# A REVIEW ON THE SOFT COMPUTING METHODS FOR THE PREDICTION OF DIABETES MELLITUS

Muni Kumar N[1], Manjula R[2]
[1]Research Scholar, [2]Associate Professor

**Abstract**
**India has got over 50 million Diabetic patients and this condition will rise to more than 205 million patients all around the world by 2035. This disease is affecting more people in the working age group and is proving to be an economic burden to any country. Diabetes is a life-long disease and is characterized by increased sugar levels in the blood. It is either caused due to lack of insulin in the blood (Type 1) or due to lack of response to insulin produced by the body (Type 2). Diabetes affects 347 million people worldwide. If current trend continues, the CDC (Centers for Disease Control and Prevention) estimates 1 in 3 adults in U.S could have diabetes by 2050. Type 2 diabetes is responsible for 90% of all diabetes cases. Timely prediction of diabetes can be a boon for the patients and a lot of research has been carried out for the same. This paper examines and reviews various soft computing algorithms that may be applied to the Diabetes dataset for early prediction of Diabetes Mellitus.**

**Index Terms: Big Data, Diabetes Mellitus, Prediction of Diabetes, Soft Computing.**

## I. INTRODUCTION

India has a high prevalence of diabetes mellitus and the numbers are increasing at an alarming rate. In India alone, diabetes is expected to increase from 40.6 million in 2006 to 79.4 million by 2030 and the projected estimate of the people with diabetes worldwide is 354 million. This statistics clearly indicates that, out of 4 diabetic people in the world, one will be Indian. Other studies have shown that the prevalence of diabetes in urban Indian adults is about 12% and the Type 2 Diabetes is 4-6 time higher in urban than in rural areas. This growth in the urban areas is because of the increase in the rates of obesity which have tripled in the last two decades due to the change in life-style and lack of physical activity. Type 2 Diabetes (T2D) is strongly associated with morbidity and mortality and carries a heavy financial burden.

Diabetes mellitus is a chronic, lifelong condition that affects your body's ability to use the energy found in food. There are three major types of diabetes: Type 1 diabetes, Type 2 diabetes, and gestational diabetes. All types of diabetes mellitus have something in common. Normally, your body breaks down the sugars and carbohydrates you eat into a special sugar called glucose. Glucose fuels the cells in your body. But the cells need insulin, a hormone, in your bloodstream in order to take in the glucose and use it for energy. With diabetes mellitus, either your body doesn't make enough insulin, it can't use the insulin it does produce, or a combination of both. Since the cells can't take in the glucose, it builds up in your blood. High levels of blood glucose can damage the tiny blood vessels in your kidneys, heart, eyes, or nervous system. That's why diabetes – especially if left untreated – can eventually cause heart disease, stroke, kidney disease, blindness, and nerve damage to nerves in the feet. A periodic test called the A1C blood test estimates glucose levels in your blood over the previous three months. It's used to help identify overall glucose level control and the risk of complications from diabetes, including organ damage.

## II. LITERATURE REVIEW

Lindstrom and Tuomilehto [14] developed a diabetes risk score model which consists of Age, BMI, waist circum- ference, history of antihypertensive drug treatment, high blood glucose, physical activity, and daily consumption of fruits, berries, or vegetables as categorical variables.

Park and Edington [17] present a sequential neural network model for diabetes prediction. The authors indicate risk factors, in the final model, including blood pressure, cholesterol, back pain, fatty food, weight index or alcohol index.

Concaro et al [5] present the application of a data mining technique to a sample of diabetic patients. They consider the clinical variables such as BMI, blood pressure, glycaemia, cholesterol, or cardio-vascular risk in the model.

Veena et al [4] used UCI PIMA indian dataset and developed a CART (Classification and Regression Tree) model on SAS Enterprise Miner. A Genetic Alogrithm is being used to impute the missing values in the data and model accuracy is 82.53% as compared to 75.82% on Neural Networks without any pre-processing.

Tarun et al [12] have combined two different techniques for the prediction of diabetes on the PIMA Indian dataset. Principal Component Analysis (PCA) algorithm is used for dimensionality reduction in statistics. Principal component analysis (PCA) is a standard tool in modern data analysis. It is a simple non parametric method for extracting relevant information from data sets. Principal components analysis method is used for achieving the simplification and generates a new set of variables, called principal components. Each principal component is a linear combination of the original variables. All the principal components are orthogonal to each other, so there is no redundant information [10]. The principal components are in the decreasing order of explanation of variance of the data. The author has combined PCA analysis to come up with Principal components which majorly describes these three variables: Plasma Glucose concentration a 2 hrs in an oral glucose tolerance test (GTT), Insulin and Body Mass Index (BMI). These 3 components are used with REP(Reduced Error Pruning Tree) and SVM models in Matlab. SVM along with PCA model had an accuracy of 93.6% as compared to 79,9% accuracy in PCA with REP model.

Stephen M. et al [13] have used data mining technique known as Apriori algorithm to come up with the cause-effect relationships between different kind of diseases and health related information. The dataset used here is the Practice Fusion diabetes challenge data. The goal here is not only just to look into diabetes as specific disease but to identify a pattern in the given dataset to come up with related diseases and then that relationship sample is rolled out to medical practitioners to know their views on the relationships, whether they do feel that such a relationship exists or it’s a mere coincidence or there’s some hidden driving factor due to which the relationship exists. There are two key parameters used for the analysis: Support and Confidence. Support is basically the ratio of observations which contain occurrences of certain disease ‘x’ to the total number of observations whereas confidence is the ratio of the number of observations where diseases ‘x’ and ‘y’ occur together to the number observations where disease ‘x’ occur. ICD-9 code descriptions are mainly used for this analysis and the idea here is to look for relationships of diseases and health problems which have high support and high confidence. There are two scenarios which are taken up: the first one has the support value threshold as 10% and the confidence threshold as 50%, this scenario got only 8 combinations of related diseases but on decreasing the support threshold to 5% in the second scenario and keeping the confidence threshold at 50%, there were 24 combinations of related diseases. The objective of this exercise is to look for relationships and patterns among diseases and those relationships will act as initial thoughts for future work for researchers in the healthcare field to check if these relationships actually exists or they’re just coincidences. C# was used for implementing the algorithm and MS-SQL server for database storage and management.

Pardha Repalli [18] has worked on the SAS Data mining Shootout challenge in 2010 where the problem statement was to diabetes prediction. The author has built bi-variate plots to do preliminary data analysis and gather insights from the data such as the diabetes rate is way higher for people above the age of 45 and they constitute for about 75% of the total diabetic population in the data set. On dividing the ages into three bins of‘less ’than ‘20 ’,‘20-45 ’and ‘greater than 45 ’there is no significant difference among the non-diabetic percentage for ‘less than 20 ’and ‘20-45 ’age group but the non-diabetic rate falls by 7% in absolute terms for people ‘greater than 45 ’. Last Dental checkups, Last checkup, Last Cholest Check, Last PSA test, Last PAP test, Last Breast Exam, Last Mammogram and Wears Seat Belts, these variables are used to consider the life style activities of the members. Max normal transformation is applied to convert the variable into uniform distribution and the data is portioned 50% as the training set and the remaining 50% as the test set.

Akkarapol Sa-ngasoongsong et al [19] has presented his analysis on the SAS Data Mining Shootout challenge in 2010. He has used three algorithms: Logistic Regression, Decision Trees and Artificial Neural Network (ANN) and two different approaches for model building. In the first approach the author had built the model on the complete population but in the second approach the author has divided the population into 3 cost buckets based on the cost that person spends on healthcare. Logistic regression model performed the best in the overall model with a misclassification rate of 22.9%. For the second approach decision tree performed the best in lowest and highest cost bucket whereas logistic regression was best in medium cost bucket. The variables that come out to be significant in both the approaches are: Age, Cholestrol last check and high blood pressure diagnosis.

Deepti Jain and Divakar Singh [11] has presented their analysis by using Feed forward neural network algorithm for prediction of diabetes on SAS Data Mining Shootout challenge dataset. The accuracy of the proposed model is 90% and the authors have focused on other key statistics measures apart from accuracy like precision, recall.

Shubhojit Das has used the SAS Data Mining Shootout challenge dataset and divided the entire population into three segments: children (age < 20 years), male and female. They have used four algorithms for building the model: Logistic regression, Decision Tree, Neural Network and an ensemble of these three algorithms. Ensemble, Neural Network and Decision performed best in the segments of male, female and children respectively based on the averaged square errors of the models. The authors had done analysis on the cost savings if the people in these 3 segments reduce their BMI just by 10.

Hao Yi Ong et al [16] has worked on the Practice Fusion diabetes classification dataset. Unlike the original competition, which assumes that the algorithm will have access to the full medical record of patients and that patients all have a standard database (e.g., exact same tests taken, same recorded variables), the authors are interested in creating a model that assumes only a part of the medical record is known as the input. The authors have used the Bayesian Networks since it does not required all the variables. The authors have evaluated top 8 bayesian networks according to the model and among them once final model is selected based on minimization of (False Positive + False Negative). This model’s performance is again tested with missing information for some of the variables.

Xiaoran Zhang et. al. worked on the Practice Fusion dataset and they have used three different algorithms for comparison of performance: GBM, SVM and Neural Network. In all the three algorithms the authors ran multiple iterations by varying the parameters like: degree of polynomial in SVM, shrinkage and depth in GBM and number of hidden layers in Neural Networks. They have also tried over sampling the data set to get a better response rate and the best model is selected based on the F-measure which combines both precision and recall. The best model is GBM with oversampling of 2.5:1, shrinkage factor of 0.2 and interaction depth of 3. These are the variables which came out to be most important in the GBM model: Hypotension, Lipoprotein deficiencies, Age, D/S Blood Pressure and Weight.

Likhitha Devireddy [7] did a very thorough experiment with a large number of algorithms, their ensembles and feature selection. Their group took 17 tables containing a wide variety of healthcare information about 9948 patients and combined it into one large dataset with 980 features per patient and then created 24 distinct versions of this dataset, applying six different methods of feature selection to the data and 3 different methods of numeric transformation. The authors then applied a variety of modeling techniques, which were evaluated using accuracy and lift. The best-performing models were two ensembles, and they achieved instantaneous lift ratios of 3.17 (at a positive prediction rate of 8.3%) and 3.012 (at a positive prediction rate of 17.97%) on unseen data. The authors also showed the general efficacy of feature selection, both as a basis for model induction and as a technique for exploring datasets.

Ariana E. Anderson et al [2] worked on the Practice Fusion diabetes data set and created three scenarios conventional model-just like the risk scores model, full HER model containing conventional information and both diagnostic and prescription information and EHR DX model conventional information and diagnostic information. Two type of algorithms were used for the model building process: Logistic Regression and Random Forest. Random Forest model performed better than logistic regression model and had an accuracy of 78.7%.

Archeana and Anita [3] has discussed how Hadoop plays an effective role in performing real time analysis on the patient health care data such as physician notes, lab reports, x-ray reports, diet regime, medicine and surgical instruments expiry dates etc. Also discussed the need for big data analytics in health care to provide patient centric services, detecting the spread of diseases, monitoring the hospitals quality and improving the treatment methods.

Meng et.al [15] have compared the performances of three data mining models i.e., Logistic regression, artificial neural networks (ANN) and decision tree models for the prediction of diabetes or pre-diabetes by considering the common risk factors and identified that decision tree model (C5.0)had the best classification accuracy, followed by the logistic regression model and the ANN with the lowest accuracy. In this study, the authors have prepared a standard questionnaire with the common risk factors of diabetes and gathered information from a total of 1487 individuals which include 735 volunteers who were confirmed to have diabetes and the remaining 752 volunteers without diabetes or pre diabetes.

A. Iyer et.al [10] carried out the diagnosis of diabetes using Decision Trees and Naive Bayes to diagnose the disease by analysing the patterns found in the data through classification analysis. Further, they proved that both the methods have a comparatively small difference in error rate and both the models are efficient in the diagnosis of diabetes using the percentage split of 70:30 of the dataset.

Sridhar & Shanthi [20] have employed the combination of Back propagation algorithm and Apriori algorithm and proved that greater accuracy is achieved in the diagnosis of Diabetes Mellitus. Further, these authors have developed a web based artificial neural network with Association rule mining for the diagnosis of Diabetes Mellitus.

Arwa et.al [1] have carried the predictive analytics for the diagnosis of Diabetes using the Artificial Neural Network on WEKA software. PIMA INDIAN diabetes dataset is used with multilayer perceptron training technique for better prognosis of the disease.

Renuka Devi & Maria [6] have analyzed various data mining techniques to predict Diabetes Mellitus in detail along with the technique employed, tools used and accuracy achieved. Also proved that the modified J48 classifier provided 99.87 % of highest accuracy using WEKA & MATLAB tool.

Eswari et.al [9] proposed a predictive methodology that uses the predictive analysis algorithm in Hadoop/Map Reduce environment to predict the diabetes types prevalent, complications associated with it and the type of treatment to be provided based on the analysis. This proposed system also provided an efficient way to cure and care the patients with better affordability and availability.

Durairaj and Kalaiselvi [8] performed a detailed survey on the application of different soft computing techniques for the prediction of diabetes. Also compared the performance of different data mining algorithms such as SVM, KNN, C4.5 and ANN. Further, the authors observed that ANN provides more accurate results than other classification techniques and ANN is identified as the best technique for the prediction of diabetes disease and classification.

## III. APACHE SPARK

Apache Spark is a cluster computing platform designed to be fast and general-purpose. Spark extends the popular MapReduce model to efficiently support computations including interactive queries and stream processing. Spark is a step ahead from Hadoop and it uses the in-memory space of the commodity hardware for computation which reduces the calculation time considerably. Spark has been developed by the AMP labs based out of University of California, Berkley. Another thing to note here is that spark works on the lazy evaluation principle. Spark supports various languages such as Scala, Java, Python and R. For the prediction of diabetes Spark's MLlib library can be used which supports machine learning.

The soft computing algorithms are quite similar for big data and traditional databases, it's just the way how data is stored and the computations are done which differentiates the big data technologies from the traditional ones. For a large data set big data tools computes the results and build the models in a shorter time as compared to the traditional tools.

## IV. SOFT COMPUTING TECHNIQUES

Soft computing is a collection of algorithms that are employed for finding a solution for very complex problems, the ones for which more conventional methods have not yielded low cost & time-feasible solutions. Soft computing methodologies have proved to be advantageous in the implementation of recommendation systems. In contrast to analytical methods, soft computing methodologies mimic consciousness and cognition in several important respects: they can learn from experience; they can universalize into domains where direct experience is absent; and, through parallel computer architectures that simulate biological processes, they can perform mapping from inputs to the outputs faster than inherently serial analytical representations. The trade off, however, is a decrease in accuracy. If a tendency towards imprecision could be tolerated, then it should be possible to extend the scope of the applications even to those problems where the analytical and mathematical representations are readily available.

## V. DATASETS

Electronic medical records (EMR) contain sensitive information such as medical details related to infectious diseases such as Human Immunodeficiency Virus (HIV) or they may contain information about mental illness and sensitive information related to fertility treatments . So, accessing and analyzing EMR databases is a tedious task and only very few health related datasets are available for research purpose. Some of the data sets available are:

- *PIMA Indian Dataset*

This is the mostly widely used data set for diabetes prediction and is freely available at the UCI machine learning repository Link. This dataset has 768 observations and 8 variables: Plasma, pressure, skin, insulin, pregnancy, mass, pedigree, age. The dataset is a collection of medical diagnostic reports of 768 women above the age of 21 from a population living near Phoenix, Arizona, USA.

- *Practice Fusion Diabetes Classification*

This is a practice competition that was hosted on kaggle.com in 2012 where the participants were given a sample of Electronic Medical Records (EMR) data for 10,000 patients Link. The papers on this dataset are the most advanced in terms of usage of machine learning techniques and model building.

- *SAS Knockout 2010*

Prediction of diabetes was the topic of SAS knockout challenge in 2010. The dataset has 50,788 records with 43 variables. The variables which turned out to be most important in the model are: High Blood Pressure, Cholest_Last_chck, Heart_disease, Los_all_teeth, Last_flushot, Years_Educ, etc. The response rate in the data set is 5% whereas the percentage of diabetic population in US is around 8.3%.

## VI. SUMMARY

The following are the observations and gaps identified during the review of various literature for the prediction of Diabetes Mellitus using Big Data and Soft Computing algorithms.

- **Accuracy vs other model parameters:** Accuracy actually can be deceptive. A better quantity to talk about can be the AUC value and its shape or rather precision and recall. AUC value is the area under the ROC (Receiver Operator Characteristic) curve. Let's take this scenario, suppose there's a dataset in which 10% of the people have diabetes, so if random tagging is done for everyone as 0 (i.e., not having diabetes, then the accuracy of the model would be 90%). So accuracy is basically how well the model is predicting both '0' and '1' where the focus should be how well we are tagging 1's (the people with diabetes) which can be obtained from precision and recall. Precision is the percentage of number of observations predicted as 1s which are actually 1s and recall is the percentage of predicted 1s which are actually 1s among the total number of actual 1s we have in the model.

- **Model Comparison vs Understanding**: the model building process: Machine learning models have different underlying assumptions and have different application areas based on the dataset. So if an algorithm does not perform well that might be because the modeler have not tuned the algorithm to fit the dataset or the algorithm cannot get better results on the given data set. The model building process requires calibration and new feature creation which gives a better insight about how the variables interact and how this interaction defines the relationship of independent variables with the target variable.

- **Model Inference:** The most important and interesting takeaways from a predictive modeling paper are how the independent variables interact with the target variable. For example how does smoking affect the diabetes trend among people, similar inferences can be drawn from other variables too. It is easier to draw such inferences from simpler models like decision trees, logistic regression as compared to SVM, Artificial Neural Networks.

- **Model statistics should be reported on the hold-out sample:** Machine learning algorithms can be greedy in nature and they can over fit the given dataset to achieve 100% accuracy. So the model performance statistics should be reported on the hold out sample which is usually called the test data set.

## VII. CONCLUSION

This paper discussed about the current statistics of Diabetic Mellitus in India and projected statistics by 2025 and showed that for every 4th persons having diabetes in world will be Indian as 25% of the world diabetic patients will be from India due to changes in the life styles. Also this paper discussed and reviewed the various methods and literature available for the analysis of Diabetic Mellitus datasets to predict and forecast the disease in pre diabetic stages, so that better diagnosis can be made to reduce the diabetic patients.

## VIII. FUTURE WORK

With respect to the observations stated in the above section, the following are some of the future enhancements for the current work. As the processing speed of Apache Spark is much better compared to the conventional MapReduce in Hadoop. Therefore, the implementation of the diabetes recommendation system using Apache Spark and HDFS is one of the future enhancement. Further, the recommendation system can be enhanced to get implemented in the Cloud environment using the cloud services such as Amazon Web Services , Google Cloud or Data Bricks so that real power of Spark can be utilized in the processing of huge amounts of data for the better decision making.

## ACKNOWLEDGEMENT

I would like to take this opportunity to express my profound gratitude and deep regard to my guide, Prof. R. Manjula, SCOPE, VIT Institute of Technology, for her exemplary guidance, valuable feedback and constant encouragement in completing this paper. Her valuable suggestions were of immense help in getting this work done. Working with her, was an extremely knowledgeable experience. Also, I would like to extend my sincere gratitude to my parents, my wife and my lovable daughter for their constant support and encouragement in completing this paper.

## REFERENCES

[1] Arwa Al-Rofiyee, Maram Al-Nowiser, Nasebih Al-Mufadi, and Mohammed Abdullah AL-Hagery. Using prediction methods in data mining for diabetes diagnosis. POSTERS, May, 2014.

[2] Ariana E Anderson, Wesley T Kerr, April Thames, Tong Li, Jiayang Xiao, and Mark S Cohen. Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general united states population: A cross-sectional, unselected, retrospective study. Journal of biomedical informatics, 60:162–168, 2016.

[3] J Archenaa and EA Mary Anita. A survey of big data analytics in healthcare and government. Procedia Computer Science, 50:408–413, 2015.

[4] Veena H Bhat, Prasanth G Rao, P Deepa Shenoy, KR Venugopal, and Lalit M Patnaik. An efficient prediction model for diabetic database using soft computing techniques. In International Workshop on Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing, pages 328–335. Springer, 2009.

[ 5 ] Stefano Concaro, Lucia Sacchi, Carlo Cerra, Mario Stefanelli, Pietro Fratino, and Riccardo Bellazzi. Temporal data mining for the assessment of the costs related to diabetes mellitus

pharmacological treatment. In AMIA, 2009.

[6] M Renuka Devi and J Maria Shyla. Analysis of various data mining techniques to predict diabetes mellitus.International Journal of Applied Engineering Research, 11(1):727–730, 2016.

[7] Likhitha Devireddy, David Dunn, and Michael Sherman. A feature-selection based approach for the detection of diabetes in electronic health record data. 2014.

[8] M Durairaj and G Kalaiselvi. Prediction of diabetes using soft computing techniques-a survey. International Journal of Scientific & Technology Research, 4, 2015.

[9] T Eswari, P Sampath, and S Lavanya. Predictive methodology for diabetic data analysis in big data. In 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15), Science Direct.

[10] Aiswarya Iyer, S Jeyalatha, and Ronak Sumbaly. Diagnosis of diabetes using classification mining techniques. International Journal of Data Mining & Knowledge Management Process, 2(1), 2015.

[11] Deepti Jain and Divakar Singh. A neural network based approach for the diabetes risk estimation. Inter- national Journal of Computer Applications, 73(10), 2013.

[12] Tarun Jhaldiyal and Pawan Kumar Mishra. Analysis and prediction of diabetes mellitus using PCA, REP and SVM. International Journal of Engineering and Technical Research, 2(8):164–166, 2014.

[13] Stephen M Kang and Peter W Wagacha. Extracting diagnosis patterns in electronic medical records using association rule mining. International Journal of Computer Applications, 108(15), 2014.

[14] Jaana Lindstrom and Jaakko Tuomilehto. The diabetes risk score. Diabetes care, 26(3):725–731, 2003.

[15] Xue-Hui Meng, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang, and Qing Liu. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences, 29(2):93–99, 2013.

[16] Hao Yi Ong, Dennis Wang, and Xiao Song Mu. Diabetes prediction with incomplete patient data. Technical report, Technical report, 2014.

[17] Jin Park and Dee W Edington. A sequential neural network model for diabetes prediction. Artificial intelligence in medicine, 23(3):277–293, 2001.

[18] Pardha Repalli. Prediction on diabetes using data mining approach. Oklahoma State University, 2011.

[19] Akkarapol Sa-ngasoongsong and Jongsawas Chongwatpol. An analysis of diabetes risk factors using data mining approach. Oklahoma state university, USA, 2012.

[20] Shanthi D Sridhar K. Medical diagnosis system for the diabetes mellitus by using back propagation- apriori algorithms. Journal of Theoretical and Applied Information Technology, 68(1), 2014.