IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 1, January 2022

DOI: 10.17148/IJARCCE.2022.11149

Hybrid Prediction System for Diabetes Disease Using Type II Dataset

Pradeep Pal[1], Swapnil Waghela[2]

PG Student, CSE, SKITM, Indore, India[1]

Assistant Professor, CSE, SKITM, Indore, India[2]

Abstract: Data mining for healthcare is an interdisciplinary subject of research that has its roots in database statistics and may be used to assess the efficacy of medical treatments.. Diabetes is a chronic disease in which the pancreas fails to make adequate insulin or body fails to utilize the insulin that is produced appropriately. In the health-care business, data analysis plays a critical role in illness identification. The proposed research is being carried out to compare the performance of various classifiers in the Ada-Boost learning environment. We employed three distinct algorithms in this regard: BPN (back propagation neural network), SVM (support vector machine), C4.5 decision tree, and classifier. The ada-boost learning approach is used to train all of the algorithms. The diabetic disease type dataset from the UCI machine learning data repository is utilized in CSV format for training and testing developed classifiers. Hence, we implemented proposed hybrid classification system using the JAVA WEKA machine learning library. The performance of the system for all the classifiers is calculated and compared in terms of Accuracy, Error Rate, Time and Memory usages based on various experiments and datasets. Furthermore we also compared our system to the traditional base Decision Stump method.

Keywords: Data Mining, Machine Learning, Classification, Dataset, ensemble learning, boosting, Ada-Boost, diabetes disease, Prediction.

I. INTRODUCTION

Data mining [1] techniques are the collection of different computational algorithms that learn of the historical data and based on the patterns of data tried to make decisions and predict the values. In this context the data mining techniques can be used in various kinds of applications where the data analysis is essential to understand the data patterns. This research work demonstrates the use of data mining technique is performed in medical domain. Using the data mining techniques here we tried to predict the diabetes disease possibility to any end user. In order to perform such task the supervised learning technique is applied on data [2] [3] [4].

The data is collection of disease symptoms as the attributes of dataset and the classes of the data instances reflect the possibility of the diabetes disease. In this data first the classification algorithm is employed and the algorithm makes the predictive model [5]. After the training of proposed model testing dataset is applied on data. The testing data attributes are processed using the trained data model for predicting the class labels of the individual data instances in testing dataset. There are different kinds of data prediction models [6] are available and provide the effective performance but the performance of prediction is not much accurate therefore in this presented work the ensemble learning technique is used for improving the accuracy of the classification algorithm.

II. LITERATURE SURVEY

Our research was inspired by extensive literature on diabetes condition detection utilising data mining methods. Researchers in the medical area use data mining methods to identify and forecast illnesses as well as provide appropriate treatment for patients. As a result, previous diabetes illness prediction research is included in this section.

L Talha Mahboob Alam et al. [7] used important features to predict diabetes, and the link between the various traits was also described. For diabetes, a variety of methods are utilised to assess relevant attribute selection, grouping, prediction, and association rule mining. The principal component analysis approach was used to pick significant features.

Fikirte Girma Woldemichael et al. [8] advocated utilizing data mining approaches to forecast diabetes. The back propagation technique is used to determine whether or not a person is diabetic. To predict diabetes, J48, naive bayes, and support vector machine were also utilized. These neural networks had an input layer with eight parameters, a hidden layer with six neurons, and an output layer with six neurons. The model's performance was improved using a 5-fold cross-validation approach and high value learning rate.

DOI: 10.17148/IJARCCE.2022.11149

Deepti Sisodia et al. [9] presented this research with the goal of developing a model that can accurately predict the risk of diabetes in individuals. To diagnose diabetes at an early stage, this experiment employs three machine learning classification algorithms: Decision Tree, SVM, and Naive Bayes. The Pima Indians Diabetes Database (PIDD), which is supplied from the UCI machine learning repository, is used in the experiments. Precision, Accuracy, F-Measure, and Recall are all used to assess the performance of the three algorithms.

Outlier rejection, data standardisation, feature selection, K-fold cross-validation, and various Machine Learning (ML) classifiers (k-nearest Neighbour, Decision Trees, Random Forest, Ada-Boost, Naive Bayes, and XGBoost) and Multilayer Perceptron (MLP) were used by **MD. Kamrul Hasan et al. [10]** to propose a robust framework for diabetes prediction. Using the Pima Indian Diabetes Dataset, all of the suggested experiments were carried out under the identical experimental settings.

This suggested ensemble classifier exceeds state-of-the-art findings by 2.00 percent in AUC in all of the extended trials, with sensitivity, specificity, false omission rate, diagnostic odds ratio, and AUC of 0.789, 0.934, 0.092, 66.234, and 0.950, respectively.

Suyash Srivastava et al. [11] suggested a diabetes prediction model based on machine learning. Various academics have attempted to forecast the diabetes machine learning algorithm, but this is a separate effort in the study work based on a certain kind of patient in a given community. A Pima Indian data sample was used to forecast risk of diabetes. Artificial Neural Network (ANN) was selected from among many Machine Learning techniques for developing a diabetes prediction model.

III. PROPOSED SYSTEM

The suggested approach for boosting prediction accuracy is based on the ensemble learning technique is demonstrated in this section.

A. Domain Overview

Machine learning and Data mining methods are used for data analysis-based prediction, classification, and other applications. This approach use computer algorithms to comprehend data patterns and forecast probable values based on the patterns learnt from training samples. The implementation of a diabetic illness prediction system is attempted in this study. The supervised learning approach is employed in this case. The supervised learning algorithms can learn from standard training samples and anticipate similar patterns of data once they've been trained on them. As a result, many types of data mining prediction models are initially deployed. SVM, C4.5 or J48 decision tree, and BPN are examples of prediction models (back propagation neural network).

These models are trained on the diabetic disease dataset before being used to classify the test dataset. The class labels for test dataset objects are predicted during the prediction process. Although all of the developed classifiers provide acceptable results during prediction, there is a potential that the performance of a classifier may be improved further. As a result, the classifier is trained and tested using the ensemble learning approach. Ensemble learning is a strategy that involves creating numerous models for a single classifier and then combining all of the models into a single instance to improve classification results.

Basically, as compared to a single instance of a classifier, the ensemble model delivers more accurate findings; hence ensemble learning outcomes are preferable to the standard classifier implementation. The fundamental overview of the proposed approach is provided in this part, and the full procedure of the suggested technique is outlined in the next section.

B. Methodology Development

Figure 3.1 depicts the suggested system design. The many phases of the proposed categorization technique's operation are shown in this figure. Essentially, the proposed work entails the implementation of several classification algorithms for completing the diabetic illness prediction job, as well as the scaling of the built classifier's performance for performance comparison.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 1, January 2022

DOI: 10.17148/IJARCCE.2022.11149

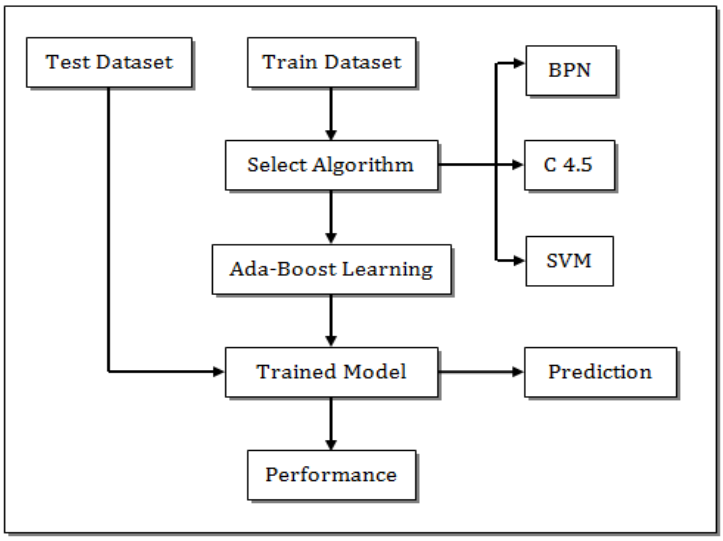

Fig. 1 Working Flow

Training Dataset: The supervised learning algorithm is used in data mining and machine learning systems to analyse data. To begin, some sample patterns must be provided as input. This data is referred to as a sample dataset or training data samples in this context. The training set for UCI data repository's diabetic disease dataset is used to train data mining techniques. The diabetic illness dataset includes symptoms as dataset attributes and class labels as an additional element.

Algorithm Selection: In data mining, SVM is a supervised learning technique. This is used to categories data samples into two main groups. The data samples are treated as a point in n-dimensional space in this learning method. In addition, the characteristics are treated as data instance coordinates in this case. This technique entails making an effort or doing calculations in order to draw a hyper plane between these coordinates in order to categorise the two data instances. The location of the hyper plane determines the categorization of two types of data samples.

C4.5 decision tree: The J48 classification algorithm is another name for the C4.5 decision tree method. The method is an improved version of the ID3 decision tree algorithm. The training samples are initially mapped into the decision structure in this classification approach. The class labels are put in leaf nodes, while the remaining characteristics are placed in the tree's branches. The test samples are utilized to traverse the tree structure in order to discover the leaf node during decision tree testing. As a result of categorization, the acquired tree leaf node is anticipated.

BPN: BPN (back propagation neural network) is a well-known machine learning and data mining approach. That method uses a weighted algorithm to estimate the values of input samples in the absence of a predefined class label. For data analysis and prediction, the overall approach has three levels. The input samples are accepted in the first layer, which is referred to as input layer; the second layer, which is referred to as the hidden layer, is used to compute weights and changes in weights. Finally, the output is collected by creating a new layer called the output layer.

Ada-Boost Learning: The Ada-boost approach is used in machine learning to improve or boost performance of a week classifier. To comprehend how this boosting strategy works, consider the following functions of a classifier:

$$H(x) = \text{sign}(\sum_{t=1}^{T} \alpha_t h_t(x))$$

Where the classifier consist of a set of T week classifiers, and the $h_t(x)$ is the output of classifier t. Similarly α_t is a weight which is determined by Ada-boost.

The classifiers are now being trained one sample at a time. The chance of individual samples appearing in a training sample increased when the training method was updated. In this case, the first classifier t is trained with equal probability throughout the whole training set, and the output for that trained classifier is computed as follows:

$$\alpha_t = \frac{1}{2} \ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

The output weight α_t is based on classifier's error rate ε_t. That is the ratio of misclassified samples over the total number of instances to be train.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 1, January 2022

DOI: 10.17148/IJARCCE.2022.11149

After computing the weight for the first classifier the update on training set weight is performed using the following formula:

$$D_{t+1} = \frac{D_t(i)\exp\left(-\alpha_t y_i h_t(x_i)\right)}{Z_t}$$

D_t is the vector that contains one weight for all the training sample instances. And Z_t is the sum of all the weights in training sample.
As a result, the chosen classifier is trained using the Ada-Boost approach in order to improve classification accuracy.

Training Model: The classifier is trained using the input training samples and returns a learned data model for prediction.

Test Dataset: For predicting the samples class labels, the trained data model accepts comparable testing patterns in the same manner as the training sample.

Prediction: The projected class labels for the input test samples are the system's final result, and performance is calculated using them.

Performance: Performance of classifiers is measured based on the classified data instances and their accuracy of categorization. At addition, in this phase, the complexity of algorithms is assessed and reported in terms of time and memory utilized.

C. **Proposed Algorithm**

This section outlines the phases in the process of training and testing the prediction system.

TABLE I PROPOSED ALGORITHM

Input: Train Data Samples ($Train_D$), Test Data Samples ($Test_D$) **Output:** class labels for Ts →C
Process: 1. $R_n = readTrainData(Train_D)$ 2. $CL = selectClassifier(i)$ 3. $T_{model} = AdaBoost.Train(CL, R_n)$ 4. $for(i = 1; i \leq (Test_D.Length; i++)$ a. $C_i = T_{model}.Classify(Test_{D_i})$ 5. end for 6. Return C

IV. RESULT DISCUSSION

This section summarizes the many parameters that were assessed throughout the experiment.**Accuracy**
The accuracy of a data model for classification or prediction may be determined by measuring its accuracy. The accuracy is essentially a ratio of the classifier's successfully differentiated patterns to the total patterns available for categorization. The formula below may be used to estimate this.

$$\text{Accuracy} = \frac{\text{Correctly Classified Sample}}{\text{Total Sample to Classify}} X100$$

The performance of the proposed diabetes disease prediction data model is represented in figure 2. The given performance of the system is estimated and compare to other classifiers. For graphical representation of the performance the X axis of graph contains the different experiments and Y-axis of graph depicts accuracy value of all classifiers performance. Our proposed hybrid method using different classifier depicting higher performance compared to base method. According to the results with the different instances of data the accuracy of the proposed model is clearly higher accuracy of the disease prediction. Thus the proposed system is acceptable for real world use of application.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 1, January 2022

DOI: 10.17148/IJARCCE.2022.11149

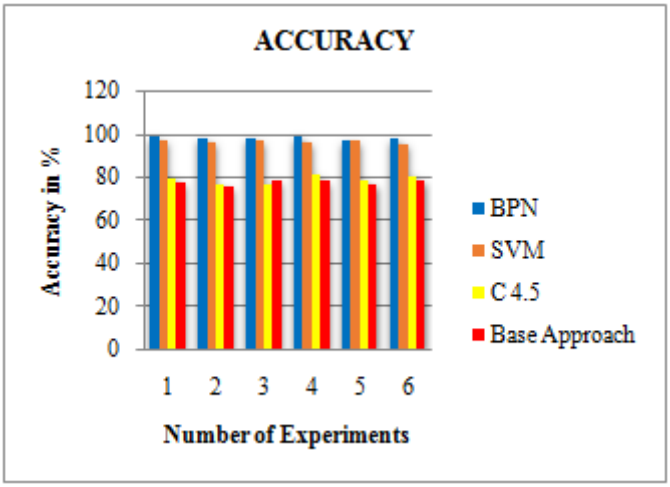

Fig. 2 Accuracy

D. **Error Rate**

Error rate is the percentage of data that is misclassified or misidentified during classification. The error rate is a metric for how accurate a categorization data model is. The following formula is used to calculate the algorithm's error rate:

$$\text{Error Rate} = 100 - \text{Accuracy}$$

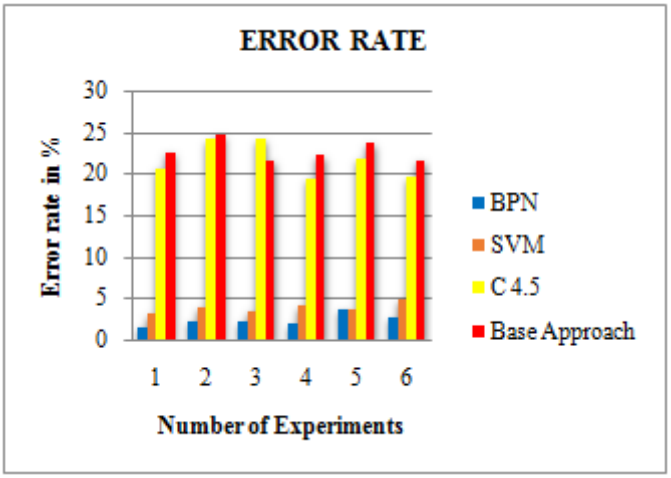

Fig. 3 Error Rate

Figure 3 show the % error rate and base findings of the proposed hybrid diabetic illness prediction system. The Y axis of the graph depicts the algorithm's calculated error rate, while the X axis depicts the experiments carried out. Proposed classifier is shown by a blue line, while SVM, C4.5, and base are represented by orange, yellow, and red colours, respectively. According to the results, the suggested model produces a lower data categorization error rate. As a result, the hybrid technique is effective and precise in disease prediction.

E. **Memory Usage**

When calculations are done in any computational system, the data is stored in main memory and used by the process. Memory consumption or algorithm space complexity refers to the amount of space needed to store data in main memory. The following formula may be used to determine the memory requirement:

$$\text{Memory Consumption} = \text{Total Memory} - \text{Free Memory}$$

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 7.39 ⁜ Vol. 11, Issue 1, January 2022

DOI: 10.17148/IJARCCE.2022.11149

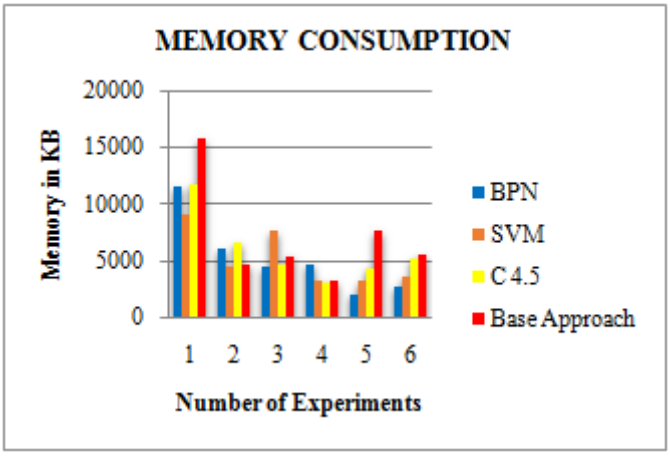

Fig. 4 Memory Consumption

Figure 4 show the memory needs for the suggested diabetic illness prediction. The performance of the suggested strategy is shown by the blue line in this figure. To show the performance, the X axis contains the data experiments and predictions, while the Y axis contains the corresponding obtained memory needs. Memory is measured in units of kilobytes (KB) (kilobytes). According to the results, the hybrid system's performance consumes less memory while processing categorization tasks. As a result, the diabetic illness prediction system suggested is more effective and efficient.

F. **Time Consumption**

The method takes some time to analyze the data based on the data models and data input size. The time complexity or time requirements of the algorithm refer to the time needs of the algorithm. The following formula is used to calculate the time:

$$\text{Time Required} = \text{End Time} - \text{Start Time}$$

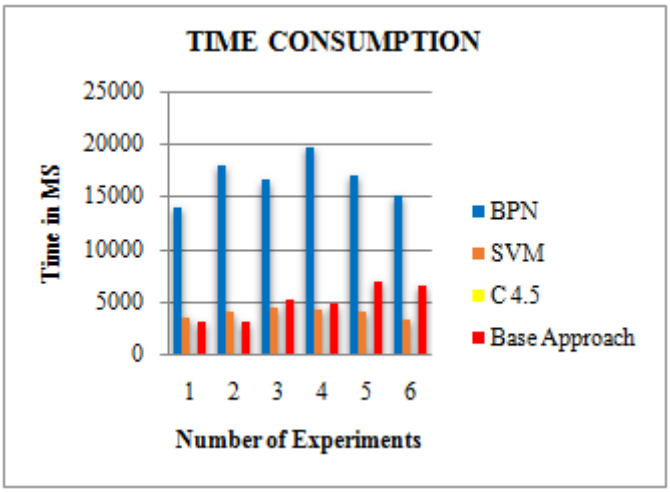

Fig. 5 Time Consumption

The time requirements of the methods, namely the suggested approach, are shown in picture 5. The given time is measured here in milliseconds (MS). The suggested classifier for diabetes prediction system's time consumption is shown by blue line on the graph. The dataset instance is shown on the X-axis, while the algorithm processing time is shown on the Y-axis.

V. CONCLUSION

Data mining and machine learning techniques have the capacity to predict. It is necessary to train the algorithms with appropriate instances in order to make them predictable. These samples are essentially pre-examined features of any item. Using these instances, the learning algorithm is taught, and during the test set input, it gives an appropriate result

DOI: 10.17148/IJARCCE.2022.11149

as a sample prediction. This capability of data mining approach is illustrated using a diabetic illness prediction system as an example. The Ada-boost training approach is used to improve the performance of classifiers in the suggested study. The Ada-boost method calculates weights for all training samples and modifies them based on the training.
Three types of classification algorithms are used in this suggested work: BPN (back propagation neural network), SVM (support vector machine), and C4.5 algorithm. All of the classifiers are combined with the Ada-Boost learning approach in order to train more effectively and perform better than before. In this context, a diabetic illness dataset from the UCI dataset repository is used for training and testing. And the prediction's whole modeling is completed. Furthermore, we compared our technique to the basic method, Decision Stump.

VI. FUTURE WORKS

The proposed work's major goal is to improve the classification accuracy of classifiers for predicting diabetes illness based on sample input. The work is scheduled to be extended in the future.

1. For Ada-boost training, the present approach uses a single classifier; however, in the near future, a group of classifiers will be employed.
2. The boosting strategy is examined in this study, and the bagging technique will be applied for performance enhancement in the near future.

REFERENCES

[1]. Chapter 3: Data Mining: an Overview, available online at: http://shodhganga.inflibnet.ac.in/bitstream/10603/11075/7/07_chapter3.pdf
[2]. Bhatia, P. (2019). Data mining and data warehousing: principles and practical techniques. Cambridge University Press.
[3]. Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006
[4]. Gultepe, Y., & Rashed, S. (2019). The use of data mining techniques in heart disease prediction. vol, 8, PP. 136-141.
[5]. Mirzajani, S. S. (2018). Prediction and diagnosis of diabetes by using data mining techniques. Avicenna Journal of Medical Biochemistry, 6(1), PP. 3-7.
[6]. Ristevski, B., & Chen, M. (2018). Big data analytics in medicine and healthcare. Journal of integrative bioinformatics, 15(3).
[7]. Alam, T. M., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., Baig, T. I., & Abbas, Z. (2019). A model for early prediction of diabetes. Informatics in Medicine Unlocked, 16, 100204.
[8]. Woldemichael, F. G., & Menaria, S. (2018, May). Prediction of diabetes using data mining techniques. In 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI), PP. 414-418, IEEE.
[9]. Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. Procedia computer science, PP. 132, 1578-1585.
[10]. Hasan, M. K., Alam, M. A., Das, D., Hossain, E., & Hasan, M. (2020). Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access, 8, PP. 76516-76531.
[11]. Srivastava, S., Sharma, L., Sharma, V., Kumar, A., & Darbari, H. (2019). Prediction of diabetes using artificial neural network approach. In Engineering Vibration, Communication and Information Processing, PP. 679-687, Springer, Singapore.