A Hybrid approach for an analysis of diabetes and prediction using machine learning techniques

[1]Dr. S Kavitha Bharathi, [2]M. Dhavamani, [3]Mr. K. Srihariakash

Abstract--*Healthcare data management and analysis applications are constructed to handle huge volume of data items. Clinical and analytical methods are applied for the Diabetes prediction process. The machine learning techniques are adapted to analyze the diagnosis data values. The classification methods are applied to discover the patterns and apply the patterns for disease prediction process. Late diagnosis of diabetes decreases the risk of coronary, kidney, nerve injury and blindness.*

The diabetes prediction is carried out with the medical bioinformatics analytics. The resampling methodology for bootstrapping is paired with the classification approaches for Naive Bayes, Decision Trees and KNN. Continuous Glucose Monitoring Systems (CGMS) are very essential to monitor the blood glycaemic levels. The potential estimation of glycaemic rates facilitates the avoidance of adverse hyperglycaemic or hypoglycaemic situations. To order to boost efficiency of diabetic treatment, glycaemic thresholds are taken into consideration. The prediction models have been developed for the same stage of patient diagnosis.

The time series data analysis model is very efficient in the diabetes prediction and control applications. The multi patient glucose level data is applied to construct the prediction model with large volume of sample data. The prediction model is applied to forecast the blood glucose level for the other patients. The bootstrapping resampling technique and Support Vector Machine (BRSVM) are integrated to construct the prediction model and future level identification process. The benchmark diabetes diagnosis data values are used in the analysis.

Keywords--*Machine learning and classification, Diabetes diagnosis, Decision tree, Support Vector Machine and Bootstrapping Resampling based Support Vector Machine*

I INTRODUCTION

Data mining is the method of exploration and translation into valuable knowledge – details capable of increasing income, rising costs or both – from dissimilar points of view. One of the research methods to evaluate knowledge is data mining applications. It helps users to research data, categorize it and summarize defined relationships from different measurements and angles. Technically data mining is the technique of broad connection repositories for finding similarities or trends of hundreds of fields. Some techniques have been used for data mining on users' social demographic info. For data mining for diabetes the correction and up-to-date data is quite necessary. Diabetes is a condition that arises when the development of insulin in the body becomes insatisfactory or because the system is not able to efficiently utilize the insulin produced such that blood glucose

[1]Department of Computer Applications, Kongu Engineering college, Erode, Tamilnadu, India

[2]Department of Mathematics, Kongu Engineering college, Erode, Tamilnadu, India

[3]Department of Computer Applications, Kongu Engineering college, Erode, Tamilnadu, India

Received: 27 Mar 2020 | Revised: 20 Apr 2020 | Accepted: 05 May 2020

is elevated. Skin cells split the food down into glucose and all body cells consume this glucose. Insulin is the hormone which guides the glucose that breaks the food into the cells of the body. Every rise in insulin production contributes to an improvement in blood sugar rates, which may contribute to tissue degradation and organ failure. In general, if the blood sugar level is above average (4.4 to 6.1 mmol / L), the individual is known to have diabetes.

Type 1, Type 2 and Gestational diabetes were classified in three major forms. While there are currently approximately 10% of Type 1 diagnoses, the number of these cases in the US has recently risen. The disease is a disease which develops at a very limited age of 20 and is therefore also known as diabetes in juveniles. In this situation, the human body's protective mechanism destroyed the pancreatic cells containing insulin. Patients of Type 1 diabetes condition will be treated by administering insulin coupled of daily blood checks and dietary limits.

Moreover, type 2 diabetes constitutes 90% of diabetes and is generally recognized as adult or non-insulin based diabetes. Around this point the body's various components are insulin-resistant and the need for insulin is raised. The pancreas does not generate the amount of insulin needed at this point. Patients must be stringent on food, regular activity and track blood pressure to maintain this variation of diabetes in order. Obesity may contribute to type 2 diabetes, such as overweight, since it isn't sufficiently involved. The incidence of diabetes is often shown to be higher because of the age. Many Type 2 patients suffer from irregular or pre-diabetes diabetes, a disease in which glucose rates are greater than average but not as severe as diabetes.

Gestational diabetes is one of the most severe diabetes that exists in expectant women because of the rise in sugar rates when the pancreas does not contain enough insulin. No intervention will handle childbirth problems. Diabetes may be regulated by dietary testing and insulin intake. All these forms of diabetes are severe and need care which can be prevented if diagnosed at an early age.

A time series is a sequence of time-listed data points. More frequently than not, a time sequence is a loop at many periods that are similarly distant. It is therefore a collection of independent details. The time series contains the ocean tide height, the sunspots count and the Dow Jones Industrial Average regular closing interest. Time series are plotted quite frequently by string maps. In numbers, signals, trends, econometrics, thematic financials, weather predictions, earthquake eprophecy, electro-encephalography, power systems, physics, communications and mostly in all fields of applied science and engineering, time series is used.

To order to derive evocative results and other distinctiveness from results, the Time Series psychoanalytical research includes approaches to evaluate time period data. Time series prediction is the implementation of a formula in order to formulate possible values dependent on historically realistic values. While regression analyzes are frequently under way in terms of evaluating hypotheses that current time series values influence the present value of any other time series, this form of time series analysis is not referred to as "time series analysis" that focuses on contrasting time series or several time series in specific time series. Application of episodic time series is the mechanism by which the actions are analyzed within one time period.

II RELATED WORK

Alessandro Aliberti et al. [2019] recommended that the blood glucose estimation mechanism be applied on a multi-patient basis. The prediction models trained a large and heterogeneous patient population and are therefore usable for the glucose-level capacity to be generated from a brand new individual. In several time series forecasting issues, on development of the Nonlinear Autoregressive Network (NAR) and on long-term memory networks (LSTM), two separate approaches have been proved efficient. Just short-term forecasts are reliable to the NAR. The LSTM is particularly strong in short and long term glucose depletion.

Sajida Perveen et al [2018] described a predictive modeling presented on machine learning methodology for metabolic syndrome and development of diabetes mellitus. The terms diabetes mellitus and the second variety diabetes mellitus are used as interchangeably. The relation of diabetes and individual risk factors of MetS, are examine. The analytical power of significant metabolomics data in predicting future risk of second type diabetes is verified using machine learning methodology. The weather data sampling approaches produce balanced instruction sets could improve the relative recital of these methods.

Manu Goyal et al [2019] identified powerful ways of detecting and locating diabetic foot ulcers on mobile devices in real-time. The profound learning methods are used for DFU position in real time. A large collection of 1775 DFU photos allows a powerful profound learning paradigm together. By demarcating the DFU electoral interest constituency with annotative tools, two medical experts created the opinion truths on the results. The faster R-CNN with the two-tier relocation of InceptionV2 software gain a high degree of accuracy with the scale of a modell. A larger data collection will further boost the opportunity to know extensively in real time localization of the DFU.

Priya B. Patel et al [2017] studied diabetes predictive data mining algorithms. Gateway to the release of body cells is the hormone. The human body will use glucose for energy. The amount of glucose in the human body is regulated by insulin. The origin of diabetes is the successful production of blood glucose. Common physical and comical diabetes reported, but it does not have an exact cure. Specific data mining techniques are being applied for diabetes mellitus prediction and diagnosis. Gaussian Naive Bayes, KNN, SVM and the Decision Tree are important data mining algorithms. The chosen data collection is based on a Pima Indian Diabetic System of Machine Learning software from University of California, Irvine (UCI).

Steffi et al., [2018] released data mining for diabetes mellitus prediction techniques. Diabetes prediction using data mining techniques is being studied. The dataset took over 768 PIMA Indian Diabetes Dataset instances to estimate the accuracy of the data mining process. The Nine Input Variables and the 1 Output Vary from Dataset Knowledge are built into five predictive models. The five models are calculated in terms of accuracy, flexibility, specificities and measurements in the F1 Score system.For the delivery of diabetes through specific risk factors a results dependent study of Naïve Bayes, Logistic Regression, Artificial Neural Networks (ANNs), C5.0 Decision Tree, and Support Vector Machine (SVM) models is comparable. The process decision tree (C5.0) produces the strongest precision division accompanied by Naïve Bayes, ANN and SVM, the logistic regression process. The artificial neural neural network mechanism was predicted by Suyash Srivastava et.al[2019]. Analyze and analysis based on the diabetes paradigm were focused on AI (Artificial Intelligence). Patients of diabetes in a local population have carried out the form of prediction. Pima Indians have been granted the possibility to predict diabetes with sample results. The Artificial Neural Network (ANN) has been chosen for simulation to forecast diabetes from multiple Machine Learning algorithms. This approach is idyllic for the delivery of high-precision diabetes possibilities as the software measurements are performed

III SUPPORT VECTOR MACHINE

A collection of similar training facts, each hallmark in one or two separate categories, an SVM algorithm constructs a straightforward prototype that assigns new models to one or the next category, constructs new examples to one category or another, and creates an SVM training algorithm. A model SVM is a space-specific example that maps the examples of split categories to the degree that they are alienated by a large, as far as possible distance. New instances are then projected into the same space and a form dependent on which side of the distance they fell predictions belong.

IV PROBLEM STATEMENT

Health sector has incredibly broad and confidential details which must be treated with considerable caution. Diabetes Mellitus is one of the rising deadly diseases in the world. As a robust diagnosis method, medical experts tend to detect diabetes. For interpreting data from multiple viewpoints and description into usable results, numerous machine learning approaches are useful. When it comes to the analysis of other data mining techniques, the quick exposure and availability of vast volumes of data will provide valuable information. The key philosophy is to build new patterns and give users more valuable knowledge and perceive such patterns. Diabetes contributes to coronary disease, metabolic dysfunction, nerve damage and blindness. The professional processing of diabetes data is a critical problem. The strategies and methodology of data mining would be presented to develop effective approaches and strategies for professional diabetes detection and for useful trends.

- Diabetes was projected by scientific bioinformatics analyzes. A UCI registry for research has been acquired for the Pima Indian diabetes project. The dataset was developed and evaluated to construct powerful models for diabetes prediction and diagnosis. In order to increase the precision and then use Naïve Bayes, Decision Trees and (KNN) to test the efficiency of the bootstrapping resample. In the latest diabetes prediction systems, the following issues are listed.
- The Naïve Bayes algorithm and decision tree algorithm achieves restricted divisions of accuracy on the diabetes prediction process.
- The diabetes prediction is carried out with single diagnosis transaction data for each patient.
- The diabetes prediction accuracy is low in the single diagnosis instance model.

Received: 27 Mar 2020 | Revised: 20 Apr 2020 | Accepted: 05 May 2020

- Continuous glucose monitoring and resampling operations are not applied on the prediction process.

V DIABETES PREDICTION WITH BOOTSTRAPPING RESAMPLING SUPPORT VECTOR MACHINE (BRSVM)

The model for diabetes prediction is based on the different patient learning processes for diagnosis. Continuous glucose monitoring systems (CEMS) have a substantial amount of knowledge to calculate the blood glycaemic significance of a diabetic patient at a fast sampling pace. In the mechanical learning methods, such data may be efficiently used to infer glycamic concentration levels, to avoid dangerous hyperglycaemic or hypoglycaemic situations and to improve diabetic care. The prediction models focused on the glucose symptoms of a large, heterogeneous legion of patients and used a totally novel patient to expect predicted glucose rates.

Throughout the diabetes prediction and management systems, the time series data analysis paradigm is quite well structured. Data from several patients' glucose rates was used to build the prediction model with wide amount of test results. The forecast model is used to estimate certain patients' blood glucose rates. To develop the prediction model and potential level recognition system, the bootstrapping resampling and vector support (SVM) framework are developed. Data values for evaluation of reference diabetes are included in the study.

The Vector Virtual System (SVM) and the Virtual Vector System (BRSVM) algorithms for Bootstrapping Resampling are used to predetermine diabetes. The PIMA Diabetes Dataset Indians of the National Institute for diabetes and digestive and renal disorders including data from diabetic patients have been identified by classification algorithms.

VI PERFORMANCE ANALYSIS

The diabetes prediction operations are construct with machine learning methods. The patient diagnosis data is gathered from the PIMA Indian diabetes database is analyzed in the prediction models. The classified techniques are applied in the diabetes prediction process. The diabetes data set is analyzed in two levels single instance-based diagnoses and multipleinstance based diagnosis models. The Support Vector Machine (SVM) and Bootstrapping Resampling based Support Vector Machine (BRSVM) techniques are used in the diabetes prediction process. The diabetes prediction accuracy level is estimated with the predicted samples and actual diabetes level in the main data set. Figure 6.1 and table 6.1 shows the diabetes prediction accuracy level between the Support Vector Machine (SVM) and Bootstrapping Resampling based Support Vector Machine (BRSVM) techniques. The Bootstrapping Resampling based Support Vector Machine (BRSVM) 10% increases the prediction accuracy level than the Support Vector Machine (SVM) technique. The Continuous Glucose Monitoring model based data analysis with Bootstrapping Resampling Support Vector Machine (BRSVM) achieves better diabetes prediction accuracy levels.

Table 1: Diabetes prediction accuracy analysis between Support Vector Machine (SVM) and Bootstrapping Resampling based Support Vector Machine (BRSVM) techniques

Transactions	SVM	BRSVM
200	72.92	84.31
400	75.26	85.63
600	78.23	87.48
800	80.89	89.79
1000	82.64	92.76

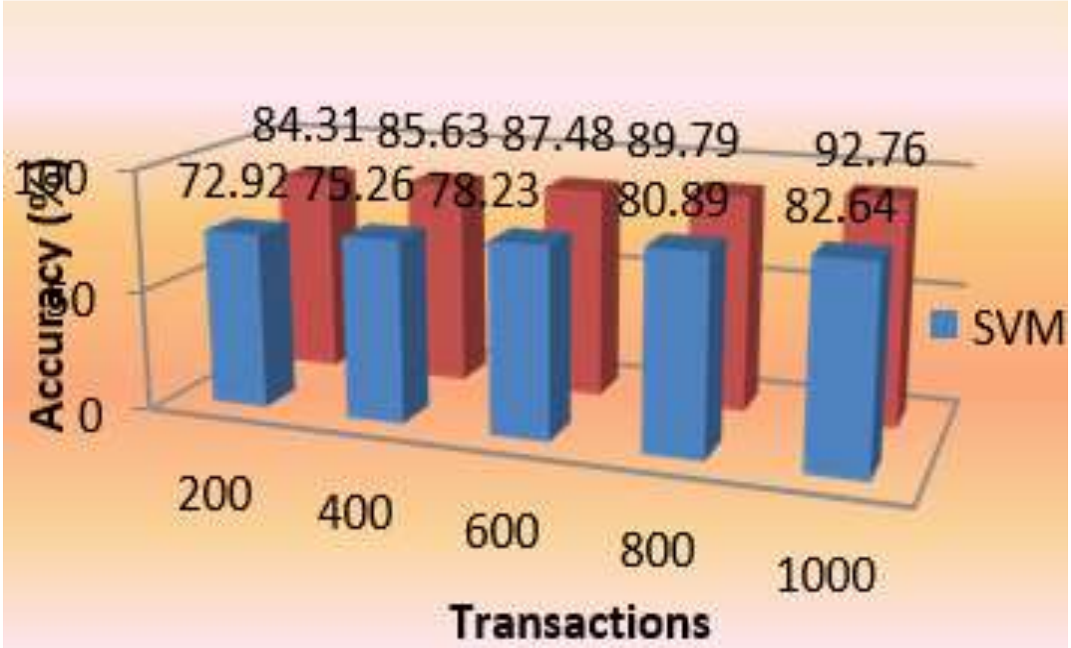

Figure 1: Diabetes prediction accuracy analysis between Support Vector Machine (SVM) and Bootstrapping Resampling based Support Vector Machine (BRSVM) techniques

VII CONCLUSION AND FUTURE WORK

The machine learning methods are implemented to classify the diabetes disease levels. The Glucose monitoring samples are collected from various patients. The samples are analyzed with the Support Vector Machine (SVM) and Bootstrapping Resampling based Support Vector Machine (BRSVM) techniques. Single and continuous Glucose Monitoring samples are used in the prediction process. The Bootstrapping Resampling based Support Vector Machine (BRSVM) increases the diabetes prediction accuracy level in a extensive manner. The diabetes prediction model can be enhanced to handle the age group based diabetes prediction and medicine impact on the diabetes treatment cycles.

REFERENCES

1. P. Yashoda and M. Kannan, “Analysis of a Population of Diabetic Patients Databases in Waikato”, *International Journal of Scientific & Engineering Research*, vol. 2, no. 5, pp. 1-5, 2011.

2. A. Ayer, J. S and R. Sumbala, “Diagnosis of Diabetes Using Classification Mining Techniques”, *arXiv preprint arXiv:1502.03774*, vol. 5, no. 1, pp. 01-14, 2015.

3. NIyati Gupta, A. Rawal and V. Narasimhan, “Accuracy, Sensitivity and Specificity Measurement of Various Classification Techniques on Healthcare Data", *IOSR Journal of Computer Engineering*, vol. 11, no. 5, pp. 70- 73, 2013.

4. Manu Goyal, Neil D. Reeves, SatyanRajbhandari and MoiHoon Yap, “Robust Methods for Real-Time Diabetic Foot Ulcer Detection and Localization on Mobile Devices”, IE*EE Journal of Biomedical and Health Informatics*, Vol. 23, No. 4, 2019.

5. K. Sharmila and S. Manickam, “Efficient Prediction and Classification of Diabetic Patients from big data using R,” *International Journal of Advanced Engineering Research and Science*, vol. 2, 2015.

6. S. Sadhana and S. Savitha, “Analysis of Diabetic Data Set Using Hive and R,” *International Journal of Emerging Technology and Advanced Engineering*, vol. 4, 2014.

7. Sassanian and G. Hari Sekaran, “Big Data Analytics Predicting Risk of Readmissions of Diabetic Patients,” *International Journal of Science and Research*, vol. 4, 2015.

8. S.Saru and S.Subashree, “Analysis and Prediction of Diabetes Using Machine Learning”, *International Journal of Emerging Technology and Innovative Engineering*, vol. 5, no. 4, 2019.

9. Alessandro Aliberti, Irene Pupillo, Stefano Terna, Edoardo Patti and Andrea Acquaviva, “A Multi-Patient Data-Driven Approach to Blood Glucose Prediction”, *IEEE Access*, 2019.

10. Priya B. Patel, Parth P. Shah and Himanshu D. Patel, “Analyze Data Mining Algorithms For Prediction Of Diabetes”, *International Journal Of Engineering Development And Research*, vol. 5, no. 3, 2017.

11. Sajida Perveen, Muhammad Shahbaz, Karim Keshavjee and Aziz Guergachi, “Metabolic Syndrome and Development of Diabetes Mellitus: Predictive Modeling Based on Machine Learning Techniques”, *IEEE Access*, 2018.

12. J. Steffi, Dr.R.Balasubramanian and K.Aravind Kumar, “Predicting Diabetes Mellitus using Data Mining Techniques”, *International Journal Of Engineering Development And Research*, vol. 6, no. 2, 2018.

13. Suyash Srivastava, Lokesh Sharma, Vijeta Sharma, Dr. Ajai Kumar and Dr. Hemant Darbari, “Prediction of Diabetes Using Artificial Neural Network Approach”, *Research Gate, ICoEVCI*, 2018.

14. DeeptiSisodia and Dilip Singh Sisodia, “Prediction of Diabetes using Classification Algorithms”, *International Conference on Computational Intelligence and Data Science*, *Elsevier*, 2018.