ABSTRACT:

The world's most deadly disease, Diabetes Mellitus, is among the fastest growing. Medical professionals are looking for a technology that can predict diabetes properly. To study data from diverse perspectives and synthesise it into useful knowledge, various machine learning algorithms can be applied. With the application of appropriate data mining techniques to large amounts of data, we will be able to obtain valuable information. The main objective is to find new patterns and then provide relevant and helpful information to consumers by analysing them. Diabetes is linked to cardiovascular disease, renal problems, nerve damage, and blindness. Data mining is critical to deal with while doing diabetes research. Methods and strategies for efficiently classifying and detecting patterns in the Diabetes dataset will be discovered using data mining techniques and methods. The purpose of this study was to predict diabetes by using medical bioinformatics. The WEKA programme was used as a diagnostic mining tool. The University of California at Irvine provided the Pima Indian diabetes database, which was utilised in the research. A model based on the data was developed for accurately predicting and diagnosing diabetes. This study compares the performance of Nave Bayes, Decision Trees, and (KNN) with bootstrapping resampling to improve accuracy.

Diabetes has spread throughout the world, affecting people of all ages, regardless of their age. As diabetes patients increase, it is due to numerous factors such as bacterial or viral infections, chemical or toxic substances in food, auto immune reactions, obesity, poor diet, lifestyle changes, eating habits, pollution, etc. To save a person's life, it is essential to identify diabetes in a timely manner. Data analytics is the process of analysing and identifying hidden patterns in large amounts of data to draw conclusions. This analytical procedure is carried out in health care by employing machine learning algorithms to analyse medical data and develop machine learning models for medical diagnostics. This study explains how to diagnose diabetes using a diabetes prediction algorithm. Furthermore, this research investigates several machine learning algorithms and strategies for improving diabetes prediction accuracy utilising medical data. Diabetes Mellitus (DM) is a group of metabolic disorders that afflict millions of people worldwide. Diabetes must be detected early in order to prevent serious complications. There have been numerous research studies on diabetes diagnosis, the majority of them rely on a single data source, the Pima Indian diabetes data set. A collection of studies on Indian women dating back to 1965 has been collected as part of the Pima Indian data set, and it has a relatively high rate of diabetes onset. The majority of prior research studies focused exclusively on one or two specialised sophisticated procedures for data testing, with no comprehensive research on multiple generic techniques. This study explores a number of Machine Learning techniques (e.g., the KNN algorithm) for the purpose of identifying diabetes and methods used for data pre-processing. Testing the accuracy of this technique will utilize the UCI ML repository data set.

Keywords—Machine Learning, Diabetes, Classification, K-nearest neighbours, Decision Trees, Naive Bayes.

INTRODUCTION

Diabetes is characterised by elevated blood sugar levels, which manifest as frequent urination thirst, appetite, and weight loss. Diabetes is diagnosed when blood glucose levels exceed 200 mg/dL two hours after loading, as well as numerous diabetes research studies that necessitate fast call detection. Patients with diabetes typically require ongoing therapy; otherwise, they risk a variety of life-threatening complications.

Early detection and treatment of diabetes can help you avoid complications. For diabetes detection, the suggested method employs machine learning algorithms. The suggested system will be a medical field application that will aid diabetic doctors and patients in recognising diabetes. The suggested approach automates the detection of diabetes using data from previous diabetes patients. People of all ages are affected by diabetes, which is a rapidly spreading disease [1]. The presence of too much sugar (glucose) in the blood causes diabetes. The two types of diabetes are type 1 and type 2. Diabetes type 1 is an autoimmune disease. In this situation, the body destroys the cells required to create insulin and absorb sugar for energy production. Obesity has no bearing on this type of cancer. A person's body mass index defines the obesity which is above his/her normal threshold [2]. Type 1 diabetes can strike at any age, including infancy and puberty. Type 2 diabetes is more likely to develop in obese individuals. The body either opposes or fails to manufacture insulin in this condition. Type 2 is more common in middle-aged and older people [1]. Other causes of diabetes include bacterial or viral infections, toxic or chemical components in food, auto immunity reactions, obesity, inadequate diets, lifestyle changes, exposure to pollution, etc. Diabetes causes a number of ailments, including cardiovascular difficulties, renal problems, retinopathy, and foot ulcers [1]. All over the world, people are affected by diabetes, which is a serious condition. Worldwide, chronic illnesses like this lead to many deaths among adults. Chronic illnesses increase costs as well. A large percentage of government and individual budgets is spent on chronic diseases [3,4]. Diabetes affected 382 million people globally in 2013 [5]. In 2012, it was the eighth biggest cause of mortality for both men and women. Diabetes is more likely to occur in high-income countries [6].

Diabetes diagnosis is regarded as a difficult subject for quantitative research. Due to constraints, the use of numerous criteria such A1c [7], fructosamine, white blood cell count, fibrinogen, and haematological indices [8] are considered ineffective. These criteria were employed in various research investigations to diagnose diabetes [[9], [10], and [11]. Chronic ingestion of booze, salicylates, and drugs have all been linked to an increase in A1C. When measured by electrophoresis, vitamin C consumption may raise A1c levels, but when measured by chromatography, levels may appear to decrease [12].

MATERIALS AND METHODS

Following are the two main types in which diabetes can be further divided, i.e, Type-1 Diabetes & Type-2 Diabetes.

Insulin-dependent and insulin-independent diabetes are two different types of diabetes.

TYPE-1

The pancreas is assaulted by the immune system in diabetes type 1. Because type 1 diabetes patients cannot produce their own insulin, they must take insulin injections throughout their lives.

Type 1 diabetes primarily affects children and adolescents, while it can sometimes affect adults. As a result of the immune system attacking pancreatic beta cells, type 1 diabetes causes them to stop producing insulin

There is no way to prevent Type 1 diabetes as it is genetic. The affect Type-1 diabetes have on people with diabetes is just about 5 to 10 percent.

TYPE-2

Insulin does not act correctly or the body does not produce enough insulin in type 2 diabetes. Diabetes type 1 patients must take medication to keep their blood sugar levels under control. At any age, a person can develop type-2 diabetes.

The risk of type 2 diabetes has increased as people grow older, but it can still cause serious health problem in children. While insulin is produced by the pancreas, it is not effectively utilized by the body. According to it, lifestyle factors influence its evolution. The vast majority of people with diabetes have type 2 diabetes.

LITRATURE REVIEW

Diabetes analysis and prediction has become an area of tremendous relevance as the global diabetes rate has increased [13]. The machine learning models were used to The Pima Indian diabetes database by Saru, S., and S. Subashree [14]. With 10 cross-validation runs of their Naive Bayes, Decision Trees, and KNN models, they used bootstrapping resampling to forecast and compare their accuracy. The proposed methodology was determined to have an accuracy of 90.36 percent. Alkaragole, Mohammed Layth Zubairi, and Sefer Kurnaz [15] investigated the accuracy of different data-mining strategies, including decision trees, Naive Bayes, SVM, and hybrid algorithms. With an accuracy of 94% and sensitivity of 91%, hybrid techniques (proposed ensemble SVM + decision tree with 100 iterations) surpassed all other algorithms.

Sneha N. and Tarun Gangil [16] investigated various classification algorithms in order to determine the best diabetes predictor. The study used five classification algorithms: random forest, KNN, decision tree, Naive Bayes, and SVM, which were all obtained from the UCI machine repository archive. The best accuracy was 82.3 percent for Naive Bayes. Aada, A., and Sakshi Tiwari [17] analysed the PIMA Indian diabetes dataset using

KNN, Naive Bayes, and decision trees, as well as bootstrapping-like approaches. After bootstrapping, SVM has the highest accuracy of 94.44 percent. On the Pima Indians dataset, Srivastava and Suyash [18] used machine learning algorithms and artificial neural networks to predict diabetes.

The PIMA Indian diabetes dataset was used in a study by Kaur, Harleen, and Vinita Kumari [19] to predict and assess diabetes trends. For prediction, they employed R data processing software and five algorithms: SVM-linear, radial basis function kernel support vector machine, k-nearest neighbor, artificial neural network, and multi-factor dimensions reduction. For diabetes prediction, the SVM-linear model had the highest accuracy of 89 percent. Md Maniruzzaman used the dataset of diabetes from the survey done by the National Health and Nutrition Examination [20], which included 6561 people, 657 of whom were diabetic. For diabetes prediction, logistic regression, Naive Bayes, decision trees, AdaBoost, and random forest were used. With logistic regression as feature selection and random forest for classification, the greatest accuracy of 94.25 percent was achieved.

Naive Bayes, decision tree, J48, and random forest with 10 fold cross-validation were used by Prasad, K.S., Reddy, N.C.S., and Puneeth, B.N. [21]. The most accurate method was Random Forest, whereas Naive Bayes had the lowest mean absolute error and root mean squared error.

EXISTING DATA MINING TECHNIQUES

Decision Tree:

The most widely used data mining technique is the decision tree. One of the most popular and straightforward classifiers is the decision tree. Data mining uses decision trees as predictive models. In this study, a decision tree is used with a classification algorithm to predict disease using patient data.

Decision trees are simple to design and understand. Decision tree-based prediction is well-organized. It is capable of handling large amounts of extent data. It is better suited to knowledge discovery searches. Finally, the Decision Tree findings are easier to understand and read [22].

Naive Bayes:

The Naive Bayes algorithm is based on the Bayes theorem and involves conjecture between predictors. The Naive Bayes allows you to easily develop models with predictive capabilities as well as a new way of navigating and analysing data. When using Naive Bayes to develop a predictive model, this technique can be applied to predictive analysis. All input attributes must be comparably independent for this to work. A Naive Bayesian model is simple to build and does not have any complex iterative parameters. Naive Bayes is an effective predictor. For really big data sets, this strategy is quite beneficial. Before and after reducing of characteristics, Naive Bayes performs consistently with the same representation of construction time. The visualisations for Naive Bayes are straightforward [23].

k-nearest neighbour algorithm (k-NN):

The k-nearest neighbour algorithm (k-NN) is an important approach to classifying objects based on the nearest training data in the feature space. Although it is the simplest of all machine learning rules, the accuracy of the K-NN rule is harmed by the availability of screaming choices.

Classification

Clustering:

Diabetes can be predicted using classification as one of the methods. Classification is the most important aspect of data mining. Classification is common in large amounts of corporate and medical data sets. Classification is a data mining function that divides a collection of items into desired categories. Classification is the process of dividing situations into groups based on a predictable characteristic. Each case has a number of attributes, one of which is the class attribute, also known as the predictable attribute. Finding a model that defines the class attribute as a function of input attributes is part of the work. Classification measures with data mining tools with detecting the problem by separating the aspects of diseases between patients and diagnosing or predicting which algorithm performs best [24]. Data is classified in order to determine the class that each case belongs to accurately. Before storing the data into the classifying model, this technique can be utilised as a pre-processing step. To locate a set of data, we must cluster them in order to collect everything based on their features. And grouping them together based on their similarities. Segmentation is another term for clustering. It's used to find unprocessed case groupings based on a set of criteria. Cases with more or less similar attribute values inside the same category. Clustering is a data mining activity that is done individually. It means that by using a single attribute you cannot model the training process. All input attributes are given equal weight. Before clustering, the attribute values must be defined to prevent higher value characteristics from impacting lower value attributes.

Neural Network:

Artificial neural networks are well adapted to issues that human are not adept at addressing, such as pattern recognition and prediction. In the medical field, neural networks have been applied to diagnosis, picture interpretation, signal interpretation, and drug development [25].

METHODOLOGY

The mechanized diabetes testing system is, which is used to consistently record the blood glucose levels of diabetics, especially in the intensive care unit. The observation frame records glucose levels via a blood glucose sensor. Information will be sent to the Specialist via WIFI. This frame is the moment, which indicates the blood glucose level regardless of whether glucose is falling or rising. The framework provides accurate results and patient information is updated daily in the cloud.

ARDUINO UNO ATMEGA 328P:

Arduino microcontroller is something however tough to utilize, open supply, and its device is smart period. The Arduino ATMEGA 328 is a well-known microcontroller chip created with the aid of using Atmel. It is an 8-piece microcontroller that has 32K of glimmer memory, 1K EEPRON and 2K of indoors SRAM of it has 14 superior information/yield pins wherein 6 may be applied as PWN yields and 16MHz earthenware resonator, an ICSP header, the USB association, 6 easy statistics sources, a force.

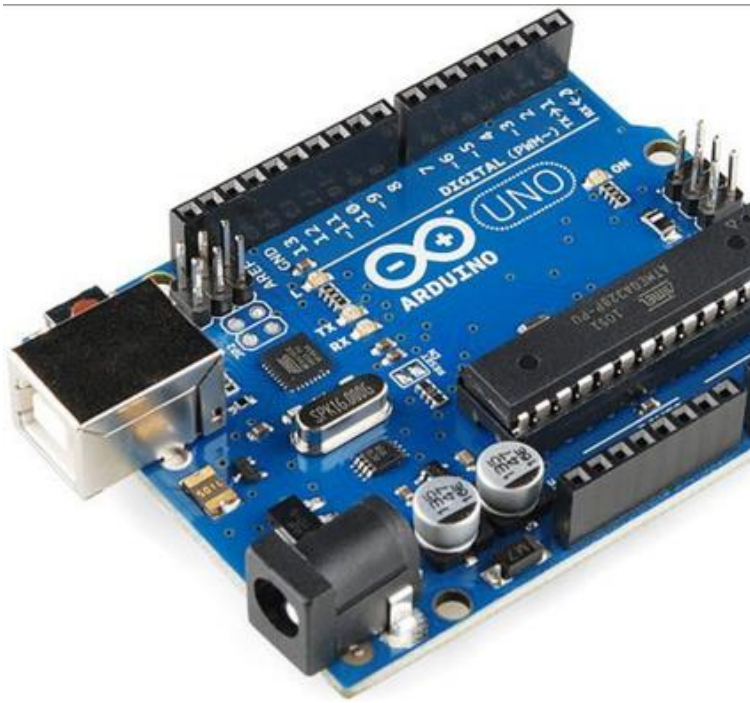

Figure 1: ARDUINO UNO ATMEGA 328

OUTCOME

HUMIDITY IN ADDITION TO TEMPERATURE SENSOR:

This is a focal temperature and viscosity sensor-driven with very little effort. Capacitive humidity sensors and thermistors are used to check for trapped air. Then it outputs a screen signal that passes through the material pin. BMP 280 Pressure Sensor: This sensor has high accuracy and ease of use, making it the perfect solution for accurate pressure

estimation. Weight varies with height and the weight estimate is very accurate.

BMP 280 PRESSURE SENSOR:

With its high accuracy and ease of use, this sensor is the perfect solution for accurate pressure estimation. Weight varies with height and the weight estimate is very accurate.

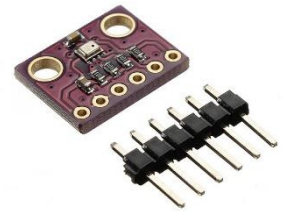

Figure 2: BMP 280 PRESSURE SENSOR

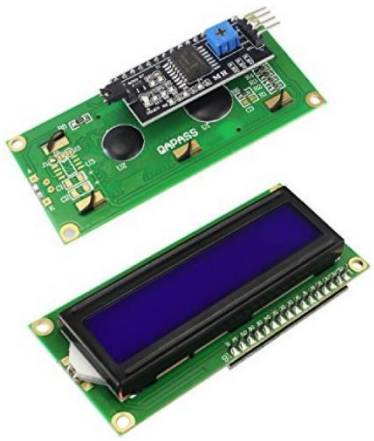

Figure 3: LCD DISPLAY

GAS SENSOR (FIGARO TGS822):

It is a electricity display case module that uses Liquid Gem to provide a clear image to the. This is usually a required module Used in DIY and circuits. 16 * 2 uses a platform with 16 fonts for each of these two lines.

Figure 4: GAS SENSOR (FIGARO TGS822)

WORKING:

The gas sensor detects CH3) 2CO from the respiration of the human body. This is sent to the data distribution unit of the effort module where the Arduino is located. By, Arduino yields will be proven to LCD or distant zone authorities. The BMPP 280 Weight Sensor measures temperature, weight, and humidity. The sensor gets the yield because the simple sign is the yield. This yield was switched to automation and was displayed on the LCD using the Arduino ATMEGA328P.

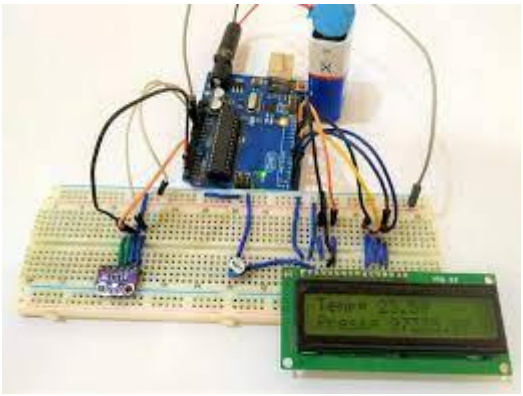

Figure 5: DISPLAYING THE PROJECT MODEL

CONCLUSION

The purpose of this paper is to examine symptoms of diabetes and gather information that can be used to assist healthcare professionals in detecting diabetes early and detecting its onset before it becomes a major problem. Data mining uses techniques such as feature selection to analyse data. Classification. All of these are used to analyse trends and forecast symptom severity. diabetes. ANOVA, Mutual Information, and other feature selection techniques to improve accuracy and reduce overhead, genetic algorithms were applied. The model's training time. Naive Bayes, SGD Classifier, KNN, Logistic Regression the Algorithms Random Forest, Decision Tree, and Support Vector Machine were used to diabetes prediction. A comparison of all the implemented algorithms was conducted by Random Forest had the highest accuracy of 93.95 percent when using Genetic Algorithm as a feature selection technique, with Cholesterol, Glucose, Chol/HDL, Systolic BP, Weight, and Hip as features and a random forest depth of 5. The ability to predict diabetes at an early stage depends on regular monitoring of patients' cholesterol, glucose, cholesterol/HDL, systolic blood pressure, weight, and hip. Diabetic complications can be prevented by early treatment of any abnormality in a number of the following measurements.

In today's world, diabetes prediction is critical, especially given its serious complications. Nowadays most people die due to diabetes. The System model focuses on detecting diabetes using only a few parameters. Physicians can utilise the system to anticipate diabetes in the early stages. So that patients can receive traditional treatments and remedies. The system used techniques like machine learning (ML) for prediction to give more exact results. There has been a lot of research regarding the diabetic imprint. It is important for hospitals and clinics to develop a prediction system to predict diabetes.

REFRENCES

[1] Kaveeshwar, S.A., and Cornwall, J., 2014, "**The current state of diabetes mellitus in India**". AMJ, 7(1), pp. 45-48.

[2] Dean, L., McEntyre, J., 2004, "**The Genetic Landscape of Diabetes [Internet].**
Bethesda (MD): National Center for Biotechnology Information (US);.
Chapter 1, Introduction to Diabetes. 2004 Jul 7.

[3] D. Falvo, B.E. Holland **Medical and psychosocial aspects of chronic illness and disability Jones & Bartlett Learning** (2017)

[4]
J.S. Skyler, G.L. Bakris, E. Bonifacio, T. Darsow, R.H. Eckel, L. Groop, *et al.*
Differentiation of diabetes by pathophysiology, natural history, and prognosis
Diabetes, 66 (2017), pp. 241-255

[5] Z. Tao, A. Shi, J. Zhao **Epidemiological perspectives of diabetes** Cell Biochem Biophys, 73 (2015), pp. 181-185

[6] W.H. Organization **World health statistics 2016: monitoring health for the SDGs sustainable development goals** World Health Organization (2016)

[7] L. Cobos **Unreliable hemoglobin A1C (HBA1C) in a patient with new onset diabetes after transplant (nodat)** Endocr Pract, 24 (2018), pp. 43-44

[8]
B. Dorcely, K. Katz, R. Jagannathan, S.S. Chiang, B. Oluwadare, I.J. Goldberg, *et al.*
Novel biomarkers for prediabetes, diabetes, and associated complications Diabetes, Metab Syndrome Obes Targets Ther, 10 (2017), p. 345

[9]
P.P. Singh, S. Prasad, B. Das, U. Poddar, D.R. Choudhury **Classification of diabetic patient data using machine learning techniques** Ambient communications and computer systems, Springer (2018), pp. 427-436

[10] A. Negi, V. Jaiswal **A first attempt to develop a diabetes prediction method based on different global datasets** 2016 fourth international conference on parallel, distributed and grid computing, PDGC) (2016), pp. 237-241

[11]
N. Murat, E. Dünder, M.A. Cengiz, M.E. Onger **The use of several information criteria for logistic regression model to investigate the effects of diabetic drugs on HbA1c levels** Biomed Res, 29 (2018), pp. 1370-1375.

[12] M.S. Radin **Pitfalls in hemoglobin A1c measurement: when results may be misleading** J Gen Intern Med, 29 (2014), pp. 388-394.

[13]. Zhou, Bin, et al."**Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4· 4 million participants**." The Lancet 387.10027 (2016): 1513-1530.

[14] . Saru, S., and S. Subashree. "**Analysis and prediction of diabetes using machine learning.**" International Journal of Emerging Technology and Innovative Engineering 5.4 (2019). \

[15]. Alkaragole, Mohammed Layth Zubairi, and Sefer Kurnaz. "**COMPARISON OF DATA MINING TECHNIQUES FOR PREDICTING DIABETES OR PREDIABETES BY RISK FACTORS**." (2019).

[16]. Sneha, N., and Tarun Gangil. "**Analysis of diabetes mellitus for early prediction using optimal features selection.**" Journal of Big data 6.1 (2019).

[17]. Aada, A., and Sakshi Tiwari. "**Predicting diabetes in medical datasets using machine learning techniques.**" Int. J. Sci. Eng. Res 5.2 (2019).

[18]. Srivastava, Suyash, et al."**Prediction of Diabetes Using Artificial Neural Network Approach.**" Engineering Vibration, Communication and Information Processing. Springer, Singapore, 2019. 679-687

[19]. Kaur, Harleen, and Vinita Kumari. "**Predictive modelling and analytics for diabetes using a machine learning approach.**" Applied computing and informatics (2020).

[20]. Maniruzzaman, Md, et al."**Classification and prediction of diabetes disease using machine learning paradigm.**" Health Information Science and Systems 8.1 (2020).

[21]. Prasad, K.S., Reddy, N.C.S. & Puneeth, B.N. **A Framework for Diagnosing Kidney Disease in Diabetes Patients Using Classification Algorithms**. SN COMPUT. SCI. 1, 101 (2020).

[22] Isha Vashi, Prof. Shailendra Mishra,"**A Comparative Study of Classification Algorithms for Disease Prediction in Health Care**", International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 9, September 2016.

[23] K. Priyadarshini1 , Dr.I.Lakshmi2 "**A Survey on Prediction of Diabetes Using Data Mining Technique**" International Journal of Innovative Research in Science, Engineering and Technology (An ISO 3297: 2007 Certified Organization) Vol. 6, Special Issue 11, September 2017.

[24] Nilesh Jagdish Vispute, Dinesh Kumar Sahu, Anil Rajput, "**ASurvey on naive Bayes Algorithm for Diabetes Data Set Problems**", International journal for research in Applied Science & Engineering Technology (IJRASET),Volume 3 issue XII,December 2015.

[25] WEBSOURCE: https://www.researchgate.net/publication/273023827**_PREDICTION_OF_DIABETES_MELLITUS_USING_DATA_MINING_TECHNIQUES_A_REVIEW**