An Efficient Method for Predicting Diabetic Disease using Machine Learning Algorithms

V. SHANMUGASUNDARAM
Research scholar
Department of Computer Application
Bharathiar University
Coimbatore-46
mvshanmugambu@gmail.com

Dr. M.PUNITHAVALLI
Professor
Department of Computer Application
Bharathiar University
Coimbatore-46
punithavalli@buc.edu.in

***Abstract*—Diabetes is one of the chronic and ongoing diseases in the world. - Diabetes is a chronic disease caused by high levels of glucose in the human body. Diabetes can cause major health issues if it is not treated. The PIDD (Pima Indians Diabetes Database) data set were used in this work. In this paper discuss about Hybrid K-Nearest Neighbor (H-KNN) algorithm. In the H-KNN algorithm combined with KNN Classifier and Voting Classifier. to predict diabetes from the dataset. The main objective of this paper is to improve the decision-making accuracy by detecting missing values and of the medical data for the accurate diagnosis of diabetes. For this purpose, it uses for handling the missing data values and outliers in the PIDD. It has been found that the proposed algorithm achieves 89% of accuracy, 82% of precision and 77% of recall**

Keywords—*Diabetic, Non-Diabetes, PIDD, H-KNN, Voting Classifier*

I. INTRODUCTION

During the last twenty years, the prevalence of diabetes has increased dramatically in many parts of the world and the disease is now a worldwide public health problem. The total number of people with diabetics is projected to rise from 171 million in 2000 to 366 million in 2030 [1]. The mortality rate is higher among people with diabetes than among the rest of the population with an excess mortality of 65 % in 2007. Excess mortality is mainly due to diabetes-related diseases developed because of poorly controlled diabetes [2]. The socio-economic impact of diabetes and diabetes care in Denmark is estimated about 22 billion DKK per year in 2008 including the health sector's annual cost for treating diabetes and diabetes-related diseases. Moreover, the social costs of production due to diabetes represent another 9.5 billion DKK yearly [3]. The term diabetes mellitus refers to a collection of metabolic diseases mostly carried on by faulty insulin secretion. Hyperglycemia (high blood sugar) and poor protein, lipid, and carbohydrate metabolism are both caused by insulin insufficiency [4]. Type 1 diabetes results from autoimmune mediated destruction of the beta cells of the pancreas. Insulin is vital for individuals with type 1 diabetes to avoid ketoacidosis, coma and death [5]. Type 2 diabetes is the most common type which often results due to excess body weight and physical inactivity in genetically predisposed individuals. Over time, diabetes can increase the risk of health-relatedproblems including blindness, kidney damage, nerve damage, amputation of lower limbs and cardio vascular disease [6] The increasing prevalence of the disease hereby also increases the social costs combined with the fact thatthe disease is associated with increased morbidity and themortality emphasizes the importance of effective diabetescare [7].

II. OVERVIEW OF ALZHEIMER'S DISEASE

A significant amount of research has been conducted in recent years to predict diabetes using machine learning techniques. Maniruzzaman [8] has suggested various cross-validation strategies and dimensionality reduction by choosing the right features to predict diabetes.

For forecasting disorders, ML algorithms are widely used in the medical industry. To get the most accurate and reliable findings, numerous studies have employed ML approaches to predict diabetes [9].

To identify diabetes, Deepti and Dilip [10] employed Decision Tree, SVM, and Naive Bayes classifiers. Finding the classifier with the maximum accuracy was the goal. For this investigation, the PIDD dataset was utilized. Cross-validation on 10 folds is used to partition the dataset. Accuracy, precision, recall, and the F-measure were used as metrics to assess performance. The Naive Bayes model had the highest level of accuracy (76.30%).

Three algorithms for supervised learning were used by

Mamuda and Sathasivam [11] The PIDD dataset was used in this study to assess performance. Data was divided into training and testing for the validation research using the 10-fold cross validation. According to the authors, Levenberg Marquardt (LM), whose Mean Squared Error (MSE) of 0.00025091, had the best performance on the validation set.

Amina Azar et al. [12] Diabetes affected among young peoples and ancient peoples. These are increased day by day and it does not curable. Data mining is used to early-stage prediction. This paper in main aim is gives the differentiation and suggest best algorithm. The PID datasets are used. The Decision tree, Naive Bayes and K-Nearest neighbour algorithms are compared and used for predict the diabetes diagnosis at early stage with highest accuracy and efficiency. The result is the decision tree is the best prediction algorithm. It gives the accuracy level is 75.65%.

VeenaVijayan.V [28] It is obvious that selecting proper algorithms for categorization improves the system's accuracy and effectiveness. The major goal of this work is to evaluate the advantages of various pre-processing methods for decision support systems based on Support Vector Machine (SVM), Naive Bayes classifier, and Decision Tree for predicting diabetes. Principal Component Analysis and Discretization are the pre-processing techniques that this study is concentrating on. The variation in accuracy assessed both with and without pre-processing methods. This study makes use of the Weka tool. The University of California, Irvine (UCI) machine learning repository served as the source of the dataset.

Rashid et al. [45] created diabetes mellitus support systems that operate automatically using classification techniques, encapsulating the aforementioned problems and also reflecting the skills of medical professionals who believe that there is a significant correlation between the adverse effects of some chronic illnesses and the carbohydrate rate. This study's implications might go beyond simply classifying diabetes mellitus patients. Thus, the following are the primary duties: It utilizes a few unrestricted variables.

III. Material and Methods

Diabetes is a most common disease faced by huge population globally. It is a chronic disease caused by the glucose presence in an individual blood stream. Insulin hormone helps glucose in the blood stream to move freely around the body for delivering energy to cells.

A. Dataset Description:

In order to compare and validate the findings, the system is tested on the most commonly used PIDD (Pima Indians Diabetes Database) which belongs to the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), United States [12]. It is part of the UCI machine learning dataset available to researchers. PIDD samples are taken from Pima Indian Heritage which contains the information of female patients having gestational diabetes of age category higher than 21 years. The data samples were gathered during the first trimester of pregnancy.

This dataset contains 768 instances and 8 attributes. The input attributes are age, glucose concentration in blood 2 hours after having breakfast (Glucose), serum insulin in blood 2 hours after having breakfast (Insulin), Body Mass Index (BMI), Number of Pregnancies (NP), Triceps Skin Fold Thickness (TSFT), Diabetes Pedigree Function (DPF) and diastolic Blood Pressure (BP).

The class label is associated with every sample for indicting whether the individual is affected with diabetes or not. The output i.e., class label of the system is either 0 or 1. 0 is interpreted as "no diabetes mellitus" and 1 is interpreted as "diabetes mellitus". There are 8 attributes that can be described as follows:

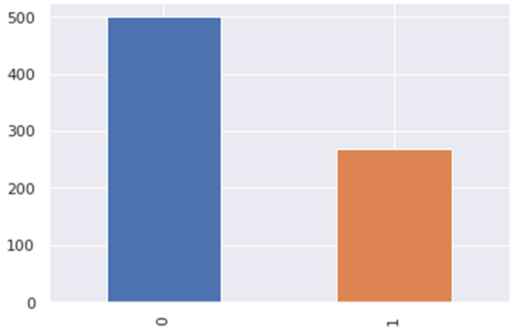
1. Number of times pregnant
2. Plasma glucose concentration 2 hours in an oral glucose tolerance test (mg/dL)
3. Diastolic blood pressure (mm Hg)
4. Triceps skin fold thickness (mm)
5. 2-Hour serum insulin (mu U/ml)
6. Body mass index (weight in kg/ (height in m)2)
7. Diabetes pedigree function
8. Age (years)

Among 768 instances gathered, 268 instances had been diagnosed with diabetes and 500 instances without diabetes.

Fig. 1. Outcome for Diabetes and Non-Diabete

The above graph shows that the data is biased towards datapoints having outcome value as 0 where it means that diabetes was not present actually. The number of non-diabetics is almost twice the number of diabetic patients

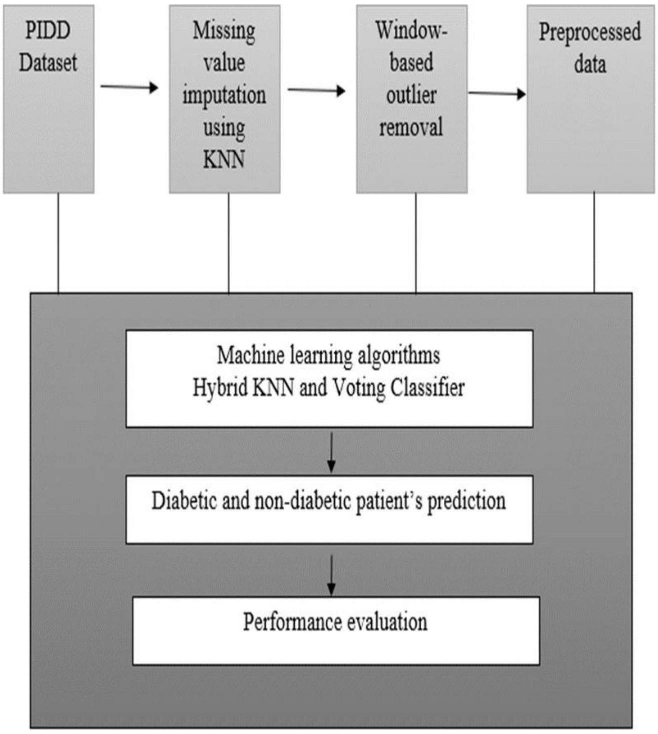
B. Hybrid KNN Algorithm



In this proposed algorithm KNN and Voting Classifier are used. KNN algorithm can be used to approximate missing values in the dataset by using statistical tools like mean and mode tofind the most suitable values. Outlier handling is achieved through KNN distance weight window.

Francis et al[11], examines the existing scientific applicability of the original cholinergic hypothesis of Alzheimer’s disease by describing the biochemical andGrid-search is used to find the optimal hyperparameters of a modelwhich results in the most 'accurate' predictions. With GridSearch we search the best "neighbors" to optimize accuracy of voting classifier. The robust and accurate prediction of diabetes is highly challenging due to the limited number of labeled data and also the presence of outliers (or missing values) in the diabetes datasets. The main objective of this paper is to improve the decision-making accuracy by detecting missing values and of the medical data for the accurate diagnosis of diabetesFor this purpose, it uses for handling the missing data values and outliers in the PIDD. It has been found that the proposed algorithm achieved well in several performance measure

Fig.2. Overall Methodology


PIDD Dataset
Missing value imputation using KNN
Window-based outlier removal
Preprocessed data
Machine learning algorithms
Hybrid KNN and Voting Classifier
Diabetic and non-diabetic patient's prediction
Performance evaluation

PIDD dataset were used in this research. In PIDD dataset class labels are used as diabetics and non-diabetics denoted by 0 and 1. Using PIDD, feature scale is not standardized; therefore, number of classes in label feature set is imbalanced. Pearson's correlation coefficient statistical method used for feature selection.

The pre- processes include missing value imputation and outlier removal from the dataset. It needs to be trained by using a training dataset as the primary phase. The training dataset should not have missing values. The data with missing values is split to the other dataset and leave the data instances with complete values as the training dataset. Data that falls beyond the accepted range is called outlier and window-based outlier identification method is applied.

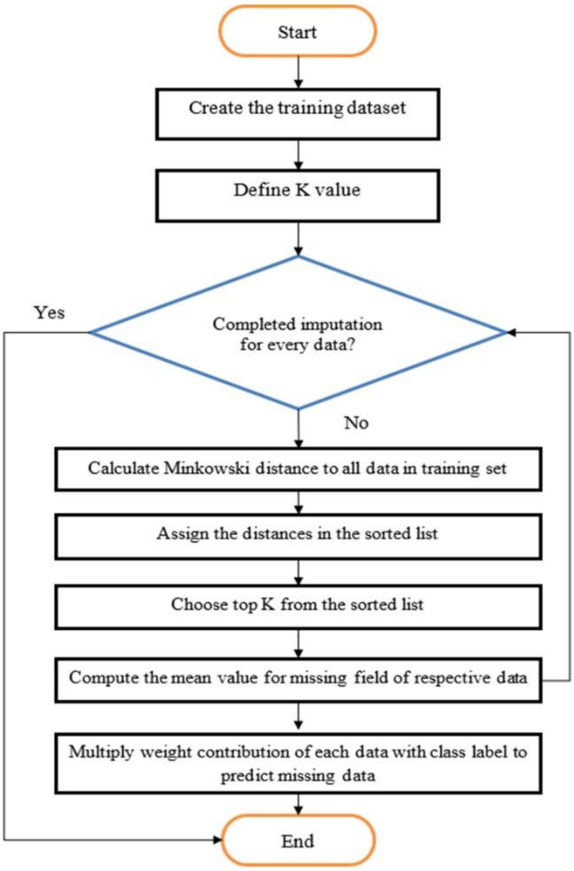
Once these preprocessing is completed, the obtained new balanced dataset is given to the machine learning classification algorithms such as Hybrid KNN and KNN for predicting the diabetic and non-diabetic patients. Figure 2 represents overall methodology flow.

c. *Missing Value Imputation And Outlier Detection*

The computation time due to distance computation is solved by using a Minkowski method which finds the distance to select the neighbors to consider for the algorithm. The generalized Minkowski distance measure is as:

d(a,b)= 〖(|u_a1-u_b1 |^q+|u_a2-u_b2 |^q+⋯+|u_ap-u_bp |^q)〗^(1/q).. (1)

Here, q is known as Minkowski coefficient. The Minkowski distance reduces to the Manhattan distance if q=1 and as the Euclidean distance if q=2. The challenge here is managing partially unknown and known information to predict a class label to which an instance belongs to. The proposed KNN-based missing value imputation is shown in Figure 3.


Start
Create the training dataset
Define K value
Completed imputation for every data?
Yes
No
Calculate Minkowski distance to all data in training set
Assign the distances in the sorted list
Choose top K from the sorted list
Compute the mean value for missing field of respective data
Multiply weight contribution of each data with class label to predict missing data
End

Outliers detection applied when data is grouped based on time. The window size plays a vital role in determining right value to fill in the missing field. Initially, first point di time series data in the KNN window is used to find the relationship between di and its nearest neighbor. When window size is high, then the number of comparisons between instances in the window and its neighbor will go high. It is recommended to set window size to be high when there is a huge number of missing values. But it incurs huge computation complexity during the training stage. It is detailed in the algorithm given below.

C. *Pearson's Correlation Coefficient*

Pearson's Correlation Coefficient helps you find out the relationship between two quantities. It gives you the measure of the strength of association between two variables. The value of Pearson's Correlation Coefficient can be between -1 to +1. 1 means that they are highly correlated and 0 means no correlation.

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

r = correlation coefficient
x_{i} = values of the x-variable in a sample
\bar{x}=mean of the values of the x-variable
y_{i} = values of the y-variable in a sample
\bar{y}=mean of the values of the y-variable

A heat map is a two-dimensional representation of information with the help of colors. Heat maps can help the user visualize simple or complex information.

IV. RESULT

To help us comparing the results fan algorithm against another one some metrics were calculated. I-KNN achieves good performance compared to KNN algorithm. These metrics are accuracy, precision, recall and f-measure. Classified examples are categorized as True Positive (TP), True Negative(TN), False Positive(FP) or False Negative(FN) depending on the classification label and the true label.

Accuracy

Accuracy is the proportion of true positive and true negative among the total number of cases examined. It is computed as follows:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision

It is the measure to find the ability of a classification model to identify only the relevant instances in the dataset. It is defined as the number of true positives divided by the number of true positives plus the number of false positives.

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

Recall

Recall can measure the model's ability to find all the data instances of interest in a dataset. The precise definition of recall is the number of true positives divided by the number of true positives plus the number of false negatives.

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

Algorithms	Diabetes	Accuracy	Precision	Recall
Hybrid KNN	Non-Diabetes	90	84	78
	Diabetes	88	80	75
	Average	**89**	**82**	**77**
KNN	Non-Diabetes	80	75	70
	Diabetes	78	70	68
	Average	79	73	69

Table1.Performance matrices

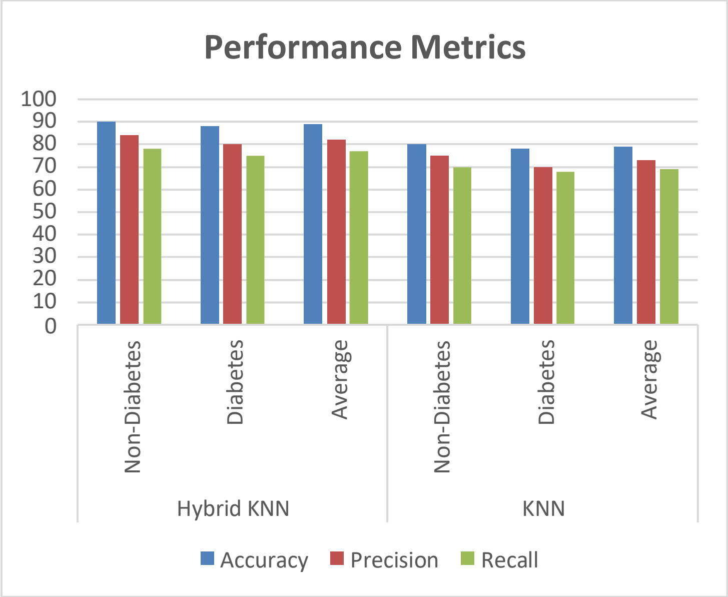


Figure3. Performance matrices forH-KNN Algorithm

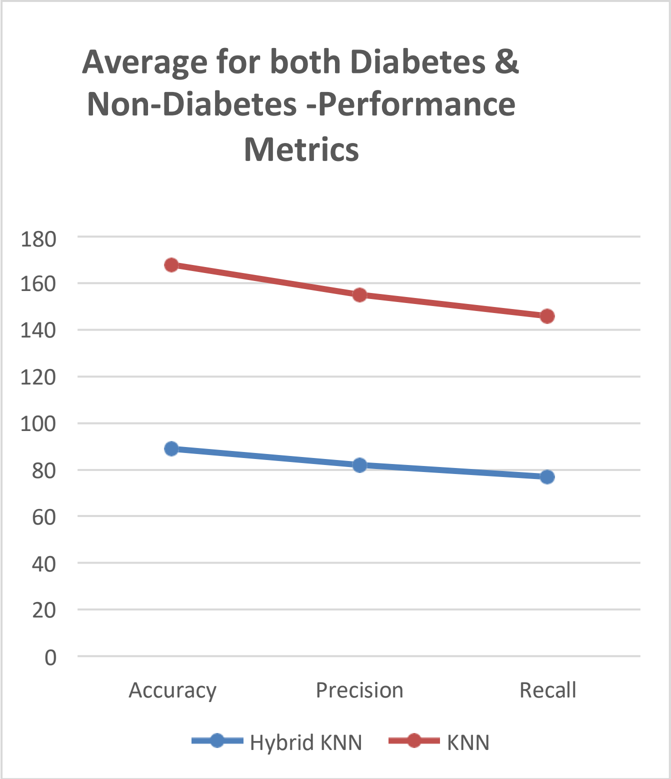


Figure 4. Statistical Analysis of Performance Metrics

From the table 1, Figure 3& 4 represent the overall performance metrics result such as accuracy, precision and recall for both diabetes and non-diabetes.

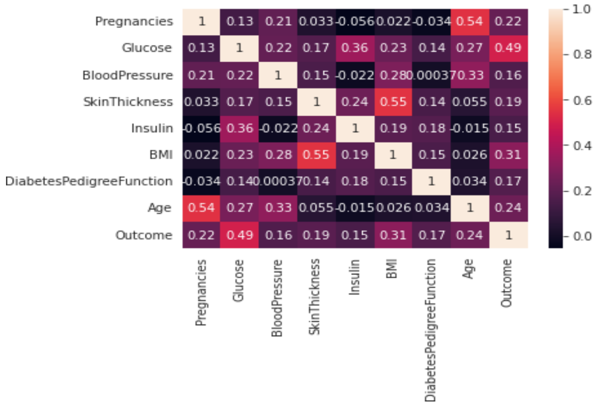


Figure 5. Heatmap for feature Selection

V. Conclusion

The early identification of diabetes is one of the major real-world medical issues. In this work, deliberate efforts are undertaken to create a system that predicts diseases like diabetes. H-KNN examined and assessed in this paper using a variety of metrics. Experiments results are achieved on Pima Indians Diabetes Database. H-KNN algorithm achieves 89% of accuracy, 82% of precision and 77% of recall. In future it can be enhanced and expanded upon to automate the analysis of diabetes using further machine learning techniques.

References

[1] Cho, N.; Shaw, J.; Karuranga, S.; Huang, Y.; Fernandes, J.D.R.; Ohlrogge, A.; Malanda, B. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res. Clin. Pr. 2018, 138, 271–281

[2] Varma, K.V.; Rao, A.A.; Lakshmi, T.S.M.; Rao, P.N. A computational intelligence approach for a better diagnosis of diabetic patients. Comput. Electr. Eng. 2014, 40, 1758–1765.

[3] Fatima, M., Pasha, M., 2017. Survey of Machine Learning Algorithms for Disease Diagnostic. Journal of Intelligent Learning Systems and Applications 09, 1–16. doi:10.4236/jilsa.2017.91001.

[4] Diagnosis and classification of diabetes mellitus Diabetes Care, 32 (Suppl. 1) (2009), pp. S62-S67

[5] Nai-Arun, N., Sittidech, P., 2014. Ensemble Learning Model for Diabetes Classification. Advanced Materials Research 931 - 932, 1427–1431. doi:10.4028/www.scientific.net/AMR.931-932.1427

[6] Han, J., Rodriguez, J.C., Beheshti, M., 2008. Discovering decision tree-based diabetes prediction model, in: International Conference on Advanced Software Engineering and Its Applications, Springer. pp. 99–109.

[7] Bamnote, M.P., G.R., 2014. Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming. Advances in Intelligent Systems and Computing 1, 763–770. doi:10.1007/978-3-319-11933-5.

[8] Maniruzzaman M, Rahman M J, Hasan M, Suri H S, Abedin M M, El-Baz A, and Suri J S, Jan 2020 “Classification and prediction of diabetes disease using machine learning paradigm," J. health information science and system., vol. 42, no. 5, p. 92 -103.

[9] Yuvaraj, N.; SriPreethaa, K.R. Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. Clust. Comput. 2017, 22, 1–9

[10] Sisodia, D.; Sisodia, D.S. Prediction of Diabetes using Classification Algorithms. Procedia Comput. Sci. 2018, 132, 1578–1585

[11] Mamuda, M.; Sathasivam, S. Predicting the survival of diabetes using neural network. In Proceedings of the AIP Conference Proceedings, Bydgoszcz, Poland, 9–11 May 2017; Volume 1870, pp. 40–46

[12] Kayaer, K., Tulay, 2003. Medical diagnosis on Pima Indian diabetes using general regression neural networks, in: Proceedings of the international conference on artificial neural networks and neural information processing (ICANN/ICONIP), pp. 181–184