Big data is enormous. While traditional data is measured in familiar sizes like megabytes, gigabytes and terabytes, big data is stored in petabytes and zettabytes. To grasp the enormity of difference in scale, consider this comparison from the Berkeley School of Information: one gigabyte is the equivalent of a seven minute video in HD, while a single zettabyte is equal to 250 billion DVDs. This is just the tip of the iceberg. According to a report by EMC, the digital universe is doubling in size every two years and by 2020 is expected to reach 44 trillion zettabytes. Big data provides the architecture handling this kind of data. Without the appropriate solutions for storing and processing, it would be impossible to mine for insights.

II. LITERATURE SURVEY

TITLE: DIABETES IN DEVELOPING COUNTRIES, AUTHOR: ANOOP MISRA

There has been a rapid escalation of type 2 diabetes (T2D) in developing countries, with varied prevalence according to rural vs urban habitat and degree of urbanization. Some ethnic groups (eg, South Asians, other Asians, and Africans), develop diabetes a decade earlier and at a lower body mass index than Whites, have prominent abdominal obesity, and accelerated the conversion from prediabetes to diabetes. The burden of complications, both macro- and micro vascular, is substantial, but also varies according to

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 5 May 2021, pp: 939-943 www.ijaem.net ISSN: 2395-5252

populations. The syndemics of diabetes with HIV or tuberculosis are prevalent in many developing countries and predispose to each other. Screening for diabetes in large populations living in diverse habitats may not be cost-effective, but targeted high-risk screening may have a place. The cost of diagnostic tests and scarcity of health manpower pose substantial hurdles in the diagnosis and monitoring of patients. Efforts for prevention remain rudimentary in most developing countries. The quality of care is largely poor; hence, a substantial number of patients do not achieve treatment goals. This is further amplified by a delay in seeking treatment, "fatalistic attitudes", high cost and non-availability of drugs and insulins. To counter these numerous challenges, a renewed political commitment and mandate for health promotion and disease prevention are urgently needed.

ADVANTAGES

- Increasing awareness of diabetics

DISADVANTAGES

- Difficult to handle large datasets

Title: Genetic Algorithm Based Feature Selection And Moe Fuzzy Classification Algorithm On Pima Indians Diabetes Dataset, Author: Vaishali R

Diabetes mellitus is a condition of chronic hyperglycemia characterized by the increased levels of glucose in blood due to defects in the secretion of insulin from the Pancreatic Beta cells. The adverse effects of Type 2 diabetes include malfunctioning of organs with permanent damage. The long term effects of diabetes may result in coma, renal failure and retinal failure, pathological destruction of pancreatic beta cells, cardiovascular dysfunction, cerebral vascular dysfunction, peripheral vascular diseases, sexual dysfunction, joint failure, weight loss, ulcer and pathogenic effects on immunity. The reduction in the amounts of insulin causes abnormalities in the levels of carbohydrates and proteins. Diabetes Mellitus is a dreadful disease characterized by increased levels of glucose in the blood, termed as the condition of hyperglycemia. As this disease is prominent among the tropical countries like India, an intense research is being carried out to deliver a machine learning model that could learn from previous patient records in order to deliver smart diagnosis. This research work aims to improve the accuracy of existing diagnostic methods for the prediction of Type 2 Diabetes with machine learning algorithms. The proposed algorithm selects the essential features from the Pima Indians Diabetes Dataset with Goldberg's Genetic algorithm in the pre-processing stage and a Multi Objective Evolutionary Fuzzy Classifier is applied on the dataset. Dimensionality is a curse to machine learning'. Medical datasets are often larger in dimensions with complex redundant features. The redundancy of features increases the possibility of noise and dependency among the features.

ADVANTAGES

- Reduces redundancy data

DISADVANTAGES

- Missing data on feature selection

III. SYSTEM ANALYSIS

EXISTING SYSTEM

Present days one of the major application areas of machine learning algorithms is medical diagnosis of diseases and treatment. Machine learning algorithms also used to find correlations and associations between different diseases. Nowadays many people are dying because of diabetics. Prediction and diagnosing of diabetic disease becomes a challenging factor faced by doctors and hospitals both in India and abroad. In order to reduce number of deaths because of diabetic diseases, we have to predict whether person is at the risk of diabetic disease or not in advance. Data mining techniques and machine learning algorithms play a very important role in this area. In this existing system, focused on how data mining techniques can be used to predict diabetic disease in advance such that patient is well treated. An important task of any diagnostic system is the process of attempting to determine and/or identify a possible disease or disorder and the decision reached by this process. For this purpose, machine learning algorithms are widely employed. For these machine learning techniques to be useful in medical diagnostic problems, they must be characterized by high performance, the ability to deal with missing data and with noisy data, the transparency of diagnostic knowledge, and the ability to explain decisions. As people are generating more data everyday so there is a need for such a classifier which can classify those newly generated data accurately and efficiently. This System mainly focuses on the supervised learning technique called the Random forests for classification of data by changing the values of different hyper parameters in Random Forests Classifier and also implement support vector machine algorithm to classify the disease whether it is appear or not.

DISADVANTAGES

- Labeled data based disease classification

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 5 May 2021, pp: 939-943 www.ijaem.net ISSN: 2395-5252

- Provide high number of false positive
- Binary classification can be occurred
- Computational complexity

IV. PROPOSED SYSTEM

Multiple opportunities for healthcare are created because machine learning models have potential for advanced predictive analytics. Diabetes is a disease which reduces the body's capability to produce insulin. In other words the body cannot retaliate to the hormone insulin production. This results in anomalous metabolism of carbohydrates and increased blood glucose levels. Early detection of diabetes becomes very important. The blood glucose levels become too high in the body when there is diabetes. Glucose is created in the body after eating food. The hormone insulin produced in the body helps balance the glucose levels and regulate blood sugar levels, deficiency of insulin causes Diabetes. Type 1 diabetes is a scenario where the body does not produce insulin at all to balance the sugar levels in blood. Type 2 is a diabetes type where the body produces insulin but does not utilize this hormone completely to balance blood sugar levels. The Type 2 diabetes is most common one. There is something called as prediabetes, this is a situation where the person can have high glucose level but not that high that he/she can be said to have diabetes. But the people who have prediabetes are prone to get type 2 diabetes. This disease can cause serious damage to many vital organs in the body like kidneys, heart, nerves and eyes. If a woman gets this disease during pregnancy then it is known as gestational diabetes. So implement deep learning based neural network algorithm can be used to predict the diabetic diseases with improved accuracy. Neural Networks has emerged as an important method of classification. Back-propagation has been employed as the training algorithm in this work. This project proposes a diagnostic system for predicting heart disease with improved accuracy. The propagation algorithm has been repeated until minimum error rate was observed. And also provide the diagnosis information to patients through SMS alert and also provide voice information to patients based on trained diabetic datasets.

ADVANTAGES

- Accuracy is high
- Parallel processing
- Multiple diabetic diseases are predicted
- Reduce number of false positive rate

V. SYSTEM IMPLEMENTATION

MODULES DESCRIPTION

1. DATASETS ACQUISITION

A data set (or dataset, although this spelling is not present in many contemporary dictionaries like Merriam-Webster) is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows. The term data set may also be used more loosely, to refer to the data in a collection of closely related tables, corresponding to a particular experiment or event. In this module, we can upload the cardiovascular datasets related to diabetic diseases which includes the attributes such as glucose, insulin level, blood pressure, BMI and so on

2. PREPROCESSING

Data pre-processing is an important step in the [data mining] process. The phrase "garbage in, garbage out" is particularly applicable to data mining and machine learning projects. Data-gathering methods are often loosely controlled, resulting in out-of-range values, impossible data combinations, missing values, etc. Analyzing data that has not been carefully screened for such problems can produce misleading results. Thus, the representation and quality of data is first and foremost before running an analysis. If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. In this module, we can eliminate the irrelevant values and also estimate the missing values of data. Finally provide structured datasets.

3. FEATURES SELECTION

Feature selection refers to the process of reducing the inputs for processing and analysis, or of finding the most meaningful inputs. A related term, feature engineering (or feature extraction), refers to the process of extracting useful information or features from existing data. Filter feature selection methods apply a statistical measure to assign a scoring to each feature. The features are ranked by the score and either selected

International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 5 May 2021, pp: 939-943 www.ijaem.net ISSN: 2395-5252

to be kept or removed from the dataset. The methods are often uni-variate and consider the feature independently, or with regard to the dependent variable. It can be used to construct the multiple diabetic diseases. In this module, select the multiple features from uploaded datasets. And train the datasets with various disease labels such as Type 1 diabetics with diagnosis information. Type 2 diabetics with diagnosis information

4. CLASSIFICATION

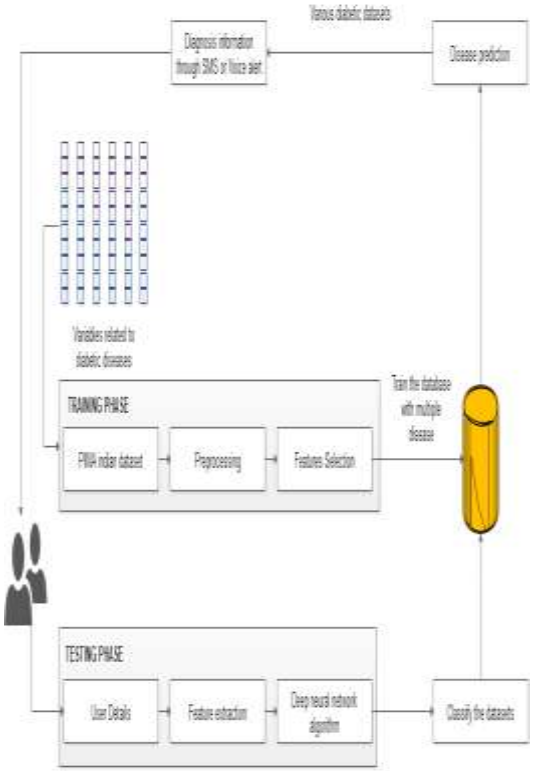
In this module implement classification algorithm to predict the diabetic diseases. And using deep learning algorithm such as back propagation algorithm to predict the diseases.A back propagation is a feed forward artificial neural network model that maps sets of input data onto a set of appropriate outputs. It consists of multiple layers of nodes in a directed graph, and each layer is fully connected to the next one. Each node is a neuron with a nonlinear activation function except for the input nodes. Back propagation utilizes a supervised learning technique called back propagation for training the network. Back propagation is a modified form of the standard linear perceptron and can distinguish data that are not linearly separable. If back propagation has a simple on-off mechanism i.e. linear activation function in all neurons ,to determine whether or not a neuron fires, then it is easily proved with linear algebra that any number of layers can be reduced to the standard two-layer input-output model. The gradient techniques are then applied to the optimization methods to adjust the weights to minimize the loss function in the network. Hence, the algorithm requires a known and a desired output for all inputs in order to compute the gradient of loss function. Usually, the generalization of back propagation Feed Forward Networks is done using delta rule which possibly makes a chain of iterative rules to compute gradients for each layer. Back Propagation Algorithm necessitates the activation function to be different between the neurons. The ongoing researches on parallel, distributed computing and computational neuroscience are currently implemented with the concepts of Back Propagation Algorithm. Back Propagation Algorithm has also gained focus in pattern recognition domain. They are so convenient in research, because of their ability in solving complex problems, and also for their fitness approximation results even with critical predictions. Back propagation is one of the Neural Network models, has the same architecture of Feed-Forward back Propagation for Supervised training. The back propagation is the most known and most frequently used type of neural network. User can provide the features and automatically predict the diseases.

5. DISEASE DIAGNOSIS

Medical decision support system is a decision-support program which is designed to assist physicians and other health professionals with decision making tasks, such as determining diagnosis of patients' data. In this module, provide the diagnosis information based on predicted diabetic diseases. Proposed system provides improved accuracy in diabetic disease prediction. Risk factors are conditions or habits that make a person more likely to develop a disease. In this module, provide the diagnosis information based on predicted diabetics diseases. Information sends to user in the form of SMS or Voice information. Proposed system provide improved accuracy in heart disease prediction

VI. SYSTEMDESIGN

SYSTEM ARCHITECTURE



VII. SOFTWARE DESCRIPTION

FRONT END: PYTHON

Python is an interpreted high-level programming language for general-purpose programming. Created by Guido van Rossum and

first released in 1991, Python has a design philosophy that emphasizes code readability, notably using significant whitespace. It provides constructs that enable clear programming on both small and large scales. In July 2018, Van Rossum stepped down as the leader in the language community. Python features a dynamic type system and automatic memory management. It supports multiple programming paradigms, including object-oriented, imperative, functional and procedural, and has a large and comprehensive standard library. Python interpreters are available for many operating systems. CPython, the reference implementation of Python, is open source software and has a community-based development model, as do nearly all of Python's other implementations. Python and CPython are managed by the non-profit Python Software Foundation. Rather than having all of its functionality built into its core, Python was designed to be highly extensible.

VIII. SYSTEM TESTING

TESTING PROCESS

Software testing is a method of assessing the functionality of a software program. There are many different types of software testing but the two main categories are dynamic testing and static testing. Dynamic testing is an assessment that is conducted while the program is executed; static testing, on the other hand, is an examination of the program's code and associated documentation. Dynamic and static methods are often used together. Testing is a set activity that can be planned and conducted systematically. Testing begins at the module level and work towards the integration of entire computers based system. Nothing is complete without testing, as it is vital success of the system.

Testing Objectives:

There are several rules that can serve as testing objectives, they are;

1. Testing is a process of executing a program with the intent of finding an error
2. A good test case is one that has high probability of finding an undiscovered error.
3. A successful test is one that uncovers an undiscovered error.

IX. CONCLUSION AND FUTURE ENHANCEMENT

CONCLUSION

In this project the problem of constraining and summarizing different algorithms of data mining used in the field of medical prediction are discussed. The focus is on using different algorithms and combinations of several target attributes for intelligent and effective diabetic disease prediction using data mining. Data mining technology provides an important means for extracting valuable medical rules hidden in medical data and acts as an important role in disease prediction and clinical diagnosis. There is an increasing interest in using classification to identify disease which is present or not. In the current study, have demonstrated, using a large sample of patients hospitalized with classification. Classification algorithm is very sensitive to noisy data. If any noisy data is present then it causes very serious problems regarding to the processing power of classification. It not only slows down the task of classification algorithm but also degrades its performance.

FUTURE ENHANCEMENTS

In future we tend to improve efficiency of performance by applying other data mining techniques and algorithms.

REFERENCES

[1]. A. Misra, H. Gopalan, R. Jayawardena, A. P. Hills, M. Soares, A. A. Reza-Albarrán, and K. L. Ramaiya, ‘‘Diabetes in developing countries,’’ J. Diabetes, vol. 11, no. 7, pp. 522–539, Mar. 2019.

[2]. R. Vaishali, R. Sasikala, S. Ramasubbareddy, S. Remya, and S. Nalluri, ‘‘Genetic algorithm based feature selection and MOE fuzzy classification algorithm on pimaindians diabetes dataset,’’ in Proc. Int. Conf. Comput. Netw.Informat. (ICCNI), Oct. 2017, pp. 1–5.

[3]. The Emerging Risk Factors Collaboration, ‘‘Diabetes mellitus, fasting blood glucose concentration, and risk of vascular disease: A collaborative meta-analysis of 102 prospective studies,’’ Lancet, vol. 375, pp. 2215–2222, Jun. 2010.

[4]. N. H. Cho, J. E. Shaw, S. Karuranga, Y. Huang, J. D. da Rocha Fernandes, A. W. Ohlrogge, and B. Malanda, ‘‘IDF diabetes atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045,’’ Diabetes Res. Clin. Pract., vol. 138, pp. 271–281, Apr. 2018

[5]. M. Maniruzzaman, M. J. Rahman, M. Al-MehediHasan, H. S. Suri, M. M. Abedin, A. El-Baz, and J. S. Suri, ‘‘Accurate diabetes risk stratification using machine learning: Role of missing value and outliers,’’ J. Med. Syst., vol. 42, no. 5, p. 92, May 2018.