

Data Gathering and Pattern Recognition for Mellitus Diabetes Classification

Shivani Patel¹, Sanjay Chaudhary², Dr. Prakash Singh Tanwar²

¹Research Scholar, Computer Science and Information Technology, Madhav University, Pindwara (Sirohi)
Rajasthan, India

²Associate Professor, Lovely Professional University, Jalandhar, Punjab, India

ABSTRACT

Data mining techniques are commonly used in medical diagnostics for pattern recognition, processing, and therapy. Diabetes is a metabolic disease in which blood sugar levels remain abnormally high for long periods of time. Diabetes mellitus (DM) is a group of metabolic diseases that puts individuals all over the world under a lot of stress. According to this research, India is home to 19% of the world's population. This overview discusses both type 1 and type 2 diabetes. To compare prior researcher approaches and procedures, a mathematical foundation is employed. The Weka open-source programmer is used to process datasets. We'll speak about getting data from various medical departments in the first half, then data cleaning and finally techniques for reducing noisy data in the second half. To select the optimum attribute, various Algorithms were utilized. Finally, we'll outline future research and compare machine learners for diabetes data categorization.

Keywords — Theoretical, Evaluation, dataset, Pre-Process, and machine learning

I. INTRODUCTION

Computer-assisted care is becoming increasingly significant in healthcare. Few other fields have as many original ideas with such a big societal influence as technology. In medicine, computer-based decision support has a long history of aiding with difficult issues such as illness diagnosis, administrative choices, and the prescription of appropriate medication. Diabetes, an incurable chronic condition, is one of medicine's most beneficial uses. It's a group of metabolic diseases in which a person's blood sugar levels are abnormally high, either because the body doesn't make enough insulin or because the insulin that is produced doesn't reach the cells.

For most businesses, data and information have become key assets. In the process of uncovering

knowledge in medical databases, data mining is a crucial step. Databases are collections of information having a specified structure and function. The programmes that create and manipulate these data are known as database management systems, or DBMS. The total process of extracting knowledge from data is known as knowledge discovery in databases. Data mining is the process of computationally extracting latent knowledge structures expressed in models and patterns from large data sets.

Type 1 diabetes, type 2 diabetes, and type 3 diabetes are the three types of diabetes. Insulin-dependent diabetes, often known as type 1 diabetes, is an autoimmune illness in which the body attacks its own pancreas (typically starting in childhood). The wounded pancreas can no longer make insulin.

Type 2 diabetes, also known as non-insulin-dependent diabetes, is the most prevalent type of diabetes in adults, accounting for almost all occurrences. Type 2 diabetes is a milder form of diabetes that usually develops later in life, although more teens are acquiring type 2 diabetes as a result of the obesity and overweight epidemic in childhood. The pancreas generates some insulin in people with type 2 diabetes. However, either the amount generated is insufficient to meet the body's needs, or the cells in the body are resistant to it. Insulin resistance affects adipose, liver, and muscle cells, causing the pancreas to work too hard to create enough insulin. Obese people are more likely to get diabetes.

Gestational diabetes is diabetes that develops as a result of pregnancy. Between 2% and 10% of all pregnancies are affected by insulin resistance. Because a mother's high blood sugar levels are passed down to her baby via the placenta, they must be kept under control to protect the baby's growth and development.

II. LITERATURE STUDY

Loannis et al. [6] used SVM, LR, and NB to predict different/various medical datasets, including diabetes datasets, utilising 5 fold cross validation. Based on their findings, the researchers assess the accuracy and performance of the algorithms and conclude that SVM delivers the superior accuracy over the other technique.

Francesco et al. [7] invented the multilayer perceptron, random forest, and Decision Tree machine learning algorithms for prediction. According to the researchers, multilayer perceptron (MP) provides excellent accuracy.

Kandhasamy and Balamurali [8] used the Artificial Neural Networks, K-Nearest Neighbour, Navier Bayes, and J48 classification methods. In this study,

the Nave Bayes algorithm provides greater accuracy in the diabetes dataset. In additional datasets on their research, the two algorithms KNN and ANN give great accuracy.

Different data mining strategies were developed by Meng et al. [9] to forecast diabetes illnesses. Three approaches are used in this: ANN, LR, and j48. The results show that the j48 machine learning approach has a higher level of accuracy.

Zhang et al. [10] used Nave Bayes Random Forest and Adaboost to train machine learning approaches to predict diabetes illnesses. Random forest outperformed other models, according to the researchers.

J48, multilayer perceptron, and Random Forest are used by Francesco et al. [11]. In compared to the others, j48 provided superior accuracy (77.5%).

On many datasets, including diabetes, Rani and Jyothi [12] employed Nave Bayes and ANN. With an accuracy rating of 77.01 percent, Nave Bayes exhibited great accuracy. Saravananathana and Velmurugan [13] employed CART, SVM, and K-NN among other machine learning techniques. According to their research, CART has a 62.28 percent accuracy, SVM has a 65.04 percent accuracy, and K-NN has a 53.39 percent accuracy.

To assess whether or not a person is diabetic, Yasodha and Kannan [14] applied machine learning classifiers to a range of datasets. The data was categorised using WEKA, and the data was analysed using a 5-fold cross validation approach, which works well on small datasets, and the findings were compared in this study. naive Bayes, J48, REP Tree, and Random Tree are among the algorithms employed. With a 60.2 percent accuracy rate, J48 was judged to be the most efficient.

By analysing and analysing data patterns via classification analysis using Decision Tree and Nave Bayes algorithms, Sumbaly et al. [15] created diabetes detection systems. In comparison to Naive Bayes, the

experimental findings demonstrate that the j48 algorithm has a 74.8 percent accuracy rate.

Gupta et al. [16] attempted to evaluate and analyse the outcomes of multiple classification techniques in WEKA, as well as identify and quantify the accuracy, sensitivity, and specificity percentages of various classification methods. The results suggest that Random Forest has the highest accuracy of 81.3 percent, 59.7% sensitivity, and 81.4 percent specificity.

Chikh et al. [17] used an enlarged Artificial Immune Recognition System 2 (AIRS2) termed modified Artificial Immune Recognition System 2 to increase the identification accuracy of diabetes diseases (MAIRS2). The K-nearest neighbour approach is replaced with fuzzy K-nearest neighbours to increase the diagnosis accuracy of diabetes diseases. The authors were able to strike a good balance between data reduction and classification accuracy. The suggested system (MAIRS2) fared better than the AIRS2 benchmark. The authors were able to get the highest classification accuracy of 80.10 percent by utilising MAIRS2.

Sharmila and Manickam [18] set out to analyse data in order to predict diabetes from patient medical records. This work uses the R tool to analyse diabetes from large medical data using the decision trees technique. The decision tree algorithm has a 79.33 percent accuracy.

In order to anticipate the most frequent kinds of diabetes, as well as related complications and therapies, Lavanya et al. [19] concentrated on constructing a prediction model by studying the strategy in a Hadoop/Map Reduce context. The proposed predictive analysis system design includes data collection, warehousing, predictive analysis, and the processing of analysed results.

Tiwari and Diwan [20] have presented a method for identifying cancer sickness patterns that is both automated and hidden. In the offered system, data

mining techniques such as association rules and clustering were used. A major job in this work is attribute-based clustering for feature selection. This method was used to vertically split the data collection. The data is divided into two clusters, one including all significant properties and the other containing all irrelevant ones.

Khaleel et al. [21] focused on data mining approaches that are critical for medical data mining, namely for detecting common illnesses including heart disease, lung cancer, and breast cancer. Some of the data mining algorithms utilised on medical data include apriori and FPGrowth, as well as unsupervised neural networks, linear programming, and association rule mining. The association rule mining approach locates elements in a collection that appear frequently. Medical mining gives the data needed to make educated diagnosis and decisions.

Chaurasia and Pal [22] forecasted diabetic illnesses using several machine learning techniques. WEKA, an information retrieval programme with a collection of machine learning techniques, is utilised. This inquiry employs Naive Bayes, j48, and bagging. j48 has an accuracy of 84.35 percent. Bagging achieves an accuracy of 85.03 percent. On this dataset, bagging provides a superior classification factor.

Parthiban and Srivatsa [23] published a study on detecting coronary artery disease in diabetic individuals. Machine learning techniques were used to do this. WEKA employs the Naive Bayes and SVM algorithms. Using SVM, the greatest accuracy of 94.60 is attained.

Sumbaly et al. [24] used two approaches to predict diabetic complications: decision trees and Naive Bayes. j48 performed Cross Validation and Percentage Split (PS) independently. Using PS, Naive Bayes provides 79.5652 percent accuracy. Using rate split test, algorithms provide the highest level of accuracy.

Alic et al. [25] compared artificial neural networks (ANN) and Bayesian networks, the two most widely

used sickness prediction methods. Because of the independent association between observed nodes, Artificial Neural Networks achieve a higher accuracy of 89.78 percent compared to Bayesian Networks' 80.43 percent.

To predict liver illness, Murthy et al. [26] employed support vector machine (SVM) and Naive Bayesian Classification techniques. This data collection contains 560 occurrences as well as 10 attributes. The comparison is based on precision and execution time. Bayes Classifier yields 61.28 percent accuracy in 1670 milliseconds. In 3210 milliseconds, SVM achieves 79.66 percent accuracy. SVM has the greatest accuracy when compared to the Naive Bayesian for predicting liver illness. The Naive Bayesian algorithm takes less time to run than the SVM method.

Baby and Vital [27] built a prototypical that can envisage the sympathetic of renal disease based on a data collection of people with kidney illness. The model examined the outcomes of numerous classification techniques, including random forests, Decision Trees, j48, and K-means algorithms, and found that RF outperformed the others.

Using machine learning approaches, Razia and Rao [28] created a framework model to identify thyroid illness. Unsupervised and supervised learning are employed to identify thyroid disease, and the framework model beats the decision tree model.

Alehegn et al. [29] employed SVM, Naive Net, Decision Stump, and the Proposed Ensemble technique to forecast diabetic diseases. The suggested ensemble technique has a 90.36 percent accuracy. In comparison to others.

The Adaboost M1 method was combined with a random committee by Kang et al. [30]. The accuracy of the prediction is 81.0 percent.

III. Proposed Methodology

The process of collecting knowledge from data using computer-based technologies is known as data mining. Data collection and preparation are the first steps in the strategy used to complete the research assignment. The Pima Indians Diabetes Database, which was gathered from several archives, served as the training dataset for data mining. The data preparation process covered all operations needed to create the final dataset from the raw data. Different modelling methodologies are chosen and used, with their parameters adjusted to provide the best results. To tackle the same data mining problem, classification algorithm techniques are employed.

Dataset

The data was gathered from Chinese hospital physical examination records. This data is split into two categories: healthy persons and diabetics. Data from two healthy people's physical examinations are available. As the training set, we used 164431 instances of data from healthy people's physical examinations. In the alternative data set, 13700 samples were picked at random as an independent test set. The 14 clinical assessment indexes include age, pulse rate, breath, left systolic pressure, right systolic pressure, left diastolic pressure, right diastolic pressure, height, weight, physique index, fasting glucose, waistline, low density lipoprotein, and high-density lipoprotein.

Another piece of data is on diabetic Pima Indians. All of the patients, in particular, are Pima Indian women over the age of 21. The dataset includes information on the time of pregnancy, plasma glucose concentration after a 2-hour oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, 2-hour serum insulin, BMI, diabetes pedigree function, and age. After missing data is eliminated, this dataset's 786 diabetes records are reduced to 92.

Pre-Process

- **Cleaning the data:** Many portions of the data may be meaningless or missing. In order to control this component, data purification is performed. It requires coping with data that is missing, noisy, and so on.
- **Data Reduction:** Data mining is a technique for dealing with massive volumes of information. Analyzing big volumes of data gets increasingly complex. To eliminate this, we use a data reduction technique. Its objective is to increase storage efficiency while minimising data storage and analysis costs. The four phases in the data cube aggregation method are data cube aggregation, attribute subset selection, numerosity reduction, and dimensionality reduction.
- **Data Transformation:** In this stage, the data is transformed into a format that can be used in the mining process. This can be done in a variety of ways: 1. Consolidation 2. Attribute Selection 4. Creation of a Concept Hierarchy 3. Disambiguation.

Recursive Feature Elimination

Recursive Feature Elimination, or RFE for short, is a popular feature selection algorithm. RFE is popular because it is easy to configure and use and because it is effective at selecting those features (columns) in a training dataset that are more or most relevant in predicting the target variable. There are two important configuration options when using RFE: the choice in the number of features to select and the choice of the algorithm used to help choose features. Both of these hyperparameters can be explored, although the performance of the method is not strongly dependent on these hyperparameters being configured well.

Machine Learning

- **Support Vector Machine:** The SVM classifier's goal is to locate the optimal hyperplane from which to divide the classification model. A extracted features and a class label are used to describe each trained model. The hyper-plane is trained and equipped to distinguish the majority of samples out of the same class from all the other samples. A binary classifier is what an SVM is. The logical concept(s) utilised for picture annotation are the classifier's output. The goal is to create a separating hyperplane that splits the collection of instances thus all locations of the same description are all on the same side of it.
- **Naive Bayes:** The Nave Bayesian method is among the most efficient and productive inductive learning approaches for machine learning techniques. The Bayes principle of likelihood function is involved. It presume that a class's properties are autonomous of one another. In other terms, it believes that the influence of a specific class's feature values is unaffected by the values of those other features. This is one of the most well-known classification methods. This is most commonly utilised in the estimate of probability that correspond to a certain group.
- **Decision tree:** The decision tree, which starts with a single node representing the training samples, employs the tree structure. If all of the samples belong to the same class, the node is transformed into a leaf, and the class is used to identify it. Otherwise, the algorithm selects the discriminating attribute as the current node in the decision tree. The training samples are divided into several subsets, each of which generates a branch based on the value of the current decision node attribute, and there are several values that generate numerous branches. The preceding phases are repeated for each subset or branch gained in the previous step, recursively

building a decision tree on each of the partitioned samples.

K-nearest neighbour (k-NN): It's a type of machine learning technique that uses the input's nearest neighbors to categories it. It's an algorithm that keeps track of all instances and categorizes new ones using a similarity metric. It's also known as case-based reasoning. ii) the number of nearest neighbors iv) Slow learning KNN calculations have been used in a variety of applications, including statistical estimation and pattern recognition. KNN is a non-parametric order method that may be classified into two types: structure-based NN techniques and structureless NN approaches. All of the data in structure-free NN techniques is categorized into training and test data. The distance between the training and sample locations is calculated, and the nearest neighbor has the least distance. N uses information structures such the orthogonal structure tree (OST), ball tree, k-d tree, closest future line, axis tree, and focus line. When all characteristics are continuous, the closest neighbor arrangement is utilized. In two phases, a simple computation of the K nearest neighbor is performed: 1. Locate the training K cases that are the most similar to the unknown instance. 2. For these K examples, choose the most common categorization.

- **Random Forest:** A random tree number of co classifier contains a number of branches, each of which has been randomly grown. Iterations of the subsequent probability across picture classes are used to identify the nodes from each node. Each grid cell has a test that divides the region of data to be categorized in the most efficient way. Through transmitting an information down every node and summing the limb patterns attained, a picture is categorized. Variation may be introduced throughout learning at 2 places: when subsampling the training data so that each tree is created using a different subset, and when picking the node tests.

IV. Results and Evaluation

This section will go through the many types of machine learners, as well as their benefits and drawbacks. Finally, a comparison classifier evaluation will be performed, based on training speed, accuracy, problem type, prediction speed, and performance.

TABLE I. COMPARATIVE ANALYSIS

Classifier	Desired Results	Prediction		Accuracy
		Diabetic	Non-diabetic	
SVM	Diabetic	102	87	77.38318
	Non-diabetic	34	312	
KNN	Diabetic	87	102	73.83178
	Non-diabetic	38	308	
Naïve Bayes	Diabetic	114	75	72.33645
	Non-diabetic	73	273	
RF	Diabetic	122	67	82.99065
	Non-diabetic	24	322	
DT	Diabetic	137	52	88.78505
	Non-diabetic	8	338	

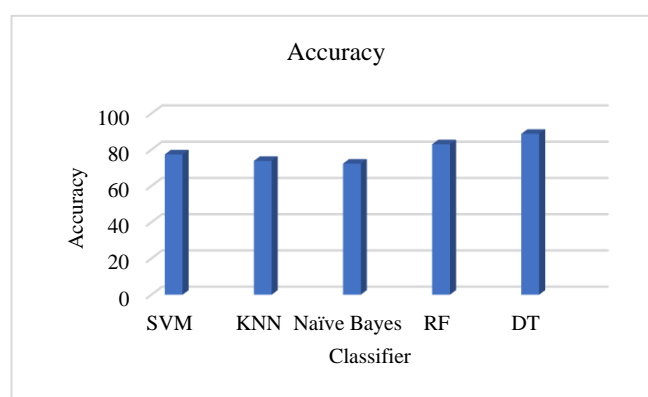


Fig. 1. Accuracy Graph

V. CONCLUSION AND FUTURE IDEA

Pima Indians diabetes was used in the comparative study with different ML. The datasets were then subjected to the multiple imputation and listwise methods. After that, the performance of various system algorithms was compared to the performance of the system described in the associated study. We discovered that the performance of is influenced by feature selection and machine learning. Comparative analysis and graph show that decision tree give 88.78% performance among all. So, in future new version of diction tree classifier that is an ensemble-based classifier, such as the extra tree classifier, is utilized to optimize the performance parameters.

VI. REFERENCES

- [1]. Abhilasha Dutta, Abhishek Kumar, Shukla S N2 , Amol Daniel, . (October 02, 2016). Impaired fasting glucose and impaired glucose tolerance in rural central Indian: A study of Prediabetes in the first-degree relatives of patients with type 2 diabetes in rural region of Malwa in Madhya Pradesh. Madhya Pradesh.
- [2]. Charu .V. Vermal, Dr. S. M. Ghosh2 (2017 IJESC). Review of Cardiovascular Disease in Diabetic Patients using Data Mining Techniques (Vol. 7).
- [3]. Devi, Dr M. Renuka. (2016). International Journal of Applied Engineering Research (Vol. 11). (J. M. Shyla, Ed.) Dean of Computer Science.
- [4]. Srideivanai Nagarajan, R. M. Chandrasekaran. (April 2015). Indian Journal of Science and Technology (Vol. 8). Design and Implementation of Expert Clinical System for Diagnosing Diabetes Using Data Mining Techniques.
- [5]. V. Umatejaswi, P. Suresh Kumar. (June 2017). Diagnosing Diabetes using Data Mining Techniques (Vol. 7). International Journal of Scientific and Research Publications.
- [6]. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2010). Machine learning and data mining methods in diabetes research. Computational and structural biotechnology journal.
- [7]. Mercaldo, F., Nardone, V., & Santone, A. (2010). Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques. *Procedia Computer Science*, 112(C), 2519-2528.
- [8]. J. Pradeep Kandhasamy, S. Balamurali (2011). Performance analysis of classifier models to predict diabetes mellitus. *Procedia Computer Science*, 47, 45-51.
- [9]. Xue-Hui Meng, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang, Qing Liu (2011). Comparison of three data mining models for predicting diabetes or pre diabetes by risk factors. *The Kaohsiung journal of medical sciences*, 29(2), 93-99.
- [10]. Xu, W., Zhang, J., Zhang, Q., & Wei, X. (2011, February). Risk prediction of type II diabetes based on random forest model. In *Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2011 Third International Conference on* (pp. 382-386). IEEE.
- [11]. Mercaldo, F., Nardone, V., & Santone, A. (2012). Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques. *Procedia Computer Science*, 112(C), 2519-2528. 40
- [12]. Rani, A. S., & Jyothi, S. (2012, March). Performance analysis of classification algorithms under different datasets. In *Computing for Sustainable Global Development (INDIACom), 2012 3rd International Conference on* (pp. 1584-1589). IEEE.
- [13]. Saravananathan, K., & Velmurugan, T. (2013). Analyzing Diabetic Data using Classification Algorithms in Data Mining. *Indian Journal of Science and Technology*, 9(43).
- [14]. P. Yasodha and M. Kannan, "Analysis of a Population of Diabetic Patients Databases in

- WekaTool", International Journal of Scientific & Engineering Research, vol. 2, no. 5, 2013.
- [15].A. Iyer, J. S and R. Sumbaly, "Diagnosis of Diabetes Using Classification Mining Techniques", IJDKP, vol. 5, no. 1, pp. 01-14, 2014.
- [16].N. NiyatiGupta,A.Rawal, and V.Narasimhan,"Accuracy, Sensitivity and Specificity Measurement of Various Classification Techniques on Healthcare Data", IOSR Journal of Computer Engineering, vol. 11, no. 5, pp. 70-73, 2014.
- [17].M. Chikh,M. Saidi, and N. Settouti, "Diagnosis of diabetes diseases using an Artificial Immune Recognition System2 (AIRS2) with fuzzy K-nearest neighbor," Journal of medical systems,vol.36, no.5, pp. 2721-2729, 2015.
- [18].K. Sharmila and S. Manickam, "Efficient Prediction and Classification of Diabetic Patients from bigdata using R,"International Journal of Advanced Engineering Research and Science, vol. 2, Sep 2015.
- [19].GyorgyJ.Simon, Pedro J.Caraballo,Terry M. Therneau,Steven S.Cha, M. Regina Castro and Peter W.Li "Extending Association Rule Summarization Techniques to Assess Risk Of Diabetes Mellitus," IEEE Transactions on Knowledge and Data Engineering ,vol 27,No.1,January 2015.
- [20].Vikas Tiwari, T. (2016). Design and implementation of an efficient relative model in cancer disease recognition". IJARCSSE.
- [21].Khaleel, M. A. (2016). A Survey of Data Mining Techniques on Medical Data for finding frequent diseases. IJARCSSE.
- [22].Chaurasia.V, P. (2016). Data Mining Approach to Detect Heart Disease. IJACSIT, 56-66.
- [23].Parthiban.G, S. (2016). Applying Machine Learning Methods in Diagnosing Heart Disease for Diabetic Patients. IJAIS, 25-30. 41
- [24].Iyer.A, S. (2017). Diagnosis of Diabetes Using Classification Mining Techniques. IJDKP, 114.
- [25].Almir Badnjević, Lejla Gurbeta, Berina Alić (2017). Machine Learning Techniques for Classification of Diabetes and Cardiovascular Diseases. Mediterranean Conference on Embedded Computing.
- [26].Baby, P. (2017). Statistical Analysis and Predicting Kidney Diseases using Machine Learning Algorithms. IJERT.
- [27].Shaik Razia and M.R.NarasingaRao "A Neuro computing frame work for thyroid disease diagnosis using machine learning techniques", Vol.95. No.9. Pages 1996-2005).
- [28].Alehegn, Minyechil, and Rahul Joshi& Dr Preeti Mulay. "Analysis and Prediction of Diabetes Mellitus using Machine Learning Algorithm." International Journal of Pure and Applied Mathematics, No. 9 (2018).
- [29].Ali, Rahman, et al. "Prediction of diabetes mellitus based on boosting ensemble modeling." International conference on ubiquitous computing and ambient intelligence. Springer, Cham, 2014.
- [30].Dr.B.L.Shivakumar(2016). Diagnosis of diabetes by Applying Data Mining Classification Techniques Comparison of Three Data Mining Algorithms (Vol. 7). (Riyad Alshammari, Tahani Daghistani, Ed.) International conference on Intelligent Computing Application,IJACSA, www.ijacsa.thesai.org.
- [31].Elsevier T. Santhanam a, M. P. (2015). Application of K-Means and Genetic Algorithm for Dimension Reduction by Integrating SVM for Diabetes Diagnosis (Vol. 47). Procedia computer Science, www.sciencedirect.com.
- [32].J.Pradeep Kandhasamy, S. B. (2015). performance Analysis of Classifier Models to Predict Diabetes Mellitus (Vol. 47). Procedia Computer Science.