Mathematical Statistician and Engineering Applications
ISSN: 2094-0343
2326-9865

Data Mining based Predictive Analysis of Diabetic Diagnosis in Health Care: Overview

M.Shanmugavalli[1],
Research Scholar,
Department of Mathematics,
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences (SIMATS),
Saveetha University,
Chennai, Tamilnadu-602105.
shanmugavallim41012.sse@saveetha.com

K. Sivakumar[2],
Professor,
Department of Mathematics,
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences (SIMATS),
Saveetha University,
Chennai, Tamilnadu-602105.
sivakumarkaliappan.sse@saveetha.com

Article Info

Page Number: 572 - 588

Publication Issue:

Vol 71 No. 4 (2022)

Abstract

The complexity of modernizing the healthcare industry's journey toward processing huge health data and accessing them for analysis and action will be considerably increased. Health research breakthroughs have resulted in a substantial quantity of data being generated from enormous electronic health records, high-throughput genomic data, for example as well as the collection of clinical information. The healthcare business confronts several obstacles, emphasizing the significance of data analytics development. In biosciences, machine learning and data mining methods are becoming more crucial in attempts to effectively convert all available data into valuable information. This is the goal of these approaches. mining techniques in today's medical studies. An overview is the goal of this work. The condition known as Diabetic Mellitus affects millions of individuals worldwide (DM). There were a lot of clinical datasets that were utilised. New hypotheses targeted at greater understanding and future study in DM may be derived from the title applications in the chosen works. According to the findings, this strategy is an effective method to treat and care for patients while also improving outcomes like as cost and availability.

Article History

Article Received*: 25 March 2022*

Revised*: 30 April 2022*

Accepted*: 15 June 2022*

Publication*: 19 August 2022*

Keywords*:* Machine Learning, Data Mining, Diabetes Mellitus (DM), Diabetic complications, Disease prediction models.

INTRODUCTION

We need to arrange and highlight the scale of big data in the healthcare industry into a nominal value with a possible solution due to its increasingly unstructured nature in general. There is a lot of promise for identifying hidden patterns in medical data sets via medical data mining. To make clinical diagnosis, these patterns might be exploited. There is a huge amount of raw medical data that may be accessed. This information must be gathered in a systematic manner. It is feasible to utilise this

information to develop a hospital information system. [2] Data mining is a user-friendly technique for identifying new and hidden patterns in data.
Data mining and machine learning are algorithm-based methods for extracting patterns from large volumes of data. As a result of the procedure, new information may be gained, but one of the most significant expenditures, data collection, is avoided. Having high blood sugar levels is a symptom of diabetes Mellitus, which is also known as diabetes. When blood sugar enters your cells, insulin acts as a conduit to move it from the bloodstream to where it may either be stored or used as a source of energy. Insulin deficiency or inability to correctly use the insulin your body produces might put you at risk of getting diabetes. Thus, diabetes has become one of the leading medical study subjects, resulting in a huge volume of data. Diabetic high blood sugar has been linked to nerve, ocular, and kidney damage if left untreated [4].

Following is a breakdown of the many types of diabetes: An autoimmune disease, type 1 diabetes is just what it sounds like. Insulin-producing cells in the pancreas are particularly targeted and killed by the immune system. [1, 2]. Around 10% of diabetics are affected by this kind of diabetes. By becoming insulin resistance and increasing blood sugar levels, Type 2 diabetes may occur. [6]. While blood sugar levels are elevated in prediabetes, they are not high enough to warrant the diagnosis of type 2 diabetes. (c) When diabetes develops during pregnancy, it is known as gestational diabetes or GD for short. When the placenta produces insulin-blocking proteins, it causes type 2 diabetes in pregnancy. In spite of its name, diabetes insipidus differs from type 2 diabetes in that it is a rare condition. If you have this condition, your kidneys aren't working properly, and you're losing a lot of fluid. Diabetic conditions have distinct signs, causes, and treatment options.
According to the findings of a large-scale survey done across India, more than 36% of those with diabetes in 2020 were over the age of 60. Notably, over 4% of those aged 20 to 29 reported having diabetes that year. This was a concerning trend that was associated with an unsustainable diet. Our survey's purpose is to discover the best model that predicts and analyses diabetics in advance, which will be patient-friendly. Large amounts of diabetes-related data may most effectively be mined and learned from using data mining and machine learning technologies. The approach employs associative and grouping techniques, as well as predictive data mining tools. In many respects, these data-mining techniques depart from classic statistical methodologies. The course of information mining is undeniably more troublesome than that of measurable procedures, which is one of the key differences between them.

I. RELATED WORK

Study and prediction were performed on healthcare datasets using various approaches and techniques, as part of the analysis of related work in this. Prediction models have been constructed using data mining techniques, machine learning algorithms, or a combination of these approaches. Fuzzy logic and an artificial neural network (ANN) were used by HumarKahramanli and NovruzAllahverdi (2008) to forecast diabetes. [3]. Aljumah et al. developed a diabetes intervention and analysis model using classification algorithms. The model was built with Support Vector

Machine and executed with Oracle Data Miner. The model's dataset was obtained from the World Health Organization (WHO). This research found that smoking is a major contributor to diabetes [7]. Many survey studies have the goal of conducting a comprehensive review of data-mining methods' uses in diabetes research [8]. The availability of patients' medical data has prompted physicians and patients to seek alternative computer-based evaluation tools to aid decision-making (Jyoti Soni et al. 2011)[10]. Analytical data from a number of patients who have the same ailment, for example, may be compared by doctors and compared to data from different regions of the country (K.Srinavas et al. 2010) [9].

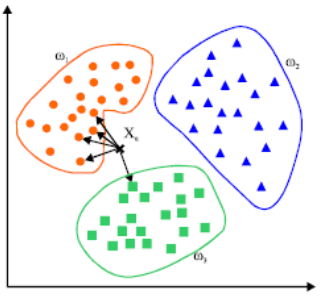
II. DATA MINING TECHNIQUES

The term "Data Mining" refers to a relatively new technique to data analysis and knowledge discovery that arose in the middle of the 1990s. However, as an interdisciplinary topic, data mining has its roots primarily in the fields of statistics and machine learning, it has now expanded to cover pattern recognition, database design, artificial intelligence, visualization, and other topics. There are two sorts of data mining approaches that may be employed efficiently: descriptive data analytics and predictive data analytics. To help consumers make sense of massive amounts of data, descriptive data analytics uses a method known as record similarity to discover previously undiscovered patterns or correlations within the data. It's all about nature in descriptive data analytics. Clustering, association, summarization, and sequence finding are all examples of this form of data mining. Training data is used to create a classification or model, which is then used in the analysis of unclassified data. There is still a lot of room for advancement in the field of predictive analytics, often known as predictive modelling and machine learning. Many different models and algorithms may be used in predictive analytics tools, which can be used to a broad range of applications. Classification, Regression, Time series analysis, and Prediction are some models of Predictive data analytics. Machine learning uses statistical concepts to enable machines to learn without explicit programming. The following are some key algorithms that have been defined.

*A. Decision Tree(DT):*If you want to make a decision, you may use a decision tree. Design, model, or representations that follow a tree-like structure have been shown to be accurate in conclusions. Machine learning is a technique that uses algorithms to learn about a target object and then make decisions based on that information. There are several applications for this kind of model, and it is employed in a wide range of fields. Each node (leaf node or terminal node) includes a class label, and the algorithmic flowchart represents a test on an attribute with each branch denoting a test conclusion. This ensures that the splitting is done according to the criteria. Decision-tree induction may be used to continuous (ordered) variables, rather than class labels. Regression trees and model trees are two of the most common types of prediction trees. As part of CART (Classification and Regression Trees), the regression tree keeps track of the average value of the predicted attribute for each set of training data. Model trees, on the other hand, have leaves that each include a regression mod-el multivariate linear equation for the predicted attribute.

B. K-Nearest Neighbors(KNN): An method known as K-Nearest Neighbors (KNN) was first proposed in the early 1950s, making it one of the oldest supervised learning algorithms in use today. The comparison of a test set to a training set that is comparable to the test set is how learning takes place. The training tuples are explained by n attributes. An n-dimensional pattern space is represented by each tuple. k-nearest neighbour classifiers look for the k training tuples in the pattern space that are the closest to an unknown training tuple. Their proximity to the k unknown training tuples makes them the k closest neighbours.

If two points or tuples have the same Euclidean distance, then$X_1 = (x_{11}, x_{12}, \dots, x_{1n})$ and $X_2 = (x_{21}, x_{22}, \dots, x_{2n})$ is $dist(X_1, X_2) = \sqrt{\sum_{i=1}^{n}(x_{1i} - x_{2i})^2}$



k-nearest neighbor

C. Naïve Bayes Classifier(NB):

Naïve Bayes is a statistical classifier which assumes no dependency between attributes. The class of unknown data sets is forecasted using Bayesian probability theory. Assumes that the existence of a given attribute does not depend on the presence of any other attribute in a class. An method known as the Nave Bayes Classifier is widely used.Bayes theorem is a tool for estimating future probabilities. $P(c_i|X)$ from $P(c_i)$, $P(X)$ and $P(X|c_i)$.Consider the following equation:

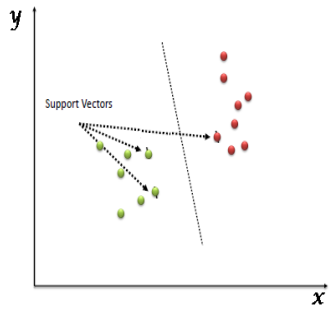
$P(c_i|X) = \frac{P(X|c_i)P(c_i)}{P(X)}$ where

- $P(c_i|X)$ is the likelihood of a particular class as predicted by a given predictor.
- $P(c_i)$ Is the probability that a class will take place.
- $P(X|c_i) = P(x_1|c_i) \times P(x_2|c_i) \times \dots \times P(x_n|c_i)$ is the probability of predicting given class, which is the likelihood.
- $P(X)$ Predictors' prior probability is a measure of how likely they are.

Naive bayes classification is simple and well-suited for high-dimensional data. Despite its simplicity, it can outperform more complex classification systems. This classifier is based on the following assumptions: It's important that the data be categorical and that each attribute occurrence is distinct, and the classifier must be able to predict effectively on large datasets.

*D. Support Vector Machines (SVM):*The Support Vector Machine can categorize both linear and nonlinear data. Using the support vector machine approach, the aim is to create a hyperplane in an N-

dimensional space (N being the number of qualities) that unambiguously classifies data points (SVM). When it comes to sorting data, decision boundaries such as hyperplanes may be quite useful. It is possible to classify data points that fall on each side of the hyperplane in a different way. In addition, the number of features affects the hyperplane's size. The hyperplane's location and orientation are affected by nearby data points, which is known as a support vector. We may increase the margin of the classifier by using these support vectors. Support vectors will alter the hy-placement. perplane'sIn order to improve our SVM, we need to consider the things listed above. Class boundary separation is maximised by optimization.



Hyperplane in Support Vector Machine

*E. Linear Regression:*It's a simple and widespread predictive analytic approach to anticipate the value of one variable depending on the value of the other. The response variable, y, and the simple predictor variable, x, are utilised in straight-line regression analysis. It's important y = b + wx to note that the constant variance of y is used to calculate the regression coefficients, which are used to calculate the line's slope and Y-intercept. More than one predictor variable may be included in multiple linear regression, which expands on straight-line regression's capabilities. As a result of linear regression, it is possible to reduce the differences between predicted and actual output values.

*F. Logistic Regression(LR):*Unsupervised learning method Logit is also known as Logistic Regression and is used to estimate the likelihood of a binary outcome in machine learning. One of two outcomes is all that can be determined in a binary situation. In predictive modelling, logical regression is used to analyse big datasets in which one or more independent variables might influence a result. Algorithms used in machine learning may categorise incoming data based on previous data. The result is stated as a binary variable with two potential possibilities. Improved algorithmic classifications are made when more relevant information is added to the dataset. As part of the extract, transform, and load (ETL) process, logistic regression may help to stage data for analysis by enabling data sets to be placed in specified stacks.

*G. Gradient Boosting(GB):*Gradient boosting is a common machine learning approach for datasets in tabular format. It is strong enough to detect any nonlinear connection between our model goal and features, and it is user-friendly enough to handle missing values, outliers, and large cardinality categorical values on your features without any additional treatment. It is a type of ensemble approach that involves combining numerous weak models to improve overall performance. So, it's easy to read their predictions and they're resistant to overfitting when utilizing an implicit feature selection

technique. Boosting algorithms are used to improve the accuracy of models in hackathons or contests. Despite the bagging learning technique, which creates models individually, gradient boosting creates models sequentially through iteration in order to reduce the error of previously learned models.

H. Random Forest(RF): The random forest is used to categorise data using a decision tree-based approach. In order to produce an interconnected forest of trees with a more accurate committee forecast than any individual tree, bagging and feature randomization are utilised. Decision trees are very sensitive to the data they're trained on, which is why even little alterations to the training set may have dramatic effects on the tree's design. In a random forest, each tree may be sampled at random with replacement from the dataset to generate unique trees. The word "bagging" is used to describe this process.

IV. ANALYSIS OF PAPERS

1.“Diabetes Prediction Using Data Mining Techniques”

This study uses the Naive Bayes classification technique to classify data that data set from the Fudawa health care center in Nigeria's Jos Plateau state. The data was pre-processed to remove duplicates. Weka tool was used to analyze noise and null fields, and the results were further refined. The front end of the application was built using the Java programming language, the Java development kit, and the Nave Bayesian Classifier. As a testing server, Wamp Server was used by the author to run Microsoft SQL (MySQL). There are two types of datasets: training and test. The following is a list of resources. For detecting and classifying the anomalies, parameters were employed. The characteristics that divide diabetes into good and negative categories are as follows: age, insulin, cigarette smoking, age first smoked, and survey location was taking place. During the course of this research, a novel diabetes prediction system was designed and implemented. This paper describes the data mining history and related concepts.

Remark: An application has been developed that uses a data mining approach of class comparison to forecast the incidence or recurrence of diabetes hazards. The naive Bayes classifier performs well and gave a prediction with less error and the efficiency of accuracy was 95%. The lack of parameters is a flaw in the study that can be remedied by discovering and integrating more constraints.

2.“Prediction of Diabetes Using Machine Learning Algorithms in Healthcare”

They address predictive analysis in healthcare by employing six distinct machine learning algorithms in this study effort, including DT, LR, RF, NB, KNN, and SVM. This study makes use of a bigger dataset that was acquired from the UCI machine learning system and comprises 768 records with 9 characteristics of females of Pima Indian ancestry. The major goals of this project are to improve prediction accuracy and performance. All six methods were analysed using Enthought Canopy, and the correctness of the dataset was determined using the standard measures. In this case, KNN and SVM have the best accuracy rate of 77%, which is the highest among the six methods. LR, NB, and DT, on the other hand, had an accuracy rate of 74%, while RF only had a 71% accuracy rate. As a result of this research, the conclusion is that the KNN and SVM are used to predict diabetes disease.

Remark:
The dataset's size and missing attribute values are two limitations of this study. Thousands of data with zero missing information will be required to construct a diabetes prediction model with 99.99% accuracy. The future researcher may concentrate on integrating alternative approaches into the utilised model in order to fine-tune model parameters for increased accuracy. Then, by putting these models to the test on a big dataset with few or no missing attribute values, further insights and improved prediction accuracy will emerge.

3."Predictive Analytics in Healthcare for Diabetes Prediction"
These findings are intended to help doctors discover Type 2 diabetes early and accurately diagnose the disease. They use bioinformatics theory and supervised machine learning approaches to increase diabetes prediction accuracy based on eight clinical indicators from the well-known PIMA dataset. They describe their technique and the stages involved in executing the program, as well as examining some of the most notable previous work in the subject. Furthermore, this research completely integrates well-known machine learning methods and compares the outcomes achieved by each technique. A common supervised learning issue is class imbalance. It occurs when the number of members in the majority and minority classes is significantly different, and it is especially prevalent in binary-valued classes, as illustrated in sample.The difference in variance between the two classes is significant, which may cause the classifiers' out-of-sample accuracy to suffer. They used all the machine learning and data mining algorithms like KNN, LR, DT, RF, Gaussian Naïve Bayes, Gradient Boosting, Keras Neural Network, Ada Boosting in their research work. The model's out-of-sample prediction accuracy of 89.94 percent, as shown by the F1 score of 0.853, shows its value in terms of its harmonic precision and recall, which is clearly strong against overfitting. Algorithms utilised for diabetes diagnosis were carefully investigated for performance and assessment.
Remark:
Blood Pressure and Pregnancy were shown to enhance the F1 score by 0.029 points when they were removed from the test and training sets (from 0.824 to 0.853). Even yet, in order to prevent overfitting, they did not eliminate any columns from their dataframe. Another interesting finding was that the gradient boosting technique prioritised the Skin Thickness value above the Blood Pressure attribute.

4."Machine Learning Classification Algorithms for Predictive Analysis in Healthcare"
The major contribution of this project is to develop the best and most appropriate method for illness prediction and diagnosis, as well as machine learning applications in healthcare systems. This article also includes an introduction of data science topics, ranging from data mining techniques to machine learning classification methods. The abstract highlights the essential parts of the entire research inside one paragraph typically. Machine learning methods applied in medicine were examined by a researcher in a study of medical research publications. Researchers use Decision Trees and Support Vector Machines in their healthcare prediction studies because they are the most accurate machine learning algorithms.
Remark: The researcher did not explain the sources of the conclusion in detail, which might be the

downside of this paper. The aforementioned algorithm is used to focus on disease detection, diagnosis, and prognosis.

5.”Diabetes Prediction using Machine Learning Algorithms”

It is our goal in this research to enhance the classification of diabetes by using a combination of a few extrinsic characteristics that are linked to diabetes and more conventional components such as blood glucose levels and BMIs. The new dataset has a better classification accuracy than the previous one. Furthermore, when a pipeline is enforced, the accuracy of the new dataset is improved when compared to the previous dataset. In addition, a pre-diabetes prediction pipeline model was applied in order to improve categorization. The dataset is classified using a number of machine learning approaches, with the greatest accuracy of 96% being achieved using Logistic Regression. Among the models tested, the AdaBoost classifier had the highest accuracy (98.8%). The accuracy of machine learning methods was evaluated using two different datasets.

Remark: To put it simply, the model improves the accuracy and precision of diabetes prediction significantly. This research might be expanded to see how likely non-diabetic persons are to develop diabetes in the coming years.

6.”Type 2 Diabetes Mellitus Prediction Model Based on Data Mining”

Improve the model's accuracy and adaptability to a variety of datasets are the key goals of this research. There are two parts to the model: the first part uses the revised K-means algo-rithm to delete erroneously grouped data, and the second part uses the improved dataset as input. We utilised logistic regression to classify the remaining data. In order to make comparisons with the findings of previous studies, researchers utilised data from the Pima Indians Diabetes Dataset and the Waikato Environment for Knowledge Analysis. According to the statistics, the model's prediction accuracy was 3.04 percent greater than that of previous research. In addition, this technique assures that the dataset is of appropriate quality.The WEKA toolbox was used to perform all of the experimental operations.

Remark: The model's performance was tested using K-fold cross-validation, which decreased the bias associated with the random sampling strategy. They compared their results to several researchers’ trials using the same dataset to show that their model's prediction exactness has improved somewhat. The level of dependability reached 95.42 %. It ensures a shorter processing time and maximal retention of original material without removing excessive amounts of data. To demonstrate the model's great accuracy, it was applied to an actual dataset.

7.”Predictive Analysis of Diabetic Patient Data Using Machine Learning and Hadoop”

We employed machine learning techniques to identify missing values and trends in a Pima Indian diabetic data collection using Hadoop MapReduce-based methodologies. According to specialists, this research will be able to anticipate the forms of diabetes that are prevalent and connected with future hazards. The patient's degree of risk might be taken into consideration while deciding on a treatment plan. Decision tree algorithms may infer missing values and uncover patterns in data sets, as this research demonstrates. Categorization systems based on decision trees may be constructed using

various data sets, as well as a set of rules that explain the separate classes of data. There are inputs and outputs that can be viewed in this approach since it is based on supervised learning. To be able to react to all possible inputs, the algorithm must be able to generate from this training data.
Remark: When plasma levels are high, the patient is diabetic; when plasma levels are low, the patient is not diabetic, as can be shown from the analysis of all the patterns. Both diabetic and non-diabetic people benefit from a medium plasma concentration. For the purpose of predicting diabetes prevalence and risk in the future, pattern matching will be utilised to find patterns in data sets. These patterns will then be used to test data sets.

8."Prediction of Diabetes Diagnosis Using Classification Based Data Mining Techniques"
The use of data mining methods to help individuals forecast diabetes has grown in popularity. Classification methods including Binary Logistic Regression, Multilayer Perceptron, and K-Nearest Neighbor are employed in this study to predict whether or not a person is diabetic, and classification accuracy is compared to categorizing data. The investigation shows that the generation of categories will change depending on the categorization technique. The accuracy of the Binary Logistic Regression is 0.69, the accuracy of the Multilayer Perceptron is 0.71, and the accuracy of the KNN is 0.80, according to the research. For accuracy, KNN is superior than Binary Logistic Regression and Multilayer Perceptrons.
Remark: When it comes to accuracy, this research shows that KNN surpasses both of the other methods. The dataset was obtained through a website, but they may experiment with real-time datasets.

9."Prediction and diagnosis of future diabetes risk - a machine learning approach"
One of the study's main goals was to identify diabetes, which is listed as one of the world's fastest-rising chronic diseases by the World Health Organization. As well as explaining the various techniques for diagnosing diabetes disease, such as gradient boosting and logistic regression, they also explained how Naive Bayes and logistic regression can be used, as well as explaining the accuracy rates of gradient boosting, and how Naive Bayes and logistic regression can be used. BMI and Plasma glucose were shown to be strongly linked in the Pima Indians diabetes dataset throughout the investigation. Consequently, This was done using the Boruta method. Machine learning methods such as Gradient Boosting are excellent for solving regression and classification issues. Prediction accuracy is better with Gradient Boosting than with the other two algorithms tested, according to findings.
Remark: National Council for Science and Technology Communications (NCSTC), the Ministry of Science and Technology (Govt. of India) in New Delhi, financed and accelerated this research effort.The implementation results reveal that gradient boosting has a prediction accuracy of 86%, which is higher than the other two strategies tested.

10."A data-driven approach to predicting diabetes and cardiovascular disease with machine learning"
The study will focus on data-driven techniques that use supervised machine learning algorithms to detect people suffering from cardiovascular disease, prediabetes, and diabetes. Using data from the

National Health and Nutrition Examination Survey (NHANES), they conduct a systematic search of all known factors. Multiple machine learning models were tested using a variety of data sets and time periods. (LR, SVM, RF and GB) (based on laboratory data). The performance of many models was integrated to construct a weighted ensemble model in order to increase detection accuracy. Analysis of a patient's data was done using tree-based models in order to discover the most essential components of the data. Extreme Gradient Boost (XGBoost) has been demonstrated to have an accuracy rate of 86.2 percent (without lab data) and 95.7 percent (with lab data) in its categorization of diabetics (with laboratory data). Patients with pre-diabetes had a higher AU-ROC score of 73.7 percent with the ensemble model, whereas XGBoost had an AU-ROC score of 84.4 percent using laboratory data. In diabetic patients, the five most important markers were: 1) waist circumference, 2) age, 3) self-reported weight, 4) leg length, and 5) sodium intake.

Remark: Patients with diabetes and cardiovascular disease may be able to be automatically identified using survey questionnaire-based machine learning models, according to this research. GB provides a small increase in accuracy, like 3%, with their ensemble models, XGBoost, and the Weighted Ensemble Model. They also indicate crucial predictors, which might be investigated further for their implications for electronic health records.

11.”Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data”

An artificial intelligence-based model for predicting negative outcomes linked to diabetic complications was developed using data from Ontario's single-payer health care system. This data was used to train a Gradient Boosting Decision Tree model using a Python-based XGBoost open source toolbox. The area under the Receiver Operating Model statistic was used to analyse discrimination, and calibration plots were utilised to measure calibration in general and within population subgroups. With the help of 700 features from multiple datasets, our model correctly predicted three-year risk of adverse outcomes due to diabetes complications (high or low blood glucose, tissue infection and retinopathy), as well as strong discrimination (average test AUC = 77.7, range 77.7–77.9) without the use of k-fold cross-validation.

Remark: They use a high-performance model to predict complications and unfavorable outcomes at the community level to demonstrate the potential of machine learning and administrative health data to aid in health planning and healthcare resource allocation for diabetes treatment.

12.”Predictive analysis of diabetic women patients using R”

Diabetes impairs women's immunity, reducing the body's capacity to fight infections. There are several factors that contribute to early mortality, such as heart attacks, obesity and other health issues, kidney and liver failure, high blood pressure, eyesight loss, and Polycystic Ovarian Syndrome (PCOS). Insulin resistance is on the rise, which is contributing to an increase in the prevalence of PCOS in females. As a result, even teens are at a higher risk of developing diabetes. This disease can potentially create complications during pregnancy. In light of these findings, identifying and predicting diabetes in

women is an important aspect of providing better healthcare services. Because of R's prominent role in data analysis and visualisation, this study studies the prevalence of diabetes in women. It does so by analysing several prediction models and establishing their accuracy with statistical consequences. Results were compared using the decision tree, logistic regression model, SVM model and random forest model for categorising results. According to the categorization findings, random forest produces the greatest outcomes. In addition, the random forest's increased classification performance allowed it to overcome the overfitting issue caused by missing values in the datasets.
Remark: Among the four used models, Random Forest gave the best accuracy of 77.06% on the PIMA Indian Diabetes Database using RStudio.

13.”Predictive modelling and analytics for diabetes using a machine learning approach”
A similar data set to the one used in a prior study is used to test numerous data mining models, including SVM, radial basis function kernel support vector machine (RBF) kernel SVM, KNN, ANN, and Multifactor Dimensionality Reduction (MDR). All categorization approaches were tested in the "R" programming studio. The dataset's features are chosen using the Boruta wrapper technique, which enables the impartial selection of essential characteristics. Precision, recall, and accuracy, as well as the F1 score and AUC, are used to evaluate the realism of a model. All of the models tested performed admirably in the real world, according to the findings of the experiments. Because our dataset is an example of an imbalanced class, we believe that using the F1 score will help us better understand the performance of our models. The F1 score strikes the sweet spot when it comes to precision and recall. In addition, the AUCs for the SVM-linear and k-NN models are also 0.90. SVM-linear and k-NN both have great AUCs for diabetes dataset classification.
Remark: A patient's diabetes status may be determined using the linear kernel Support Vector Machine (SVM linear) and the K-Nearest Neighbor (k-NN).Thus, using a small number of parameters, they were able to attain improved accuracy and precision using the Boruta feature selection technique.

14.”A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques”
According on the findings of a literature review, the authors have developed an intelligent framework for diabetes pre-diction. Decision trees (DT) and random forests (RF) are two of the most often used methodologies in the literature for predicting diabetes, and this research uses a framework to develop and assess these models. Machine learning techniques are used to identify training procedures, model evaluation methods, and issues linked to diabetes prediction, as well as the answers they provide. It was developed after a comprehensive review of the literature and a study of their application to diabetes, according to the framework's methodology. Diabetes prediction research and development experts may find this study's findings helpful.
Remark: Evaluation of algorithm performance is accomplished via the use of the Receiver Operating Characteristics (ROC) plot. The use of ROC in healthcare diagnosis and prognosis has been shown to be successful. When the reference point on the ROC chart is in the top left corner, the system or model is regarded excellent. In order to comprehend the exceedingly sensitive and fewer FP reference values,

we need reference points. Using the region below the ROC as a normalisation point is the most straightforward method. An upright test procedure is one that has an AUC larger than 0.5. It depicts the ROC value of LR, with 86 % achieving a high score when compared to others. Because of this, we may conclude that RF is capable of properly diagnosing disease.

15.”Type 2 Diabetes with Artificial Intelligence Machine Learning: Methods and Evaluation”

Predicting type 2 diabetes using machine learning has been proposed in a number of research. As a result, it was impossible to compare the algorithms in the papers since they employed different datasets and assessment measures. Diabetic risk factors are caegorised in this study, and 35 machine learning algorithms (with and without feature selection) are tested to predict diabetes type 2 prevalence using a homogeneous setting. The advantages and disadvantages of classification algorithms were discussed in detail in this article. For the assessment, they employed three real-world diabetes datasets like PIMA Indian, UCI, and MIMIC III and nine feature selection techniques. The preparation of these datasets was specified. For diabetic and non-diabetic people, they determine the accuracy, F-measure, and execution time for model construction and validation of the algorithms under consideration. The article goes into detail about the models' performance analysis. The accuracy and F-measure of classification methods for all datasets with and without feature selection discussed in the article are shown in a bar chart.

Remark: To test the performance of these algorithms, they used three diabetes datasets in an integrated setup to analyse accuracy, F-measure, and number of iterations. A balanced dataset with and without feature selection is best served by the Bagging-LR technique, whereas an unbalanced dataset benefits the most from RF's superior accuracy. Furthermore, they classify type 2 diabetes risk variables in order to examine the most critical factors for diabetes forecasting.

16.”Risk Prediction of Diabetes: Big data mining with a fusion of multifarious physical examination indicators”

The author of this paper developed a computer method for predicting diabetes risk based on a variety of different types of physical examination data. Between 2011 and 2017, researchers in Luzhou, China, collected health data and follow-up information on the city's healthy population and diabetes sufferers. The three kinds of physical examination indicators that were statistically analysed were demographics, vital signs, and laboratory data. With eXtreme Gradient Boosting (XGBoost), a model with an AUC of 0.8768 was developed to tell diabetic patients apart from healthy ones. A diabetes risk assessment based on logistic regression increased the model's usability and versatility in clinical and real-world settings. As a result, it was discovered that patients' illness control was impacted by a variety of crucial follow-up markers. An online diabetes risk assessment system was developed to help with diabetic cascade screening and self-management of a healthy lifestyle. The purpose of this system is to provide recommendations for the management of human health.

Remark: In this work, feature selection approaches such as mutual information (MI), analysis of variance (ANOVA), and Gini impurity(GI) were used to assess the value of the aforementioned attributes .

17.”A Novel Hybrid Approach for Diagnosing Diabetes Mellitus using Farthest First and Support Vector Machine Algorithms”

Farthest First (FF) clustering and Sequential Minimal Optimization (SMO) classifier method are used in this study to propose a hybrid approach for identifying DM. A procedure known as FF clustering is used to split the data into a number of groups. The processing time was considerably shortened when the dataset size was minimized. The clustering output is sent into the SVM classifier. It accurately divides patients into diabetes and non-diabetic groups, i.e., tested positives and negatives.

Remark: The same PIMA Indian dataset was utilised for the experiment, and the findings demonstrated that their suggested integrated strategy attained a classification accuracy of 99.4% for classifying diabetes and non-diabetic patients. It demonstrates that a multimodal model combined with data-mining methods might assist clinicians in making better clinical judgments when identifying diabetes patients.

18.”Type2 diabetes mellitus prediction using data mining algorithms based on the long-noncoding RNAs expression-a comparison of four data mining approaches”

Type 2 diabetes mellitus (T2DM) was diagnosed using four different classification models, including K Nearest Neighbor (KNN), Support Vector Machine (SVM), Logistic Regression (RR), and Artificial Neural Networks (ANN). The models' diagnostic abilities were compared. Six LncRNA variants and demographic data were used to test the algorithms (LINC00523, LINC00995, HCG27 201, TPT1-AS1, LY86-AS1, and DKFZP). Using standard measures, the author aimed to determine the optimum data mining strategy based on the expression of six lncRNAs for predicting T2DM.

Remark: SVM and logistic regression were found to have the highest AUC, among the methods tested. The mean AUC values of KNN and ANN were the highest, with the lowest standard deviations of AUC values. It was found that KNN and SVM had the best sensitivity and specificity values. This study's findings might help us learn more about how lncRNAs can be employed as biomarkers for the early identification and diagnosis of T2DM.

19.”Deep learning based big medical data analytic model for diabetes complication prediction”

Various diabetes strategies have been developed to anticipate diabetic complications. However, the accuracy of categorization and prediction in the present approaches is not very great. As a result, in contrast, this research offers a Deep Learning (DL)-based model to improve diabetes prediction accuracy. The proposed model utilises data collection, pre-training, feature extraction, Deep Belief Network (DBN), validation, and classification methods to assist predict diabetic complications. The show itself is the last piece of the puzzle. DBN's DL-based Big Medical Data Analyt-ics model is compared to conventional methods in terms of precision, accuracy and recall. The data for this model's inputs comes from a variety of diabetes archives. The attributes are then extracted using the RBM pre-training approach. The DBN-based SL categorization includes diabetic retinopathy, diabetic nephropathy, diabetic cardiovascular disease, and amputation.

Remark: As compared to current approaches like SVM and ANN, the proposed DBN-centered prediction model predicts diabetic complications with an accuracy of 81% for training and 81% for testing, respectively, with an 81.20 percent accuracy rate. As a consequence, the data shows that the DBN is more effective. The researchers in this study focus on the detrimental impacts of T2D rather than other diabetes diseases like type 1 and type 2. As a result, this research will be broadened in the future to include new forms of diabetic sickness complications as well as improved classification algorithms for the benefit of society. A cloud-based patient monitoring system will also be developed.

20."Machine learning models for prediction of co-occurrence of diabetes and cardiovascular diseases - a retrospective cohort study"

A total of more than 2,000 individuals participated in the Diabetes Complications Screening Research Initiative (DiScRi), which gathered information on almost 200 different traits. First, they used two machine learning models (logistic regression and Evimp functions) investigate diabetes and cardiovascular disease risk factors using an advanced multivariate adaptive regression spline model. They then used a correlation matrix to eliminate variables that were duplicated.In level 2, they developed their model using a classification and regression approach. Even though the survey focused on diabetes, some papers like this talked about heart diseases, which are a side effect of diabetes.

Remark: The ML model's accuracy, sensitivity, and specificity are all 94.09 percent for detecting the co-occurrence of DM and CVD. If you're participating in a screening process, our ML model can accurately identify the presence of diabetes and cardiovascular disease (CVD). Preventive therapy for diabetes and cardiovascular disease may be more effective if they are diagnosed earlier, which might save money in the long run.

21."A patient network-based machine learning model for disease prediction - The case of type 2 diabetes mellitus"

Researchers in this research used a dataset of de-identified administrative claim data from 1,028 people with type 2 diabetes and 1,028 people without diabetes to create a network of unique patient networks and a machine learning approach for disease prediction. Based on graph theory, the "patient network" may be used to represent the underlying connections between health issues for a group of individuals with the same ailment. The "patient network" attributes (e.g., centrality measure) indicate patients' latent qualities for risk prediction. Analyzing the data is done by eight different types of machine models (LR, KNN, SVM, NB, DT, RF, XGBoost, and ANN).

Remark: The proposed framework using machine learning classifiers worked well, with AUCs ranging from 0.79 to 0.91, according to extensive trials. In comparison to the other models, Random Forest outperformed them allwith the most important elements of the model being eigenvector centrality and network closeness centrality, as well as patient age. Our model's exceptional performance suggests exciting prospective applications in healthcare services. Moreover, we show that the retrieved hidden characteristics are critical in illness risk prediction. The suggested method provides critical insight into chronic illness risk prediction, which may help healthcare professionals and other stakeholders.

V. MEASURING AND EVALUATING METRICS

The usual model evaluation is used in many articles to measure the authors' work toward the proposed model. In general, the confusion matrix is used to analyse the assessment and show the number of properly and erroneously classified classes, with the right classifications marked by Truthful Positive(TP), as well as Truthful Negative (TN). A false positive (FP) arises when a result is anticipated as positive (yes) but is really negative (no), and a false negative (FN) When a bad consequence is expected yet the actual outcome is positive.

The confusion matrix:

		Predicted Class	
		+	-
True Class	+	**TP**	**FN**
	-	**FP**	**TN**

Accuracy is the percentage of correctly categorised items to the total number of correctly categorised items.

$$Accuracy = \frac{TP + TN}{N},$$

$$Where\ N = TP + FP + TN + FN$$

$$Specificity = \frac{TN}{TN + FN}$$

The precision (positive predictive value) with which the data instances are classified is as follows:

$$Precision = \frac{TP}{TP + FP}$$

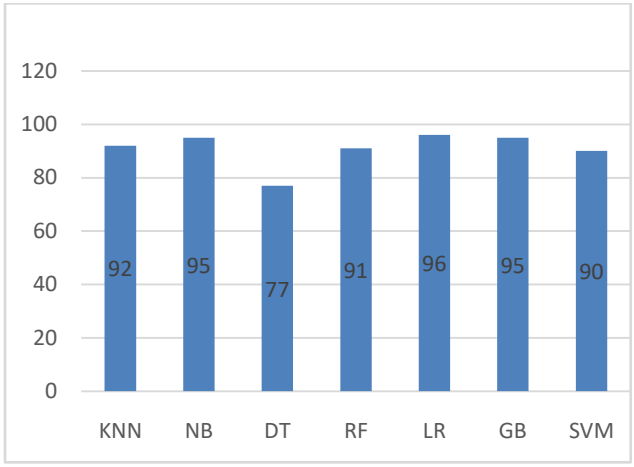
The following is the definition of recall, sometimes referred to as the true positive rate or sensitivity:

$$Recall = \frac{TP}{TP + FN}$$

We require a statistic that considers both accuracy and recall. The F1-score is a measure that takes accuracy and recall into consideration and is defined as follows:

$$F1\ Score(Measure) = \frac{2 * Precision * Recall}{(Precision + Recall)}$$

Conclusions: All of the publications in the selected 21 examined alternative methods and improved accuracy. Some of the most commonly used models are KNN, NB, DT, RF, LR, GB, and SVM. Models such as Adaboost, ANN, J48, and Linear Regression were also employed. Among them, several authors picked the LR models, which have been shown to be the most accurate. Following that, the NB and GB algorithms provided the most accurate results. However, I've compiled a list of the most frequently used models and their greatest accuracy levels in the graph below.



Overall accuracy of all of these algorithms

So, if you're looking for the greatest diagnostics for diabetes, go no further. The best way to acquire diagnosis accuracy is to use the Linear Regression technique. The PIMA Indian Data Set is the most widely utilised dataset. We may use the genuine clinical dataset for improved accuracy.In the future, we may design more new models to illustrate the same.

REFERENCES

[1]. Han and Kamber, "Data Mining:Conceots and Technique", margen Kaufmann Publisher,2006-Elsevier inc.

[2]. Asao K, Sarti C, Forsen T et al., "Long-term mortality in nationwide cohorts of childhood-onset type 1 diabetes in Japan and Finland".
Diabetes Care 26:2037–2042, 2003.

[3].HumarKahramanli and NovruzAllahverdi, "Design of a Hybrid System for the Diabetes and Heart Disease", Expert Systems with Applications: An International Journal, Volume 35 Issue 1-2, July, 2008.

[4].Kaul, K., Tarr, J. M., Ahmad, S. I., Kohner, E. M., &Chibber, R. "Introduction to diabetes mellitus. In Diabetes" (pp. 1-11). Springer New York, 2013.

[5]. American Diabetes Association. "Diagnosis and classification of diabetes mellitus." Diabetes care 31.Supplement 1: S55-S60, 2008.

[6].Namayanja, J., &Janeja, V. P., "An assessment of patient behavior over time periods: A case study of managing type 2 diabetes through blood glucose readings and insulin doses". Journal of Medical Systems, 2012.

[7].Aljumah, A. A., Siddiqui, M. K., &Ahamad, M. G. "Application of classification based data mining technique in diabetes care". Journal of Applied Sciences, 13(3), 416-422, 2013.

[8]. Miroslav Marinov, M.S.et.al., "Data-Mining Technologies for Diabetes: A Systematic Review" Journal of Diabetes Science and Technology, Volume 5, Issue 6, November 2011.

[9]. K.Srinivas, Dr.G.Raghavendra Rao , Dr.A.Govardhan, "Analysis of Coronary Heart Disease and Prediction of Heart Attack in coal mining regions using data mining techniques", The 5th

International Conference on Computer Science & Education Hefei, China., p(1344 - 1349). August 24–27, 2010.

[10]. Jyoti Soni.et.al., " Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.

[11]. Chauraisa V., and Pal, S.,"Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology (IJACSIT), 2, (4), pp 56-66, 2013.

[12]. Srinivas, K., "Analysis of Coronary Heart Disease and Prediction of Heart Attack in coal mining regions using data mining techniques", IEEE Transaction on Computer Science and Education (ICCSE), p(1344 - 1349), 2010.

[13]. "American Diabetes Association. Screening for type 2 diabetes". Diabetes Care, 27:S11– 4, 2004.

[14]. Zimmet P, Shaw J, Alberti KG, "Preventing type 2 diabetes and the metabolic syndrome in the real world: a realistic view". Diabetic Med 2003;20:693–702.

[15]. Ferchak V, Meneghini LF. Obesity, bariatric surgery and type 2 Diabetes: a systematic review. Diabetes Metab Res Rev 2004;20:438 – 45.

[16]. Retnakaran R, Cull CA, Thorne KI, Adler AI, Holman RR, "Risk factors for renal dysfunction in type 2 diabetes", U.K. Prospective Diabetes Study 74. Diabetes 55:1832–1839, 2006.