International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 09 Issue: 10 | Oct 2022 www.irjet.net p-ISSN: 2395-0072

Early Stage Diabetic Disease Prediction and Risk Minimization using Machine Learning Techniques: A Review

Rachna K. Somkunwar

Associate Professor, Department of Computer Engineering
Dr. D. Y. Patil Institute of Technology, Pimpri, Pune-18

Abstract - *The health of the entire population is impacted by diabetes, a chronic illness that continues to be a major global concern. It is a metabolic illness that causes high blood sugar levels and numerous other issues, including heart and nerve issues, stroke, renal failure, and many others. Over the years, numerous researchers have made an effort to develop an accurate diabetes prediction model. But because there aren't enough relevant data sets and prediction strategies, this field still has a lot of unresolved research problems, which forces academics to apply ML-based techniques. In this study, machine learning algorithms have been utilized to diagnose people with diabetes so that we can better care for them. The prediction of diabetic illness patients is taken into consideration. The diabetic disease patient's dataset includes diverse attributes like glucose levels, insulin levels, etc. The dataset in this study is initially examined using established machine learning techniques like decision trees, KNN, and random forests. Decision Tree accuracy was 83.11%, Random Forest accuracy was 88.42%, K-NN accuracy was 60.2%, and SVM (Linear) accuracy was 74.45%. The accuracy of the Decision Tree technique is enhanced from 83.62% to 89.5% by using the Pearson Correlation, while the accuracy of the K-NN approach is increased from 60.2% to 60.8%. In this study, the existing algorithm is improved, and we increase accuracy by about 74% in comparison to the accuracy attained by the original KNN algorithm, which was 60.8%.*

Key Words*:* **Machine Learning, Diabetic Retinopathy, component, Decision Tree, KNN, Pearson correlation, Random Forest**

1. INTRODUCTION

Diabetes is a condition in which there is insufficient insulin, which impairs blood sugar metabolism and causes blood sugar levels to keep rising. Patients with diabetes are unable to efficiently convert the eaten carbohydrates into the glucose sugar needed to fuel daily activities. As a result, the blood sugar level gradually rises. Consequently, glucose doesn't reach all of the body's cells and instead stays in the bloodstream [1].

Diabetes is a metabolic condition that develops when the pancreas fails to produce the required quantity of insulin over an extended period of time. According to WHO, the premature mortality rate from diabetes increased by 5% between 2000 and 2016. From 2000 to 2010, the premature death rate due to diabetes fell or was limited in affluent nations, but rates rose from 2010 to 2016. In contrast, rates rose over the whole period in developing countries. In 2014, 8.5% of individuals and senior citizens aged 18 and older had diabetes. However, up until 2016, 1.6 million deaths or incapacities were directly attributable to diabetes. From 108 million in 1980 to 422 million in 2014, the number of diabetic patients increased. People over the age of 18, the prevalence of diabetes increased globally from 4.70% in 1980 to 8.50% in 2014. (WHO official site) [2].

To identify these fatal diseases, a sophisticated ML-based diagnostic system is needed. Patients with diabetes can be successfully diagnosed at an early stage using an ML-based expert decision system. For the purpose of predicting diabetes, researchers used a variety of different datasets. An adequate dataset with the required features for training and validation is required for ML-based frameworks. The ability of the ML model to predict outcomes properly is increased by choosing pertinent and relevant characteristics from the dataset. The dataset utilized in the suggested system was assembled by the hospital in Sylhet, Bangladesh and is available in the (University of California Irvine) UCI Machine Learning repository [3].

The technologies of machine learning and deep learning are directly related to solving problems in the real world. Machine learning uses classification algorithms to help in diabetes prediction. The various classification methods enable us to distinguish the crucial characteristics that are more strongly associated with the prediction of diabetes. We employed a Convolution neural network from a Deep Learning algorithm for the prediction of diabetic retinopathy. By classifying qualities, we are effectively able to distinguish between dependent and non-dependent traits in patients, allowing us to determine which trait contributes to diabetes and which does not. An correct diagnosis is necessary because the number of diabetic patients is growing day by day over time. In recent years, diabetes has become one of the top causes of death in developing nations. Both the government and private citizens are funding research efforts to find a cure for the serious disease. Finding the best classification system for Diabetes prediction is what drives us to do this.

2. LITERATURE SURVEY

This section reviews the most recent research in the field, provides insights into the problems, and looks for gaps in the methods that are currently used. The literature in this area focuses on the use of machine learning's classification algorithm in the healthcare industry to design and create an effective learning healthcare system for better diabetes diagnosis. Our source for the pertinent data is a literature review.

An adaptive neuro-fuzzy inference system (ANFIS) and principal component analysis were used to describe an intelligent method for improving the accuracy of diabetes diagnosis (PCA). To be more specific, the PCA technique is utilized to minimize the number of characteristics in the diabetic data set. The ANFIS classification approach is crucial for the early detection of diabetes [4]. Reduced attributes provide the foundation of the ANFIS classification model [5].

Disease Prognosis analyzed data using the K-mean algorithm and machine learning. K-mean algorithm is the proposed system's foundation for both structured and unstructured data. A 0.95 accuracy level was attained. [6] The Pima Indians Diabetes Database, which is compiled from data collected by the National Institute of Diabetes and Digestive and Kidney Diseases, is the dataset used for diabetes disease prediction utilizing machine learning on big data in healthcare. The input dataset is initially preprocessed in the recommended approach using the WEKA tool. The Naive Bayes, Support Vector Machine, Random Forest, and Simple CART algorithms for machine learning are employed. The SVM's accuracy, 0.7913, is the highest [7]. Pima Indians Diabetes Database (PIDD), a dataset supplied from the UCI machine learning repository, is used for the prediction of diabetes using classification algorithms. Decision Tree, SVM, and Naive Bayes are the algorithms that are employed. The accuracy that Naive Bayes provides is the highest, at 0.7630. In the future, more diseases may be predicted or detected using the developed system and the machine learning classification algorithms [8]. This paper [9] makes use of the Pima Indians Diabetes Data Set. The updated J48 classifier is used to boost the data mining process' accuracy rate. This section analyses the effectiveness of several classifiers on the input dataset. Also presented is the suggested system that helps to enhance diabetes prediction. The J-48 classifiers were created using the data mining application WEKA as a MATLAB API. J48, Decision Tree, Naive Bayes, Multiclass Classifier, Random Tree, Random Forest, and Multilayer Perception are the algorithms that are employed. The proposed algorithm's accuracy value is 0.9987. More data sets will soon be used to verify the suggested algorithm. [10] An Effective Rule-based Diabetes Classification Using ID3, C4.5, and CART Ensembles. The datasets utilized are the BioStat Diabetes Dataset and the Pima Indian Diabetes Dataset (PIDD) (BDD). Information gain, gain ratio, and gin index are utilized as foundation classifiers in the proposed work to create several decision trees with changing splitting criteria. These distinct classifiers are then blended via a variety of ensemble techniques. Experimental findings indicate that, when compared to alternative decision tree classifiers, the suggested methodology has achieved the highest accuracy. Similar ensemble techniques can be used in the future on datasets related to other diseases as breast cancer, heart disease, and liver illness. [11] A review of current literature reveals that there has been a significant amount of study done on the diagnosis of diabetes. However, the Pima Indian Diabetes Dataset has been used for the majority of the research (PIDD).

3. PROPOSED SYSTEM

This section analyses the effectiveness of several classifiers on the input dataset. Also presented the suggested system that helps to enhance diabetes prediction.

3.1.1. Disease Diagnosis Using Machine Learning

We are motivated to examine the effectiveness of various machine learning algorithms in the prediction of the diabetic condition since effective decision-making by medical professionals is essential. After preprocessing the dataset, the suggested system discussed the various classifiers. Standardization and the elimination of missing values were the preprocessing procedures applied.

3.1.2. Dataset Description

The 15k records in the diabetes dataset, which was acquired from Kaggle, have the following attributes:

- Number of times Pregnant
- Plasma Glucose Concentration
- Skin fold Thickness
- Diastolic Blood Pressure
- 2-hour Serum Insulin
- Body Mass Index
- Diabetes Pedigree Function
- Age (in Years)

It contains a pedigree component that provides information about the family's history of diabetes. The dataset has labels on it. Label 0 indicates someone who is not diabetic, whereas Label 1 indicates a diabetic. Since the dataset is labeled, supervised learning was employed to train the model. We used 30% of the dataset for testing purposes and 70% of the dataset for training purposes for our model.

4. MACHINE LEARNING AND DEEP LEARNING APPROACH

The proposed system, which employs a machine learning technique, used the following steps:

Steps:

1. Initialization of the classifier.
2. Train the classifier: All classifiers in scikit-learn uses a fit(X, Y) method to fit the model (training) for the given train data X and train label Y.
3. Predict the target: Given a non-label observation X, predict(X) returns the predicted label Y.
4. Evaluate the classifier model.

The suggested system is implemented using the Deep Learning methodology in the manner described below:

Steps for Implementation:

1. Collecting the Dataset
2. Importing Libraries and Splitting the Dataset
3. Building the CNN
4. Full Connection
5. Data Augmentation
6. Training our Network
7. Testing

4.1.1 Data Preprocessing

The process of converting raw data into a comprehensible format is known as data preparation. Real-world data is frequently lacking in specific behaviors or patterns, inconsistent, incomplete, and likely to contain several inaccuracies. Such problems are solved via data preparation. Raw data are processed before being further processed. Here are a few preprocessing techniques:

• **Standardization**: A standardization approach is used to modify the dataset so that it has a standard deviation of one in order to increase the accuracy of the Machine Learning Algorithms dataset during preprocessing.

• **Replacing Missing Values**: If a column's missing values are numerical, they can be removed. The average of all the variable's cases was used to replace them.

4.1.2 Feature Selection

The process of selecting features that contribute the most to our prediction variable or our intended output is known as feature selection.

The following methods for feature selection have been employed:

Pearson correlation: We can determine the relationship between two quantities with the use of a coefficient. It provides us with a measurement of how strongly two variables are associated. The Pearson's Correlation Coefficient's value ranges from 0 to 1. 1 denotes a strong link between them, while 0 denotes no association. We are comparing the attributes in our dataset to the results. The characteristics of glucose, insulin, BMI, and the number of pregnancies are closely related to the outcome.

Feature Selection Using Random Forest: Since random forest uses decision trees to categorize test data, it can be used for feature selection as well. In this method, the importance of each attribute is determined by taking into consideration the decision tree's priority, which truncates the dataset.

Recursive feature elimination for feature selection: In this method, the accuracy is checked by selecting a number of attributes in advance. The attributes that provide the best accuracy are determined by adding each attribute one at a time, checking accuracy recursively, and looking at all possible combinations of adding attributes.

4.1.3 Algorithm

Machine Learning Algorithms

For prediction, the following algorithm has been applied:

Decision Tree: For classification tasks where the dataset is partitioned into smaller subsets, decision trees are a supervised learning technique. An related decision tree is created along with the dataset's split. The decision nodes and leaf nodes are present in the finished tree. The classification or judgment value is represented by the leaf node, and the predictor node is represented by the root node.

ALGORITHM STEPS:

1. Pick the best attribute. The best attribute is the one that separates the dataset.
2. Ask the relevant question.
3. Follow the answer path.
4. Go to step 1 until you arrive at your answer.

Random Forest: During training, numerous decision trees are generated using the Random Forest algorithm, and the output is a class that represents the average of all the classes. Over fitting of the dataset by the decision tree is overcome by Random Forest.

ALGORITHM STEPS:

1. Randomly select "k" features from total "m" features

2. Among the "k" features, calculate the node "d" using the best split point

3. Split the node into daughter nodes using the best split 4.

4. Repeat the 1 to 3 steps until "l" number of nodes has been reached

5. Build forest by repeating steps 1 to 4 for "n" number of times to create an "n" number of trees.

K-Nearest Neighbor: The k-nearest neighbor algorithm is one used for supervised classification. The number of nearest neighbors is shown above as "k". This method generates a collection of named points and applies those labels to a new point. The neighbor that is closest to the new point is designated as its label.

ALGORITHM STEPS:

1. Determine parameter k = nearest neighbor.

2.Calculate the distance between the query instance and allthe training samples.

3.Sort the distance and determine nearest neighbors based on k-th minimum distance.

4. Collect the groups of nearest neighbors.

Support Vector Machine: The discriminative classifier known as the support vector machine, or SVM for short, is used for classification tasks. Finding the hyper plane in an N-dimensional space, where N is the number of features, and using that hyper plane to clearly classify the characteristics, is the basic objective of SVM. Any shape, such as a line or a curve, can be the hyper plane. SVM algorithms including linear, polynomial, and radial basis function SVM are employed.

Linear case: We should now consider the case of two classes' problem with N training samples. Each sample is described by a Support Vector (SV) Xi composed by the different "band" with n dimensions. The label of a sample is Yi. For a two classes' case, we consider the label -1 for the first class and +1 for the other. The SVM classifier consists in defining the function : (f (x) = sign((ω, X) + b)) which finds the optimum separating hyper plane, where ω is normal to the hyper plane, and b / ω is the perpendicular distance from hyper-plane to the origin.

Non-Linear case: If the case is nonlinear as the first solution is to make a soft margin that is particularly adapted to noise data. The second solution that is the particularity of SVM is to use a kernel. The kernel is a function that simulates the projection of the initial data in feature space with a higher dimension θ : KnH. In this new space the data are considered as linearly separable. To apply this, the dot product (xi, xj)) is replaced by the function: K(x, xi) = (φ(x), φ(xi)) Then the new function to classify the data are: f (x) = sign(Σβi.αi.K(x, xi)NSi= 1 + b) . Kernels are commonly used: The polynomial kernel: K(x, xi) = ((x.xi) + 1)p . The sigmoid kernel: K(x, xi) = tanh((x.xi) + 1).

5. MODIFIED ALGORITHM

Modified Algorithm (Proposed Algorithm)

Maximum class predictions performed on a test data are applied to that class, which is chosen as the class for the testing data based on predictions made by other algorithms on test data. KNN, Decision Tree, and Random Forest algorithms are employed in that voting process. This is a combination of these three algorithms since one of the algorithms that incorrectly predicted the test data could be correctly predicted by another algorithm.

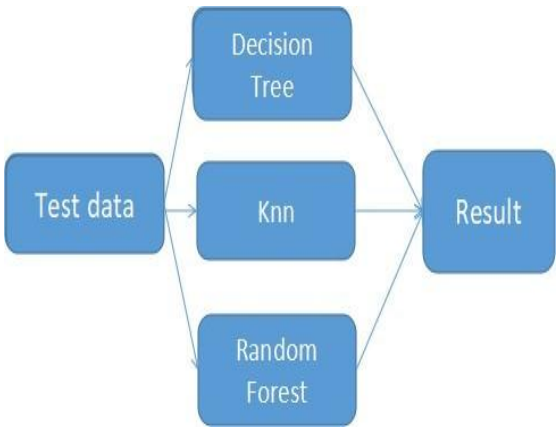

Fig -1 : Abstract of Proposed Algorithm

The more accurate version would define a function based on the accuracy of these algorithms to take the importance of the vote into account, meaning that rather than assuming that each vote is equally important, the vote with the highest accuracy should be given priority by defining a suitable function. However, the basis for this suggested method is the equity of the votes cast by three algorithms. Fig. 1 depicts the suggested algorithm's abstract.

We are motivated to examine the effectiveness of various machine learning algorithms in the prediction of Diabetes since effective medical professionals need to make decisions.

After preprocessing the dataset, we implemented various classifiers in the suggested system. Here, normalization and the elimination of missing values were utilized as preprocessing approaches.

6. SIGNIFICANCE OF PEARSON CORRELATION

We can determine the link between two quantities using Pearson's correlation coefficient. It provides us with a measurement of how strongly two variables are associated. The Pearson Correlation Coefficient's value ranges from -1 to +1. 1 denotes a strong link between them, while 0 denotes no association. We are comparing the attributes in our dataset to the results. The characteristics of glucose, insulin, BMI, and the number of pregnancies are closely related to the outcome.

7. RESULTS

The accuracies achieved by applying the algorithms Decision Tree, Random Forest, K-NN, SVM(Linear), SVM(Radial), SVM-RFE(Recursive Feature Elimination) on the preprocessed diabetic dataset are 83.62%, 93.31%, 60.2%, 74.45%, 64.93%, 74.45% respectively shown in the figure 2.

The accuracy rates for the methods Decision Tree, Random Forest, K-NN, and SVM (Linear) after applying Pearson Correlation to the diabetes dataset are, respectively, 89.5%, 92.64%, 60.8%, and 74.02%.

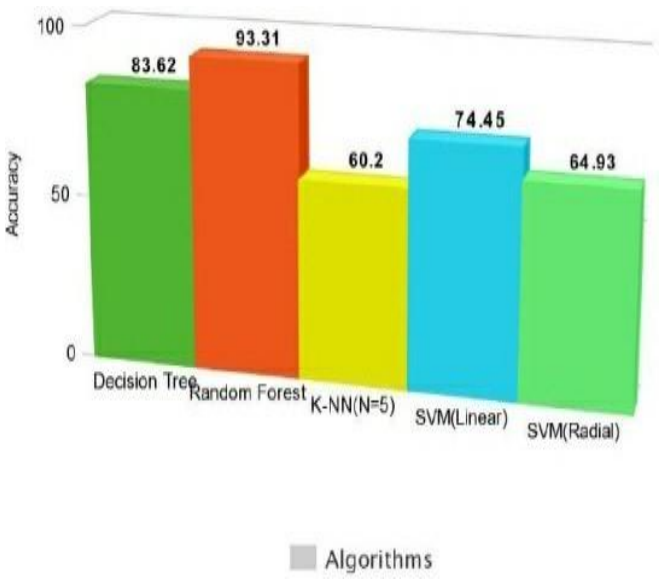

Fig -2 : Histogram depicting the accuracy of several algorithms

The accuracy rates obtained using the methods Decision Tree, Random Forest, K-NN, and SVM(Linear) on the diabetes dataset are 83.11%, 88.42%, 60.2%, and 74.45%, respectively. Random Forest has the maximum accuracy (93.31%), which is obtained. After using the Pearson Correlation, the accuracy of the Decision Tree is improved from 83.62% to 89.5%. After using the Pearson Correlation, the accuracy gained in the case of K- NN increased from 60.2% to 60.8%. The updated (proposed) algorithm's accuracy is estimated to be 74%. The proposed algorithm achieves an accuracy of 88%.

8. CONCLUSIONS

The use of ML technique is thought to be beneficial in disease diagnosis. The patients benefit from early diagnosis and treatment. In this article, a few existing machine learning (ML) classification models for the accurate prediction of diabetic patients have been discussed. On the classification issue, an expression of correctness has been found. In this study, diabetes disease has been predicted using classifier machine learning techniques such Decision Tree, Random Forest, K-NN, and SVM (Linear and Radial). In comparison to the other algorithms, the Random forest performs better overall at predicting diabetic illness.

REFERENCES

[1] T. M. Alama, M. A. Iqbala, Y. Ali et al., "A Model for Early Prediction of Diabetes," Informatics in Medicine Unlocked, vol. 16, Article ID 100204, 2019.

[2] Srivastava, Rashmi, and Rajendra Kumar Dwivedi. "A Survey on Diabetes Mellitus Prediction using Machine Learning Algorithms." ICT Systems and Sustainability. Springer, Singapore, 2022. 473-480.

[3] A. Frank and A. Asuncion, UCI Machine Learning Repository, Oct. 2010.

[4] K. Dwivedi, "Analysis of decision tree for diabetes prediction," International Journal of Engineering and Technical Research, vol. 9, 2019.

[5] K. Polat and S. Güneş, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," Digital Signal Processing, vol. 17, no. 4, pp. 702–710, 2007.

[6] Akash C. Jamgade, Prof. S. D. Zade, Disease Prediction Using Machine Learning,IRJET,2019

[7] Ayman Mir,Sudhir N. Dhage ,Diabetes Disease Prediction using Machine Learning on Big Data of Healthcare,IEEE ,2018

[8] Deepti Sisodia ,Dilip Singh Sisodia, Prediction of Diabetes using Classification Algorithms, ICCIDS, 2018

[9] .Gaganjot Kaur, Amit Chhabra, Improved J48 Classification Algorithm for the Prediction of Diabetes, IJCA, 2014

[10] Saba Bashir,Usman Qamar, Farhan Hassan Khan, M.YounusJaved , An Efficient Rule-based Classification of Diabetes Using ID3, C4.5 and CART Ensembles, IEEE, 2014

[11] Supaporn Phetarvut, Nantiya Watthayu, Nantawon, Suwonnaroop, Factors Predicting Diabetes Selfmanagement Behavior among Patients with Diabetes Mellitus Type 2,Journal of Nursing Science , 2011.