LEVERAGING BIG DATA CLOUDS TO IDENTIFY DIFFERENT FORMS OF DIABETES

Mrs. B. Laxmi[1], Divya Salandri[2], M. Nikitha[3], B.S. Vinutha Reddy[4], K. Sonali[5]

Article History: Received: 11.02.2023 **Revised:** 26.03.2023 **Accepted:** 11.05.2023

Abstract

Technological advancements in the recent years has seen an unparalleled rise. Today, technologies like big data, machine learning, robotics, deep learning, and many other AI enabled softwares have eased the ways to develop and implement many disease detecting and monitoring systems and applications. And one such disease is Diabetes. It is vital to develop efficient methods for the diabetes diagnosis and treatment due to the life-long and extensive harm that comes with diabetes. Systems that have been previously proposed suffer from networking and intelligence issues. The goal of this work is to develop a customised, intelligent, and cost-effective diabetes diagnosis solution using machine learning techniques including random forest, SVM, ANN, XgBoost, and AdaBoost. One of the main benefits of cloud-based diagnosis is that it enables real-time monitoring and analysis of a patient's health data, allowing medical personnel to spot patterns and trends that may not be obvious using conventional diagnostic techniques.

Keywords: Accuracy Score, Insulin, Diabetes, Ensemble Algorithm, Efficiency.

[1]Assistant Professor, Department of Computer Science & Engineering, Sridevi Women's Engineering College, Hyderabad, Telangana, India.
[2,3,4,5]Final Year B.Tech, Department of Computer Science & Engineering, Sridevi Women's Engineering College, Hyderabad, Telangana, India.

Email: laxmikalpanaswec19@gmail.com[1], divyasalandri16@gmail.com[2], nikithasonu2002@gmail.com[3], vinuthabana@gmail.com[4], ssonaly39@gmail.com[5]

DOI: 10.31838/ecb/2023.12.s3.294

1. Introduction

Diabetes is a chronic illness which affects a huge masses of people and is characterised by elevated blood glucose (sugar) levels. The hormone insulin, that is created by the pancreas, generally controls blood sugar levels in the body. High blood sugar levels are due to either insufficient or inefficient use of the insulin that the body produces in diabetics. Approximately 422 million people worldwide suffer from diabetes. It is important to remember that 90% of cases of diabetes are type 2 diabetes mellitus. Additionally, teenagers and adults are more likely to have this diagnosed. It's critical to advance strategies for the prevention and treatment of this fatal disease because diabetes has a significant influence on the world's economy and well-being. Diabetes is also brought on by elements including an unhealthy lifestyle, vulnerable situations, and accumulated stress from society and the workplace. If mistreated or badly managed, it can have detrimental long-term effects on one's health, including damage to the eyes, kidneys, nerves, and blood vessels as well as an elevated risk of heart disease and stroke. The standard course of treatment includes taking various drugs, altering one's lifestyle, and checking blood sugar levels. The use of cloud technology in the diagnosis of diabetes is a rapidly evolving field that offers many potential benefits. Cloud-based diagnosis of diabetes involves using computer systems and remote servers to store, analyse, and interpret data related to blood sugar levels, insulin levels, and other health metrics. This data can be collected from a range of sources, including wearable devices, glucose meters, and electronic health records. Another benefit of cloud-based diagnosis is thatit allows for more efficient sharing andcollaboration betweenhealthcare providers, enablingthem to worktogether to developcomprehensivetreatment plans and coordinate care. By lowering the need for additionaltesting and pointlesstreatments, this can assist to improve patient outcomes and lowerhealthcareexpenditures. Overall, cloud-based diagnosis of diabetes has the potential to revolutionise the way in which diabetes is diagnosed and managed, offering greater accuracy, efficiency, and convenience for both patients and healthcare providers. However, it is necessary to ensure that data privacy and security protocols are in place to secure patient information and maintain confidentiality.

Related Work

1. Statistics show that diabetes, which affects one in eleven people, is one of the most common metabolic illnesses. Nevertheless, one in two adults who have diabetes do not receive treatment, and by 2040, one in ten adults will have the condition. This study created a blend of Adaptive Neuro-Fuzzy Inference System (ANFIS) model to classify people with diabetes (Pima Indians Diabetes Dataset) based on diabetic patients data set. The Neuro-Fuzzy ANFIS modelling was implemented using the MATLAB Toolbox and the ANFIS Fuzzy Logic Toolbox. In order to evaluate the algorithm's performance, specificity, accuracy, and sensitivity were taken into consideration. The accuracy for training data for the proposed neural network was 85.35%, while for testing data it was 84.27%, both utilising the Pima Indians Diabetes Database.

2. To maintain the body's necessary blood glucose level, each patient receiving insulin-dependent diabetic mellitus (IDDM) medication needs to take the sufficient amount of insulin at the proper times. The real-time decision rules for generating IDDM therapy are computed using a data stream mining technique in this article depends on the patient's blood glucose reactions and the prescription records for the patient's insulin. Decision criteria are based on the most recent health conditions that are continuously observed by the patient, as opposed to using a population's historical data repository acquired over years. The criteria are therefore adaptable and better predict if a medical implication will develop since glucose levels fluctuate as a result of numerous medical consequences, such as changes in lifestyle, the use of various drugs, or other external variables. An experiment using a computer simulation is done to find the data stream algorithms that are the most effective in terms of accuracy and speed.

3. The amount and variety of services being offered by different vendors has increased as a result of the growing use of mobile computing and smartphone technology. These mobile service providers provide support for several fresh services that are comparable in terms of functionality but have different quality standards. Quick selection and composition of services from the services pool are needed to simplify an automated service procedure. In order to swiftly provide customers with the required service composition, more efficient solutions are required due to the ambient and dynamic nature of the mobile environment. In a short amount of time, it can be difficult to select the best required services from the numerous sets of dynamic services. In this study, the subject is approached as an optimisation problem. An approach is brought up by combining particle swarm optimisation and k-means clustering. It runs concurrently on the Hadoop platform through MapReduce. Parallel processing can create the ideal service composition considerably more quickly than other methods. This is important for managing vast volumes of hybrid data and services from different sources in the cell phone context. The effectiveness of this proposed solution for big data-driven service composition is proven through modelling and simulation.

4. The promise of cloud-supported cyber-physical systems (CCPSs) has piqued the interest of both academics and industry. Physical equipment including cameras, sensors, mics, speakers, and GPS units can be seamlessly integrated with cyberspace thanks to CCPSs. This makes it possible for a variety of new systems or applications that need to follow a patient's location, including patient monitoring or health monitoring. In order to gather, detect, analyse, and disseminate enormous amounts of medical and user-location data for intricate processing, these systems combine a large number of physical devices, such as sensors, with localisation technologies (e.g., GPS and wireless local area networks). However, these systems have a number of issues, including communication, massive processing, universal access, and patient placement. In order to support massive real-time data communications and processing in the cyber or cloud environment, a scalable, universal infrastructure or system is needed. Using cellphones to gather speech and electroencephalogram signals, this research suggests a scalable, quick, and effective cloud-supported cyber-physical localisation system for patient monitoring. The recommended strategy has been shown to beat other comparable approaches in terms of error estimates since it localises using Gaussian mixture modelling.

Existing System

Modern diabetes monitoring systems and applications may now be developed and deployed, mostly due to recent breakthroughs in wearable computing, artificial intelligence, big data, and wireless networking technologies, including 5G networks, the Internet of Things and big data analytics. Due to the long-term and systemic harm that diabetics experience, effective guidelines for the diagnosis and treatment of diabetes must be developed.

Disadvantages of Existing System

Real-time data collection is challenging, and the setup is cumbersome. Diabetic patients' multidimensional physiological markers are not consistently tracked. For the management of diabetes, including its management and treatment, there are no present recommendations. A system for data exchange and individualised analysis of massive volumes of data from numerous sources, including lifestyle, sports, food, and other characteristics, is lacking in the diabetes detection model.

Proposed System

In this paper, the model leverages reasonably priced 5G technology to monitor the health of people with diabetes. Because of their busy professions or unhealthy lives, many people these days are diagnosed with diabetes, but they are completely kept in dark of it until they have symptoms or they go to a doctor for diagnosis. Before that, the illness will already be advanced, making prediction impossible. This study uses a range of techniques, including ensemble algorithms, decision trees, SVM, and ANN.

Advantages of Proposed System

The technology in this paper, makes it possible to accurately, affordably, and sustainably detect diabetes. Here, the data exchange mechanism is very efficient for both the social and data spaces.

System Architecture

The above architecture depicts the various actors and the different systems involved in the process. This paper has the main system named

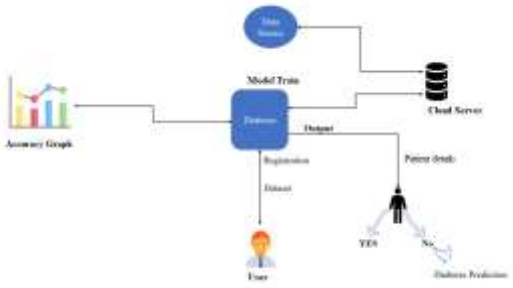

Fig. 1 System Architecture

Diabetes, where the models are generated for the project. Once the system is ready with the models, the accuracy graphs are generated and the system is ready to predict the results. The user will now, insert the datafile and the system will predict the results based on the values.

Main System (Diabetes):

This server records, trains, and generates an accuracy graph for the dataset model using a variety of techniques, including decision trees, SVM, ANN, and ensemble algorithms.

Test case No.	Test case Name	Description	Expected O/P	Actual O/P	Status
1.1	Preprocess the dataset	The admin uploads the dataset into the system	The dataset length is displayed.	The dataset length is displayed.	1
1.2	Preprocess the dataset	The admin not uploading any dataset	The dataset length is not displayed.	The dataset length is not displayed.	1
2.1	Generating accuracy graphs	Once the algorithms are performed on the dataset, graphs are generated	The accuracy graphs are displayed.	The accuracy graphs are displayed.	1
2.2	Generating accuracy graphs	Directly hold onto the accuracy graphs button	Nothing is displayed.	Nothing is displayed.	1
3.1	Start the cloud server	Once the model is generated, the cloud server is started	The work frame for the clod server is displayed.	The work frame for the clod server is displayed.	1

Cloud Server :
The test data is uploaded into this application, transferred to the primary system, and the output data is sent to
this application. Since there are no sensors in this paper to collect data, the test data that is uploaded is taken as sensitive data. Users don't need to disclose their details in this paper because we use the local host.

Modules
Data exploration: The first and foremost step in data analysis is exploration of the data, which involves visualising and examining data to find patterns or areas that require more research. With the use of a user-friendly interface, anyone inside an organisation may become familiar with the data, identify trends, and generate clever queries that could motivate additional in-depth, helpful study. In this module, we upload the dataset and make decisions.

Data Processing: Data preprocessing, which is a part of data preparation, deals with any sort of processing carried out on raw data to prepare it for the next action. Since ages, it has been an important first step in the data mining process.

Splitting data into train & test: The data is divided into training and testing data in a 9:1 ratio using this module.

Model generation: Through the use of machine learning, a computer programme can learn on its own and choose whether to complete a task without any intervention of humans. When a programmer incorporates machine learning capabilities into a software programme to attempt and evaluate a broad variety of different techniques, they do it in accordance with a set of rules known as the machine learning model generation process, tools, and settings in order to precisely acquire the desired result. In this situation, the generated model can give us the most accurate results.

Testcases:

Algorithms:
Random Forest, SVM, ANN, XgBoost, AdaBoost, Ensemble algorithms are used in this paper.

Random Forest: It operates by combining a huge number of decision trees during the training phase, with the class coming from either classification or regression that represents the mean forecast of all the trees combined. Due to the fact that it can give feature importance metrics to help explain the predictions provided by the model, it is renowned for its excellent accuracy, usability, and interpretability.

SVMor Support Vector Machine: By maximising the margin, or distance, between the nearest points in each class, this potent and extensively used machine learning technique determines a decision boundary between the classes. One of SVM's primary benefits is its ability to handle both linear and non-linear data by converting the input into a higher-dimensional

space where a linear decision limit may be constructed.

ANN or Artificial Neural Network: It is a machine learning method which is depended on how the human brain works and is organised. One of ANNs' major benefits is their capacity to learn from huge, complicated datasets and generate predictions based on brand-new, unforeseen data.

XGBoost or eXtreme Gradient Boosting: It is a method of ensemble learning that combines the results of several weak models, often decision trees, to yield a more reliable and accurate result. Additionally, it offers feature importance metrics that can be used to justify the model's predictions.

AdaBoost, or Adaptive Boosting: It operates by repeatedly training weak models using examples from the incorrectly categorised previous models.Depending on how well it performs, each weak model is given a weight, with models that perform well receiving more weight and models that perform poorly receiving less. The final forecast is then created by adding all the weak model predictions and weighting them according to their accuracy.

Ensemble Algorithm: An ensemble algorithm is a machine learning technique that aggregates the results of various models to create a single, more reliable prediction. The premise behind ensemble algorithms is that by integrating the results of various models, each with their own advantages and disadvantages, the final result will be a more accurate and reliable prediction than any one of themodel alone.

Fig. 3 Accuracy of the models

The first picture tells about the accuracy score of the algorithms after the model is trained. In this way, only the algorithms which give us a high accuracy score are chosen for further predictions. Here, the SVM with GridSearch model has the highest accuracy among the other models. Therefore, this model analysis the inputs and results in the most accurate prediction.

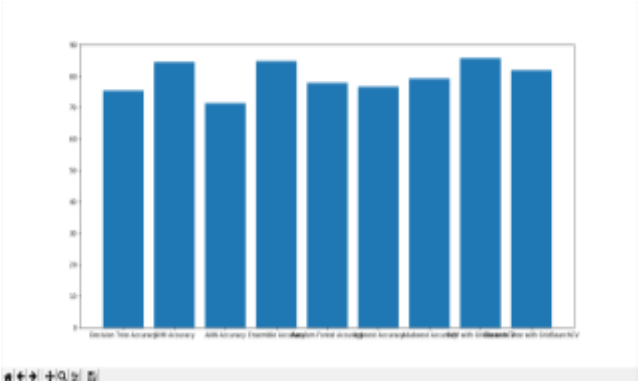

Fig. 4 Graphs of different models

The second picture is about visualising the data in the form of graphs. Here, the difference between the accuracies of the various models is cleared depicted.

Fig.5 Results of the sample input data given by the user

The third picture in this paper shows the predictions displayed after the user gives certain inputs. This model certainly tells whether the patient is diagnosed with Diabetes or not and it also mentions the type of diabetes.

2. Conclusion

In this study, a system for 5G-Smart Diabetes is initially proposed that can identify the type of diabetes that a patient will have. Compared to earlier methods, this one can provide intelligent, affordable, and ongoing diabetes diagnosis. This software is more secured and the data is end-to-end encrypted. In conclusion, cloud-based solutions offer a promising avenue for the diagnosis of diabetes. By leveraging the power of cloud computing, healthcare providers can securely store, access, and analyse huge volumes of patient data, including imaging studies, test reports, and electronic health records. This allows for a more comprehensive and accurate assessment of a patient's risk for developing diabetes or their current diabetic condition. Cloud-based solutions can also facilitate remote patient monitoring and telemedicine, allowing for real- time feedback and intervention to improve diabetes management. Additionally, cloud- based solutions can enable collaboration between healthcare providers and researchers, accelerating the development of new treatments and therapies for diabetes. Overall, the use of cloud-based solutions in diabetes diagnosis and management has the potential to enhance the quality of care, reduce costs, and improve outcomes for patients.

Future Scope

The future scope for the diagnosis of diabetes using cloud-based solutions is vast, and there are several potential areas where it could be further developed and applied:

Artificial intelligence and machine learning: By using AI and machine learning algorithms to find patterns and trends in patient data, diabetes can be diagnosed earlier and with greater accuracy.

Wearable devices and IoT: Integration with wearable devices and the Internet of Things (IoT) can enable continuous monitoring of a patient's blood glucose levels, allowing for real-time adjustments to their treatment plan.

Cloud-based clinical trials: Cloud-based solutions can enable large-scale clinical trials to be conducted remotely, allowing for the rapid collection and analysis of data to advance diabetes research.

Improved data security and privacy: With the growing concern over data privacy and security, cloud-based solutions could be further developed to ensure that patient data is kept secure and confidential.

Overall, the future scope for the diagnosis of diabetes using cloud-based solutions is promising, and further development and research in this area could result in significant improvements in diabetes management, early detection, and patient outcomes.

3. References

S. Mendis, “Global Status Report on Noncommunicable Diseases 2014,” WHO, tech. rep.; http://www.who.int/ nmh/ publications/ncd-status-report-2014/en/, accessed Jan. 2015.

F. Florencia et al.,IDF Diabetes Atlas, 6th ed., Int’l. Diabetes Federation, tech. rep.; http:// www.diabetesatlas.org/, accessed Jan. 2016.

M. Chen et al., “Disease Prediction by Machine Learning over Big Healthcare Data,” IEEE Access, vol. 5, June 2017, pp. 8869--79.

O. Geman, I. Chiuchisan, and R. Toderean, “Application of Adaptive Neuro-Fuzzy Inference System for Diabetes Classification and prediction},” Proc. 6th IEEE Int’l. Conf. E-Health and Bioengineering, Sinaia, Romania, July 2017, pp. 639--642.

S. Fong, et al. “Real-Time Decision Rules for Diabetes Therapy Management by Data Stream Mining,” IT Professional, vol. 26, no. 99, June 2017, pp. 1--8.

B. Lee, J. Kim, "Identification of Type 2 Diabetes Risk Factors Using Phenotypes Consisting of Anthropometry and Triglycerides Based on Machine Learning," IEEE J. Biomed. Health Info., vol. 20, no. 1, Jan. 2016, pp. 39--46.

M. Hossain, et al., "Big Data-Driven Service Composition Using Parallel Clustered Particle Swarm Optimization in Mobile Environment," IEEE Trans. Serv. Comp., vol. 9, no. 5, Aug. 2016, pp. 806--17.

M. Hossain, "Cloud-Supported Cyber- Physical Localization Framework for Patients Monitoring," IEEE Sys. J., vol. 11, no. 1, Sept. 2017, pp. 118--27.

P. Pesl, et al., "An Advanced Bolus Calculator for Type 1 Diabetes: System Architecture and Usability Results," IEEE J. Biomed. Health Info., vol. 20, no. 1, Jan. 2016, pp. 11--17.

M. Chen et al., "Wearable 2.0: Enable Human-Cloud Integration in Next Generation Healthcare System," IEEE Commun. Mag., vol. 55, no. 1, Jan. 2017, pp. 54--61.