Prediction of Diabetes using Artificial Neural Network Classification Technique

[1] Dr. M.Manimekalai, [2] S. Divya

[1]Professor, Director and Head, [2]Research Scholar
[1,2]Department of Computer Science
[1,2]Shrimati Indira Gandhi College, Tiruchirappalli, Tamil Nadu, India-620002

Abstract : Diabetes is one of the major diseases of the population across the world. Diabetes is a chronic disease that occurs either when the pancreas does not produce enough insulin or when the body cannot efficiently use the insulin it produces. In 2014, 8.5% of adults aged 18 years and older had diabetes. In 2012, diabetes was the direct cause of 1.5 million deaths and high blood glucose was the cause of another 2.2 million deaths. Over the time, diabetes can damage the heart, blood vessels, eyes, kidneys, and nerves. Early diagnosis can be made through a relatively inexpensive method of computation. In this paper, Machine Learning, a branch of Artificial Intelligence is used to analyze and make the diabetes prediction model. Various researchers have also been done to predict the diabetes machine learning algorithm, but this is an additional effort in the research work based on a specific type of patient in a specific community. In this research work, a data sample of Pima Indians was taken to predict the possibility of diabetes. Among several algorithms of Machine learning, Artificial Neural Network (ANN) was chosen for building the model to predict diabetes. This model is ideal for predicting the possibility of diabetes with 92% accuracy while tested with the sample test data. This model can achieve more accuracy if it is trained with a large sample training data in the future**.**

***IndexTerms* - Diabetes, Data Mining Techniques, Feature Selection, Classification, Artificial Neural Networks, Machine Learning.**

I. INTRODUCTION

Diabetes is a dangerous disease and greatly affects human life periods. Diabetes is transmittable from mothers to unborn children. Its effects include premature Death, Strokes, Heart diseases, blindness, and kidney failure. This objective of this paper is to propose an easy technique for Prediction of Diabetic patients. A diabetic person is identified with low levels of blood sugars and insulin in their body. Diabetes can be classified into three types namely Type 1, Type 2 and Gestational [1]. Type 1 – Diabetes was previously known as "Insulin-Dependent Diabetes Mellitus". Diabetes can be formed at any age, but needs to be diagnosed below 20 years. This type of diabetes is formed insulin producing cells or beta cells in the pancreas get destroyed Type 2 - Diabetes was previously known as non–insulin-dependent diabetes as it was diagnosed in patients above 20 years of age. Gestational – Diabetes, can occur in pregnant women, when pancreas does not create required amount of insulin in the body. These three types of diabetes need treatment and when detected and treated early, complications associated with them can be avoided [2]. Data mining is a powerful tool for data analysis in its process of discovering interesting pattern from huge amounts of data like massive datasets or data warehouses. This work uses data mining techniques for the prediction of the above listed diseases in patients' databases [3].

Diabetes Mellitus (DM) [4] is a set of related diseases in which the body cannot regulate the amount of sugar in the blood. In a healthy person, the blood glucose level is regulated by several hormones, including insulin. Insulin is produced by the pancreas, a small organ between the stomach and liver. The pancreas secretes other important enzymes that help to digest food. Insulin allows glucose to move from the blood into liver, muscle, and fat cells, where it is used for fuel.

Hereditary and genetics factors, Infections caused by viruses, Stress, Obesity, Increased cholesterol level, High carbohydrate diet, Nutritional deficiency, Excess intake of oil and sugar No physical exercise, Overeating, Tension and worries, High blood pressure, Insulin deficiency, Insulin resistance [5].

II. RELATED WORKS

Dagliati, Arianna, et al [6] One of the areas where Artificial Intelligence is having more impact is machine learning, which develops algorithms able to learn patterns and decision rules from data. Machine learning algorithms have been embedded into data mining pipelines, which can combine them with classical statistical strategies, to extract knowledge from data. Within the EU-funded MOSAIC project, a data mining pipeline has been used to derive a set of predictive models of (T2DM) complications based on electronic health record data of nearly one thousand patients. Such pipeline comprises clinical center profiling, predictive model targeting, predictive model construction and model validation. After having dealt with missing data by means of random forest (RF) and having applied suitable strategies to handle class imbalance, we have used Logistic Regression with stepwise feature selection to predict the onset of retinopathy, neuropathy, or nephropathy, at different time scenarios, at 3, 5, and 7 years from the first visit at the Hospital Center for Diabetes (not from the diagnosis). Considered variables are gender, age, time from diagnosis, body mass index (BMI), glycated hemoglobin (HbA1c), hypertension, and smoking habit. Final models, tailored in accordance with the complications, provided an accuracy up to 0.838. Different variables were selected for each complication and time scenario, leading to specialized models easy to translate to the clinical practice.

Shaik, Ashraf Ali, Ch Prathima, and Naresh Babu Muppalaneni [7] Making use of estimating methods in the field of medicine has been the powerful research recently. Diabetic retinopathy is a retinal disease which causes huge blindness. Recurrent screening for prior disease detection has been a highly labor force—and resource—powerful process. So computerized diagnosis of these diseases through estimating methods would be a great remedy. Through this paper, a novel estimation strategy for computerized disease prognosis is suggested, which utilizes retinal image analysis and mining methods to accurately differentiate between the retinal images as normal and affected. Eighteen feature relevance and three variations algorithms were analyzed and used to identify the contributing features that provided better conjecture results.

Das, Himansu, Bighnaraj Naik, and H. S. Behera [8] The aim of this research is to predict diabetes based on some of the DM techniques like classification and clustering. Out of which, classification is one of the most suitable methods for predicting diabetes.

In this study, J48 and Naïve Bayesian techniques are used for the early detection of diabetes. This research will help to propose a quicker and more efficient technique for diagnosis of disease, leading to timely and proper treatment of patients. We have also proposed a model and elaborated it step-by-step, in order to make medical practitioner to explore and to understand the discovered rules better. The study also shows the algorithm generated on the dataset collected from college medical hospital as well as from online repository. In the end, an article also outlines how an intelligent diagnostic system works. A clinical trial of this proposed method involves local patients, which is still continuing and requires longer research and experimentation.

Singh, Pankaj Pratap, et al [9] Health care data are often huge and complex because it contains different attributes and also missing some values. Data mining techniques can be used to extract knowledge by constructing models from data such as diabetic patient datasets. This research aims at finding solutions to diagnose the disease by analyzing the patterns found in the dataset through data mining. In addition, the neural network approach is also used for classifying the existing diabetic patient data for predicting the patient's disease based on the trained data that can lead to find the different level of diabetes in the patients. It is also compared with the association rule mining based approach for classification of data to validate the correctly classified cases.

Husain, Adil, and Muneeb H. Khan [10] An Ensemble model using majority voting technique was developed by combining the unweighted prediction probabilities of different machine learning models. Also, the model is evaluated and validated for real user input data for user friendliness. The overall performance was improved by Ensemble Model and had an AUC (Area under Curve) of 0.75 indicating high performance.

III. PROBLEM IDENTIFICATION

Nowadays, diabetes has become a common disease to the mankind from young to the old persons. The growth of the diabetic patients is increasing day-by-day due to various causes such as bacterial or viral infection, toxic or chemical contents mix with the food, auto immune reaction, obesity, bad diet, change in lifestyles, eating habit, environment pollution, etc. Hence, diagnosing the diabetes is very essential to save the human life from diabetes. The data analytics is a process of examining and identifying the hidden patterns from large amount of data to draw conclusions.

IV. RESEARCH METHODOLOGY

In this research work, Machine Learning, a branch of Artificial Intelligence is used to analyze and make the diabetes prediction model. Various researchers have also been done to predict the diabetes machine learning algorithm, but this is an additional effort in the research work based on a specific type of patient in a specific community. In this research work, a data sample of Pima Indians was taken to predict the possibility of diabetes. Among several algorithms of Machine learning, Artificial Neural Network (ANN) was chosen for building the model to predict diabetes. This model is ideal for predicting the possibility of diabetes with 92% accuracy while tested with the sample test data. This model can achieve more accuracy if it is trained with a large sample training data in the future.

4.1 Chi-Square Feature Selection Technique

Feature selection is an example of the common prominent and frequent techniques in data pre-processing and has converted a crucial part of the machine learning method it has also distinguished as attribute selection, variable subset or variable selection in statistics and machine learning. It is the process of identifying relevant and eliminating irrelevant features, redundant or noisy data. This method rushes up data mining algorithms, improves predictive accuracy and understandability. Irrelevant features are those that provide no valuable knowledge, and irrelevant features present no further information than the presently selected features. Chi-Square analysis has used in this project work to carry out the pre-processing step [11][12] [14][15].

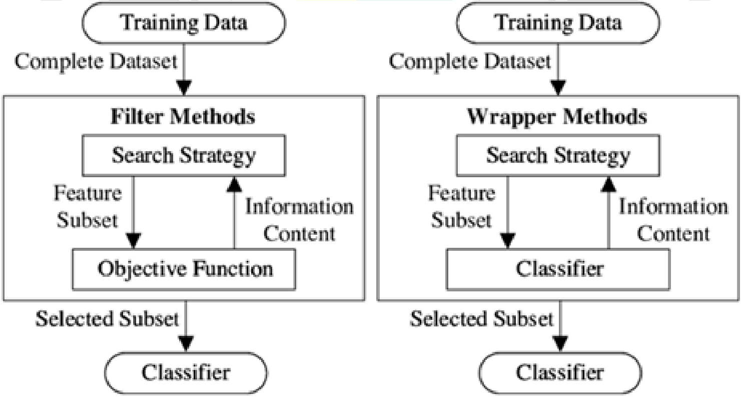

Figure 1: Types of Feature Selection Techniques

4.2 Artificial Neural Network Classification Technique

Artificial Neural Network (ANN) is an efficient computing system whose central theme is borrowed from the analogy of biological neural networks. ANNs are also named as "artificial neural systems," or "parallel distributed processing systems," or "connectionist systems." ANN acquires a large collection of units that are interconnected in some pattern to allow communication between the units. In order to form a feed-forward multi-layer in MLP, the collection of non-linear neurons is connected to one another. This technique is known to be very useful for prediction and classification issues. Cross-validation is used to determine the 'optimal' number of hidden layers and neurons which were relied on the experimental design of the Intrusion Detection classification framework. These units, also referred to as nodes or neurons, are simple processors which operate in parallel [13] [16][17][18].

4.3 Naïve Bayes Classification Technique

In statistics, the correlation coefficient indicates the strength and direction of a relationship between two random variables. The commonest use refers to a linear relationship. In general, statistical usage, correlation or correlation refers to the departure of two random variables from independence [16][17][18].

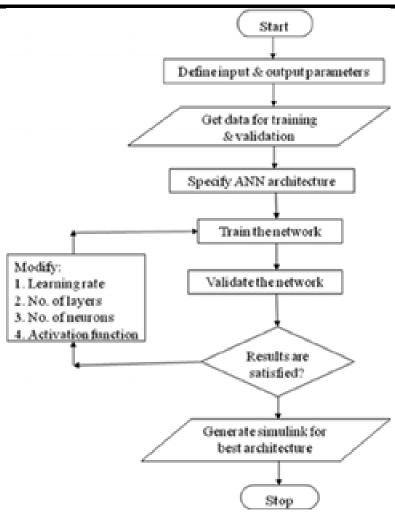

Figure 2: Flowchart of Artificial Neural Network

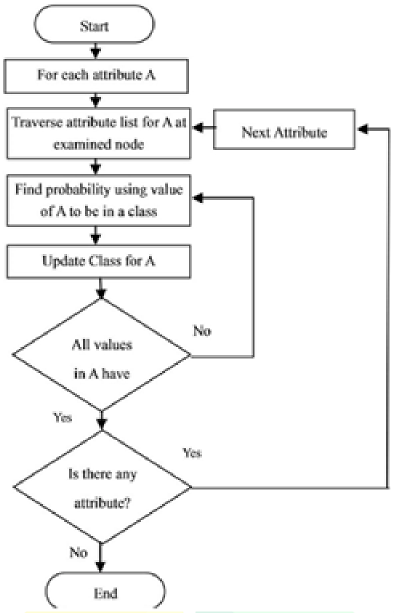

Figure 3: Flowchart for Naïve Bayes Classification Technique

IV. RESULTS AND DISCUSSION

5.1 Description of the Dataset

Table 1 depicts the description of the diabetes dataset used in this research work.

Table 1: Features given in Diabetes dataset

Feature index	Feature name
1	Regular insulin dose
2	NPH insulin dose
3	UltraLente insulin dose
4	Unspecified blood glucose measurement
5	Unspecified blood glucose measurement
6	Pre-breakfast blood glucose measurement
7	Post-breakfast blood glucose measurement
8	Pre-lunch blood glucose measurement
9	Post-lunch blood glucose measurement
10	Pre-supper blood glucose measurement
11	Post-supper blood glucose measurement
12	Pre-snack blood glucose measurement
13	Hypoglycemic symptoms
14	Typical meal ingestion
15	More-than-usual meal ingestion
16	Less-than-usual meal ingestion
17	Typical exercise activity
18	More-than-usual exercise activity
19	Less-than-usual exercise activity
20	Unspecified special event

Table 2a gives the classification accuracy and error rate analysis of the feature selection method and original dataset using Naïve Bayes classifier. From table 2a proposed hybrid Feature Selector gives lesser classification accuracy, kappa statistic value. Error rates are reduced for original dataset itself than the existing methods. Table 2b, Table 2c depicts the detailed accuracy of the diabetes dataset using NB classifier. The detailed accuracy like True Positive Rate, False Positive Rate, F Measure and ROC values are higher for the proposed feature selector method than the existing methods.

Table 2a: Classification Accuracy of SU, PSO and Hybrid Feature Selector using Naïve Bayes Classification Method for Diabetes Dataset

Dataset Name	Naïve Bayes Classification Method	
	Original	Feature Selection
Classification Accuracy	70%	67.2727 %
Kappa Statistics	0.3483	0.1915
Mean absolute error	0.3528	0.4378
Root mean squared error	0.4548	0.4655
Relative absolute error	76.9231 %	95.4548 %
Root relative squared error	95.0277 %	97.2659 %

Table 2b: Detailed Naïve Bayes Accuracy by Class for Original Diabetes dataset

Class	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
0	0.761	0.41	0.771	0.761	0.766	0.746
1	0.59	0.239	0.575	0.59	0.582	0.746
Weighted Average	0.7	0.35	0.702	0.7	0.701	0.746

Table 2c: Detailed Naïve Bayes Accuracy by Class for Feature Selection Processed Diabetes dataset

Class	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
0	0.887	0.718	0.692	0.887	0.778	0.635
1	0.282	0.113	0.579	0.282	0.379	0.637
Weighted Average	0.673	0.503	0.652	0.673	0.637	0.636

Table 3a gives the classification accuracy and error rate analysis of the Feature Selector, and original dataset using ANN classifier. From table 3a proposed Feature Selector gives higher classification accuracy, kappa statistic value. Error rates are reduced for proposed method than the existing methods. Table 3b , Table 3C depicts the detailed accuracy of the dermatology dataset using ANN classifier.

Table 3a: Classification Accuracy of original dataset and Feature Selector using ANN Classification Method for Diabetes Dataset

Dataset Name	ANN Classification Method	
	Original	Feature Selection
Classification Accuracy	47.8142 %	90.9836 %
Kappa Statistics	0.2841	0.7225
Mean absolute error	0.2898	0.056
Root mean squared error	0.4402	0.1872
Relative absolute error	78.362 %	32.8346 %
Root relative squared error	102.3606 %	64.5845 %

Table 3b: Detailed ANN Classifier Accuracy by Class for Original Diabetes dataset

Class	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
0	0.686	0.194	0.628	0.686	0.656	0.767
1	0.222	0.109	0.333	0.222	0.267	0.61
2	0.54	0.308	0.397	0.54	0.458	0.639
3	0.316	0.1	0.453	0.316	0.372	0.748
Weighted Average	0.478	0.189	0.471	0.478	0.466	0.697

Table 3c: Detailed ANN Accuracy by Class for Feature Selection Processed diabetes dataset

Class	**TP Rate**	**FP Rate**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
0	0.997	0.069	0.983	0.997	0.99	0.979
1	0	0	0	0	0	0.372
2	0.884	0.074	0.613	0.884	0.724	0.914
3	0.077	0.012	0.333	0.077	0.125	0.853
Weighted Average	0.91	0.065	0.885	0.91	0.889	0.958

6. CONCLUSION

In this work, feature selector is the goal of reducing redundant and irrelevant features. The data classification has improved by picking only the most relevant features. The performance of the feature selector has estimated concerning three quality criteria such as the number of selected elements, the detection performance of classifiers, and time is taken to build the model with the dataset from University of California Irvine (UCI) diabetes dataset. The proposed system can more explicitly state as follows:

- Feature selector for choosing only most relevant features for supervised. The system has pointed at making enhancements over the present work in three aspects such as the decrease in feature set, increase in classification accuracy, and finally, reducing the running time of reaching the goal.

- The result of feature selector imparts higher classification accuracy rate for some dataset with minimum selected features and minimum running time.
- The proposed features and learning paradigm hybrid feature selector are promising strategies to be applied to any data classification problems.

REFERENCES

[1] Jiawei Han and Micheline Kamber, “Data Mining: Concepts and Techniques”, 2ndedition.Vol.2, No.6, pp..251-261, June 2012.

[2] E.Papagcoriou, I.Kotsioni, A.Lions, “Data Mining: A new Technique in medical research”, Vol.4, No.2, pp.114-118, 2013.

[3] Mohammed J.Zaki, Jason T.L. Wang, Hannu T.T. Toivonen, “Biokdd01: Workshop on data mining in bioinformatics”, Vol.12, No.4, pp 121-134, Feb 2012.

[4] Khalid Raza “Application of data mining of data mining in bioinformatics”, Indian journal of computer science and Engineering, Vol.2 No.1, pp 72-88, Sep2013.

[5] Hasan sai, kuldeep vasnik”Screening for type-2 diabetes: Report of a world health organization and international diabetes”, Vol.5 No.3, pp.114-128, Aug 2014.

[6] Dagliati, Arianna, et al. "Machine learning methods to predict diabetes complications." *Journal of diabetes science and technology* 12.2 (2018): 295-302.

[7] Shaik, Ashraf Ali, Ch Prathima, and Naresh Babu Muppalaneni. "A Computational Approach to Predict Diabetic Retinopathy Through Data Analytics." *Internet of Things and Personalized Healthcare Systems*. Springer, Singapore, 2019. 105-112.

[8] Das, Himansu, Bighnaraj Naik, and H. S. Behera. "Classification of diabetes mellitus disease (DMD): a data mining (DM) approach." *Progress in Computing, Analytics and Networking*. Springer, Singapore, 2018. 539-549.

[9] Singh, Pankaj Pratap, et al. "Classification of Diabetic Patient Data Using Machine Learning Techniques." *Ambient Communications and Computer Systems*. Springer, Singapore, 2018. 427-436.

[10] Husain, Adil, and Muneeb H. Khan. "Early Diabetes Prediction Using Voting Based Ensemble Learning." *International Conference on Advances in Computing and Data Sciences*. Springer, Singapore, 2018.

[11] Poornappriya, T. S., and M. Durairaj. "High relevancy low redundancy vague set based feature selection method for telecom dataset." *Journal of Intelligent & Fuzzy Systems*, Preprint: 1-18.

[12] M. Durairaj, T S Poornappriya, “Choosing a spectacular Feature Selection technique for telecommunication industry using fuzzy TOPSIS MCDM.”, *International Journal of Engineering & Technology*, 7 (4) (2018) 5856-5861.

[13] M. Durairaj, T. S. Poornappriya, “Importance of MapReduce for Big Data Applications: A Survey”, *Asian Journal of Computer Science and Technology*, Vol.7 No.1, 2018, pp. 112-118.

[14] M. Lalli, V.Palanisamy,(2016), “Filtering Framework for Intrusion Detection Rule Schema in Mobile Ad Hoc Networks”, International Journal of Control Theory and Applications –(IJCTA),9(27), pp. 195-201, ISSN: 0974-5572

[15] M. Lalli, V.Palanisamy,(2017), “Detection of Intruding Nodes in Manet Using Hybrid Feature Selection and Classification Techniques”, Kasmera Journal, ISSN: 0075-5222, 45(1) (SCIE)(Impact Factor:0.071).

[16] M. Lalli, V.Palanisamy, (Sep 2014), “A Novel Intrusion Detection Model for Mobile Adhoc Networks using CP-KNN”, International Journal of Computer Networks & Communications- (IJCNC), Vol.6, No.5, ISSN:0974-9322.

[17] M. Lalli, “Statistical Analysis on the KDD CUP Dataset for Detecting Intruding Nodes in MANET”, *Journal of Applied Science and Computations*, Volume VI, Issue VI, JUNE/2019, 1795-1813.

[18] M. Lalli, “Intrusion Detection Rule Structure Generation Method for Mobile Ad Hoc Network”, *Journal of Emerging Technologies and Innovative Research*, June 2019, Volume 6, Issue 6, 835-843.