FUDMA Journal of Sciences (FJS)
ISSN online: 2616-1370
ISSN print: 2645 - 2944
Vol. 7 No. 1, February, 2023, pp 253 - 260
DOI: https://doi.org/10.33003/fjs-2023-0701-1301

A DIABETES PREDICTION CLASSIFIER MODEL USING NAIVE BAYES ALGORITHM

***Folasade Mercy Okikiola, Olumide Sunday Adewale, and Olumide Olayinka Obe**

Department of Computer Science, School of Computing, Federal University of Technology, Akure, Ondo State. Nigeria

*Corresponding authors' email: sade.mercy@yahoo.com

ABSTRACT

One serious health condition which has made people to suffer from uncontrollable high blood sugar is diabetes. The problems of existing detection approaches are data imbalance, feature selection, and lack of generic framework for diabetes classification. In this research, developed an ontology-based diabetes classification model using naïve Bayes classifier was developed. The model is divided into five modules: data collection, feature selection, ontology construction, classification, and document query. The data collection module adapted PIMA Indian Diabetes Database to predict diabetes. The feature selection module employed multi-step approach for selecting the most important features from dataset. For automatically constructing ontology rules based on the chosen features, the ontology generation module used a decision tree classifier. Based on the user's question, the classification module employed a Nave Bayes classifier to automatically classify the built ontology as having diabetes. Based on the ontology-based nave Bayes classification, the document query module searches and returns the anticipated documents requested by users. The proposed model using a 10-fold cross validation performed better in diabetes in precision, accuracy, recall and F1-score of 96.5%, 93.55%, 79.2% and 87.0%, respectively. Benchmarking tools included K-Nearest Neighbor (KNN), Decision Tree (DT), Multilayer Perceptron (MLP), Logistic Regression (LR), Hidden Markov Model (HMM), Support Vector Machine (SVM), Naive Bayes (NB), Random Forest (RF), and Deep Convolutional Neural Network (DCNN). With an area of 0.9578 in compared to other relevant methods, the created model suggested a more accurate test. They demonstrated that the model's cost-effectiveness for predicting diabetes outweighs its value.

Keywords: Classification, Decision Tree, Diabetes, Naïve Bayes, Ontology

INTRODUCTION

Diabetes Mellitus is a long-term illness that causes the body's improper digestion of carbohydrates, which raises blood glucose levels in humans as a result of reduced insulin production (DM). In general, diabetes is a serious medical illness that results in a person's uncontrollably high blood sugar levels (Kushwaha, J. S., Gupta, V. K., Singh, A., & Giri, 2022; Pranata et al., 2021). The most common symptoms of this are increase thirst, increase hunger, and regular urination. Age, obesity, inactivity, hereditary factors, way of life, height, weight, and high blood pressure, and poor diet are among the causes of diabetes (Parveen, S., Patre, P., & Minj, 2023). Detecting diabetes early is the only remedy to avoid complications (Ahlqvist et al., 2018; Kumar & Christopher, 2016), and ensure timely and effective treatments (Dremin et al., 2021; Ogurtsova et al., 2022; Vijayan & Anjali, 2016). Late detection of diabetes could cause high risk of diseases such as kidney disease, stroke, eye problem, and nerve damage. According to the 2017 statistics, about 425 million people are diagnosed with diabetes. The annual death rate due to diabetes condition has been estimated around 2-5 million patients. Insulin-Dependent Diabetes Mellitus (IDDM), another name for type 1 diabetes, is characterised by insufficient insulin synthesis by the body, necessitating the injection of insulin into the patient. Type-2 diabetes, also known as Non-Insulin-Dependent Diabetes Mellitus, is characterised by the body's cells' inability to properly utilise insulin (NIDDM). Type-3 diabetes, also known as gestational diabetes (GD), is characterized by a rise in blood sugar levels in pregnant women as a result of a delayed diagnosis.

Numerous other diseases' problems have their roots in the ignorance of diabetes and its symptoms. This emphasizes how crucial it is to create a predictive model that can automatically generate knowledge about diabetes and its forecasts. The methods currently in use for predicting diabetes have a number of issues. The challenging task of creating techniques for the early identification and prediction of diabetes is one of the issues (Hatua et al., 2021; Komi et al., 2017). Moreover, manual ontology development, which is difficult and time-consuming, is still used to gather domain knowledge about a topic (Kiv et al., 2022; Yun et al., 2021). The manual ontology generation process is more difficult due to a particular topic's potential size, complexity, and dynamic nature (Yun et al., 2021). Another issue is the accurate creation of a general domain ontology that can identify diabetes in patients (Thakkar et al., 2021). In this research, an ontology-based diabetes classification model using naïve Bayes classifier was developed. The most crucial features from the diabetes dataset were chosen using a multi-step feature selection process. When diabetes is not identified early enough for effective and prompt treatment, it is one of the chronic health disorders that worsens over time and eventually results in mortality (Bhutta et al., 2021). Research has showed that diabetes is a root cause for many other disease complications (Mandal et al., 2021). Different ML and ontology-based techniques have recently been used in medical science to design an automated system that can detect diabetes in patients. There are various conventional techniques for predicting diabetes that make use of physical and chemical examinations (El Massari et al., 2022). However, detecting and predicting diabetes early is quite difficult for practitioners in the medical field (Hatua et al., 2021; Komi et al., 2017). Therefore, there is need for a structured and intelligent model to query knowledge about diabetes disease and its symptoms for early prediction. The use of an automated knowledge sharing method for identifying diseases early can help people to understand how to avoid diabetes and how they can completely manage it. Also, a lot of information about diabetes patients' medical histories is created, and smart approaches can be utilized to retrieve this crucial data for diabetes prediction. The lack of knowledge about diabetes and its symptoms is a root cause for

many other disease complications. This research aims to develop an ontology-based diabetes classification model using naïve Bayes classifier. The objectives are to design an ontology-based diabetes diagnosis system using Naive Bayes and decision tree model, implement the ontology-based model and evaluate the performance above using confusion matrix, accuracy and efficiency. Section 2 explains the works related to ontology-based medical diagnosis and artificial intelligence. I In Section 3, the research approach is made clear. In Section 4, the outcomes of the application of this research are displayed. The study's results are shown in Section 5.

Related Works

There are now studies being done in the area of diabetes prediction systems based on ontologies. Some of these articles were reviewed and discussed in this session. Krishnamoorthi et al. (Krishnamoorthi et al., 2022) presented a novel diabetes healthcare disease prediction framework using machine learning techniques. The presented framework is divided into four phases. For the research's data collection phase, the Pima Indian Diabetes database was altered to make diabetes predictions. Data visualisation is utilised in the second phase to highlight the proportion of people who have diabetic issues and makes the data easier to interpret by showing it as a bar chart. The third stage, preprocessing, comprises eliminating outliers and standardizing the data. The classification phase, the fourth, is where diabetes is diagnosed using various machine learning techniques. The authors adopted machine learning algorithms because of their simplicity and popularity. The Pima Indian Diabetes database simulation results demonstrated that logistic regression outperformed alternative machine learning algorithms. Using association rule mining, the data also demonstrated a substantial correlation between glucose and BMI and diabetes. The use of a structured dataset, the absence of feature selection techniques, and the lack of a universal model for automatic diabetes prediction are the study's main weaknesses. A random forest classifier strategy was created by Oza & Bokhare (Oza, A., & Bokhare, 2022) developed a diabetes prediction using logistic regression and k-nearest neighbour. For the purpose of predicting diabetes, the study used well-known machine learning techniques including K-nearest neighbor and logistic regression. The performance and accuracy results from the two independent machine learning models were obtained and compared to determine the best model for the prediction of diabetes. The authors showed that the adapted machine learning models performed competitively with logistic regression the better model for the prediction of diabetes. The study provide good accuracy and high confidence level in the prediction of diabetes. Although, there is no generic model for automatic diabetes prediction and the authors did not state explicitly the feature selection method, number, and type of features selected.

Ranjitha et al. (Ranjitha et al., 2022) developed a diabetes prediction model using artificial neural network. The study adapted the use of the PIDD for their experimentations. The study modified a back propagation algorithm-based ANN model for predicting diabetes. The training and testing sets were created from the customized PIDD dataset. The ANN network was constructed using several neuron types at different epochs, and the outcomes were compared with comparable models. It was observed that the accuracy of the adapted ANN reaches up to 99.23% compared to related methods. The developed method provides high accuracy and it is robust to error. Although, there is no generic model for automatic diabetes prediction and the time complexity of the model is relatively high. Alex et al. (Alex et al., 2022) developed a deep convolutional neural network (DCNN) for diabetes mellitus prediction. The PIDD dataset was modified for the study's experiments. First, by applying the oversampling method, the impact of imbalance class on prediction accuracy was reduced (SMOTE). Afterwards, predictions are produced using a DCNN classifier, and they are evaluated using a particular set of evaluation markers. The results of the developed DCNN algorithm for diabetes mellitus prediction showed better and superior accuracy results when compared to the other related models. Although, there is no generic model for automatic diabetes prediction and the authors did not state explicitly the feature selection method, number, and type of features selected.

In the articles reviewed, no generic model for automatic diabetes prediction and the efficiency of the approaches were low. This research would help address this gap.

METHODOLOGY

An ontology-based diabetes classification model utilising a naive Bayes classifier was created in this study. Figure 1 shows the five components that make up the created model: data collection, feature selection, ontology development, classification, and document query. The popular PIDD was modified for the diabetes prediction by the data gathering module.

Data Collection Module

The dataset used for the developed model is PIMA Indian Diabetes Database (PIDD) available at (*PIMA Indian Diabetes Database*, n.d.). 768 diabetic individuals' records are included in the PIDD, which was originally created by the National Institute of Diabetes and Digestive and Kidney Diseases. Diabetes was the outcome examined; 258 people tested positive and 500 people tested negative. Therefore, the dataset consists of one target (dependent) variable and 8 attributes. The dataset was pre-processed into the Attribute Relation File Format (ARFF) suitable for the machine learning algorithms.

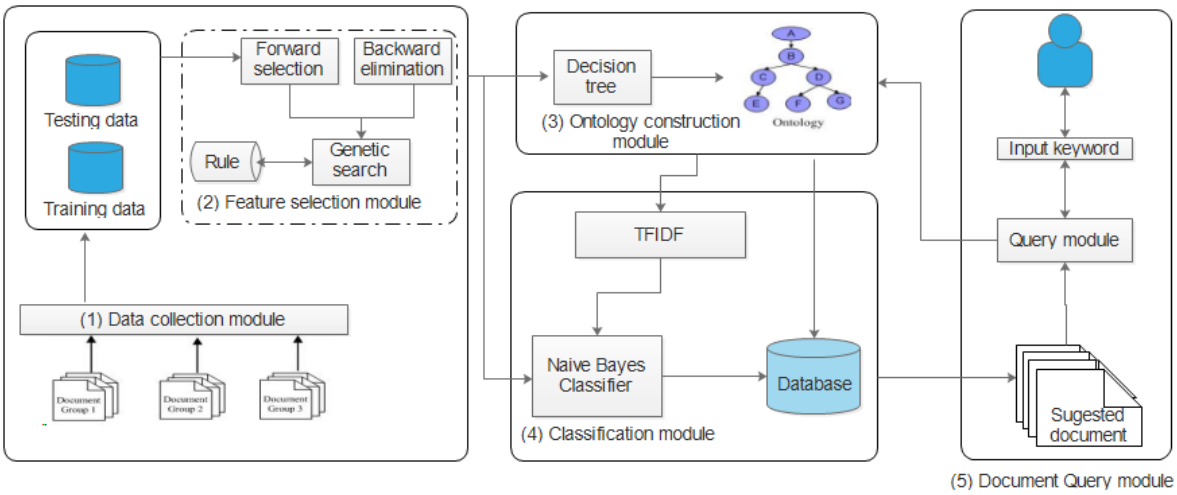

Figure 1: Architecture of the ontology-based diabetes classification model using naïve Bayes classifier

Feature selection module

The preprocessed dataset's most crucial characteristics were chosen by the feature selection module using a multi-step feature selection process. The multi-step selection method consists of Forward Selection, Backward Elimination, Genetic Search, and Rule Evaluation (FS-BE-GS-RulEval) methods.

Definition 1 (Forward selection**):** Starting from the empty set (Equation 1), sequentially select the next best feature X^+ that results in the highest accuracy J(Y_k+ X^+) when combined with the features Y_k that have already been selected (Equation 2).

$$Y_o = \{\emptyset\} \quad (1)$$

$$X^+ = \underset{X \notin Y_k}{\operatorname{argmax}}[J(Y_k + X)] \quad (2)$$

Definition 2 (Backward elimination): Starting with the full set (Equation 3), sequentially remove the worst feature X^- that results in the smallest decrease in the value of the accuracy J$(Y_k - X)$ as in Equation 4.

$$Y_o = X \quad (3)$$

$$X^- = \underset{X \notin Y_k}{\operatorname{argmax}}[J(Y_k - X)] \quad (4)$$

Definition 3 (Genetic search):Genetic search is influenced by natural evolution. This genetic search employs a linear combination of accuracy and simplicity terms as the fitness function.

$$Fitness(X) = \frac{3}{4}A + \frac{1}{4} = \left(1 - \frac{S+F}{2}\right) \quad (5)$$

Where X is a feature subset, A is the average cross-validation accuracy of the classifier, S is the number of instances or training samples, and F is the number of subset features.

Definition 4 (Rule evaluation): If there are many feature subsets ($F_>$) with equal fitness values, the rule-based engine returns a feature subset (V_i) with fewer features (X_F); otherwise, it returns the feature subset with the greatest fitness value (F_{hi}) to the basic classifier as in (6).

$$R = \begin{cases} V_i, if\ V_i \in F_> \cap X_F \\ V_i, if F_{hi} \cap \emptyset \end{cases} \quad (6)$$

Ontology construction module

The rules for the automatic generation of the ontology-based diabetes prediction were defined using the J48 decision tree model on the selected features. An ontology-based intelligent system for patient diabetes prediction was created using the decision's outcomes. The decision tree technique was used to identify the criteria that distinguish positive diabetic patients from negative diabetic patients. A decision tree can be represented via a recursive split of the instance space. According to a certain discrete function of the values of the input attributes, each internal node in a decision tree divides the instance space into two or more sub-spaces. Equation 7 provides the metric for determining the appropriate split depending on the extent to which examples belong more to one class than the others.

$$Entropy\ (t) = -\sum_{i=0}^{c-1} p(i|t)\ log_2\ p(i|t) \quad (7)$$

where:

c is the number of classes

$p(i|t)$ denotes the fraction of instances belonging to class *i* at a given node *t*

Gain ratio is the splitting criterion used by the J48 decision tree method to assess the quality of a split. The criterion is defined as in Equation 8.

$$Gain\ ratio\ (D, A) = Entropy\ (D) - \sum_{v\ \in\ Values(A)} \frac{|D_v|}{|D|}\ Entropy\ (D_v) \quad (8)$$

where:

$Values(A)$ denotes the set of all possible values for attribute *A*

D_v denotes the subset of dataset *D* having value *v* for attribute *A*

Classification module

The Bayesian classifier, which is based on the Bayesian theorem, is a straightforward and useful classifier that suggests document classification using a probabilistic approach. The categorization module obtains the user-supplied keywords and determines the weights that should be assigned to each testing document. The Term Frequency Inverse Document Frequency (TFIDF) was used to compute the frequency and weights for each testing document.

The Term Frequency (TF) is used to compute the frequency of each keyword in the testing document (Equation 9) and Inverse Document Frequency (IDF) is used to compute the weight of each keyword appearing in the testing document as shown in Equation 10

$$TF = \frac{count}{l} \quad (9)$$

$$IDF = log\left(\frac{n}{k}\right) \quad (10)$$

where l represents the total number of keywords in the document, count is the frequency of a given keyword in the document, n is the number of documents and k is the number of documents the keyword appears in. Thus, the product of TF and IDF denoted as TFIDF is used to compute the weighted term frequency score for a given keyword in a document as shown in Equation 11.

$$TFIDF = \frac{count}{l} * log\left(\frac{n}{k}\right) \quad (11)$$

The testing document is then classified using a Bayesian classifier using these keyword weights. The probability of a given document/symptom X_i belonging to a category or class C_j of diabetes is as shown in Equation 12. In other word, the probability of C_j given input features X_i is given as follow:

$$P(X_i|C_j) = \prod_{i}^{n} P(W_{ki}|C_j) \quad (12)$$

where X_i is a document, i=1 to m, C_j is a category, j=1 to n, and W_{ki} is the weight of keyword W_k in X_i.

The probability that a given document or symptom belongs to class Y = C, can be calculated as the product of probability that each of values of the i document's attributes belong to class C, as shown in Equation 13.

$$P(X|Y = C) = \prod_{i=1}^{n} P(X_i|Y = C) \quad (13)$$

When the dataset contains numerical inputs, the probability that a given value of the attribute belongs to class $y_{i,}$ can be determined using the Gaussian distribution function as in Equation (14).

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i - \mu_i)^2}{2\sigma_y^2}\right) \quad (14)$$

Finally, the probability that a given document or symptom will be classified in class C, can be calculated as in Equation 15.

$$P(C|x_1, \dots, x_n) = \frac{P(C)P(x_1, \dots., x_n|C)}{P(x_1, \dots., x_n)} \quad (15)$$

Document query module

This module's goal is to find and show the desired documents that users have requested using the ontology and the naive Bayes classifier.

Algorithm 1: Multi-step feature selection

Input: Training Dataset X$\{x_1, x_2, \dots., x_k \mid$ xi ∈ C$\}$

Output: X_r //reduced features

Process:

1. Begin
2. $Y \leftarrow X(x_1, x_2, \dots., x_k)$
3. $Y_o = \{\emptyset\}$ //forward selection
4. $X^+ = \underset{X \notin Y_k}{\text{argmax}}[J(Y_k + X)]$
5. $Y_{k+1} = Y_k + X^+, k = k + 1$ //update
6. $Y_o = X$ // Backward elimination
7. $X^- = \underset{X \notin Y_k}{\text{argmax}}[J(Y_k - X)]$
8. $Y_{k+1} = Y_k - X^-; k = k + 1$ //update
9. $P = X^+ + X^-$ // feature pool
10. Begin by randomly generating an initial population P.
11. Define a probability distribution p over the members of P where p(x) / f(x).
12. Select population members x and y with respect to p.
13. Apply crossover to x and y to produce new population members x' and y'.
14. Apply mutation to x' and y'.
15. Insert x' and y' into P′ (the next generation).
16. If |P′| < |P|, goto 13 & 14.
17. Let P ← P′
18. If there are more generations to process, goto 12.
19. Return x ∈ P for which f(x) is highest.
20. If two feature subsets have the same fitness value
21. //return the feature subset with least number of subset features
22. return X_r
23. End

Algorithm 2: BuildTreeClassifie

Input: Attribute_set F{ $f_1,f_2,\dots,f_n$ },
Training Dataset $X_r\{x_1,x_2,\dots,x_n \mid x_i \in C\}$

Output: Decision Tree – root

Process:

1: **IF** stopping_cond(X,F) = true **THEN**
2: leaf = **createNode()**
3: leaf.label = **Classify(X)**
4: return leaf
5: ELSE
6: root = **createNode()**
7: root.test_cond = **find_best_split(X,F)**
8: let V = {v|v is a possible outcome of root.test_cond}
9: **FOR EACH** v ∈ V **DO**
10: X_v = **{**x |root.test_cond(x) = v and x ∈ X**}**
11: child = **buildTreeClassifier(X,F)**
12: add child as descendant of root and label the edge as v
13: **END FOR**
14: **END IF**
15: return root

Algorithm 3: Prunning Algorithm

Input: Attribute_set F{ $f_1,f_2,\dots,f_n$ }, SplitRatio S
Training Dataset $X_r\{x_1,x_2,\dots,x_n\}$

Output: Pruned Decision Tree

Process:

1: GrowingSet G ← set of training data used to build the tree
2: PruningSet P ← set of training data for validating the tree
3: P,G = **splitExamples**(S,X,F) //split training data
4: tree = **buildTreeClassifier**(G,F)
5: **WHILE(true)**
6: prunedTree = **bestSimplification(tree,P)** //pruning process
7: IF (accuracy(prunedTree,P) < accuracy(tree,P))
8: **BREAK;**
9: tree = prunedTree
10: **END WHILE**
11**:** return tree

Algorithm 4: Classification

Input: Testing documents, keyword set KS, and ontology
Output: Categorized training documents C
Process:

1. **For each** (testing document, TD) Do
2. **Retrieve** the keywords T_{x1}, T_{x2}, T_{x3}, …, T_{xm} from the TD
3. **Calculate** the frequency of each keyword T_{xm} that appeared in the TD, where T_{xm} ∈ KS and $1 \leq i \leq$ m.
4. TFIDF (T_k, Di) = T_kF *ID_iF// Use TFIDF formula to calculate the weights W_{x1}, W_{x2}, W_{x3}, …, W_{xm} for keywords T_{x1}, T_{x2}, T_{x3}, …, T_{xm}.
5. // Use Bayesian classifier to classify the TD according to the weights W_{x1}, W_{x2}, W_{x3}, …, W_{xm}
6. $P(X|Y = C) = \prod_{i=1}^{n} P(X_i|Y = C)$
7. $P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} exp\left(-\frac{(x_i-\mu_i)^2}{2\sigma_y^2}\right)$
8. $P(C|x_1, \dots, x_n) = \frac{P(C)P(x_1,\dots,x_n|C)}{P(x_1,\dots,x_n)}$
9. **return** C
10. **End**

Algorithm 5: Generating_Ontology
Input: Decision tree rules, all training documents, and a threshold value α.
Output: A domain ontology.
Initialization: Let KS and TD be the set of keywords and training documents respectively.

Let count_k_i = 0 denotes the number of training documents with k_i in them.

Let val(k_i) = count_ k_i / |TD|.

1. **Begin**
2. **Retrieve** keywords for the training documents by Decision tree rules and get a keyword set, KS. In addition, the value of count_k_i is obtained for each k_i in KS.
3. **Delete** each keyword k_j from KS when val(k_i) < α.
4. **Use** Decision tree rules to generate a binary relation between the attributes in KS
5. **Compute** the hierarchy relationship among concepts (keywords) in KS; that is, the categories, attributes, association rules, and concept of extension for keywords are generated and formed a concept relationship diagram.
6. **For each** node (concept) N_i in the concept relationship diagram **Do**
7. **Let** N_i_PS be the set of parent nodes of N_i, where N_i is a node in the concept relationship diagram.
8. **If** | N_i_PS | > 1, **Then**
9. **Select** any node Np to be the parent of N_i, where count_kp = max{count_k_j | $N_j \in N_i$_PS}.
10. **Delete** edge (N_i, N_m), where $N_m \in N_i$_PS – {Np}.
11. **Return** domain ontology
12. **End**

Algorithm 6: Query
Input: A series of keywords inputted by a user.
Output: The suitable documents D.

1. **querySet** :={q_1(X),..., q_i(X)}
2. **For each** (keyword $q_i \in$ querySet inputted by the user) Do
3. **While** (TD ≠ ∅) do
4. **s**=**Search** (documents belonging to the category q_i from the domain ontology, TD).
5. **If(**$q_i \equiv$ s) Do
6. **D=**getDocument(TD)
7. **If(**querySet≠ null**) goto step 4**
8. **Return** D
9. **End**

RESULTS

The method of the Naive Bayes classifier-based classification module is demonstrated in the example below. Let's assume that there are 768 training documents, two categories C1 and C2 belonging to documents 18 and 19, and a testing document TD with three keywords Tk1, Tk2, and Tk3 exist in the ontology. The TD has the keywords Tk1, Tk2, and Tk3 six times, three times, and nine times, respectively. The training papers contain the keywords Tk1, Tk2, and Tk3 sixteen times and 26 times, respectively. The weights of Wk1, Wk2 and Wk3 are 6 * log (768 / 26) =8.82, 3 * log (768 / 9) =5.79, and 9 * log (768 / 16) =15.13 by TFIDF formula. P(Tk1| C1) = 8.82/ 18 =0.49, P(Tk2| C1) = 5.79/ 18 =0.32, P(Tk3| C1) = 15.13 / 18 =0.84 , P(Tk1| C2) = 8.82/ 19 =0.46 , P(Tk2| C2) = 5.79 / 19 =0.30 , and P(Tk3| C2) = 15.13 / 19 =0.79 . Finally, P(TD|C1) = 0.49* 0.32* 0.84 =0.13, P(TD| C2) = 0.46*0.30*0.78 =0.10 . Since 0.13 > 0.10, TD document is classified to C1.

True positive rate (TP): The number of cases that were appropriately categorized in the typical class is shown here. This is denoted in (16).

$$True\ positive\ rate = \frac{TP}{TP + FN} \quad (16)$$

False positive rate (FP): This is the number of cases that were misclassified as belonging to the usual class. It is denoted in (17).

$$False\ positive\ rate = \frac{FP}{FP + TN} \quad (17)$$

Precision: is an indicator of how accurately a certain class that was anticipated to be positive turned out to be positive. It is denoted in (18).

$$Precision = \frac{TP}{TP + FP} \quad (18)$$

Recall: is a measure of the number of labelled instances that are correctly detected by a prediction model as depicted in (19).

$$Recall = \frac{TP}{TP + FN} \quad (19)$$

F-measure: is the precision and recall depending on a specific threshold's harmonic mean. It is employed to evaluate the classification's quality as shown in (20).

$$F - measure = \frac{2(precision * recall)}{precision + recall} \quad (20)$$

Accuracy: is the percentage of correctly classified instances over the total number of instances. It is denoted in (21).

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \quad (21)$$

ROC area: It is a metric used to check the quality of classification algorithms. The AUC is a plot of the true positive ratio and the false positive ratio.

DISCUSSION

This section compares the effectiveness of the built ontology-based diabetes classification model using naive Bayes classifier and provides thorough explanations of the experimental findings. K-fold (k = 10) cross-validation was employed to address the issue of imbalanced data and overfitting in classification and to reinforce the simulation findings. Ten subsets of the dataset with similar sizes were randomly selected from the training data. The built ontology-based diabetes classification model was tested using one subset and a naive Bayes classifier, with the remaining nine subsets serving as training data. Table 6 shows the accuracy and error analysis between the developed ontology-based diabetes classification model using naïve Bayes classifier and other related methods. Different methods were implemented on the same dataset and computational platform. The results of the correctly classified (accuracy) and incorrectly classified (error) were recorded. The record showed that the developed approach achieved better accuracy and error rate of 719 (93.55%) and 49 (6.46%) respectively compared to other related methods. This results showed that the developed approach was able to correctly predict diabetes in patient at high accuracy rate. The reason for the better performance of the developed approach can be attributed to the automatic generation of rules for the ontology construction and the ability of the ontology to provide straightforward interpretable decision rules through naïve Bayes model. The results also showed that Hidden Markov Model (HMM)

achieved the lowest results with accuracy and error rate of 268 (34.90%) and 500 (65.10%) respectively.

The results showed that most of the methods on the testing set obtained F-measure of at least 51.7% classification rates. The F-measure of the developed approach is better with 87% compared to other related methods. Also, in the developed method, a large number of the instances were correctly classified as either tested negative or tested positive, while only a small amount of the instances was misclassified. This can be explained by the large score differences between the true positive and the false positive rates with 0.792 and 0.019 respectively. The true positive of the developed approach is slightly higher than most of the other methods, plus the false positive being lower than most of the other methods. The precision of the developed approach is higher than that of the closest result of the decision tree with 96.5% > 84.8%, plus the recall of the developed approach is slightly lower than that of the multilayer perceptron with 79.2% <81.3%. The results showed that the developed approach is best for diabetes classification. Overall, the results showed relatively imbalanced performances among the different methods with SVM producing the worst results for diabetes prediction.

Figure 2 shows the accuracy and error analysis. The accuracy of the developed approach is higher than that of the closest result of the decision tree with 719(93.55%) >646(84.12 %), plus the error of the developed approach is lower than that of the closest result of decision tree with 49(6.46%) <122 (15.89 %). The inferences from the comparison showed that the developed approach outperformed the other related methods.

Figure 3 shows the comparative analysis of the different methods in terms of the confusion matrix. The true positive, false positive, and recall of the HMM with 1.000, 1.000, and 1.000 respectively exhibited overfitting problem despite the application of the 10-fold cross validation. The HMM also produces the worst results with precision, F-measure, and ROC area of 0.349, 0.517, and 0.500 respectively, when compared with the other methods. On the other hand, the developed approach produces the best overall results with TP rate, FP rate, Precision, Recall, F-measure, and ROC area of 0.792, 0.019, 0.965, 0.792, 0.870, and 0.958 respectively. These results of the developed approach showed its ability to correctly classify diabetes classes in patients compared to the other methods. The developed approach did not exhibit overfitting and imbalance problems due to the application of the 10-fold cross validation. Figure 13 displays the Area Under Curve (AUC) for the various diabetes classification algorithms. AUC is a statistic that can be used to evaluate the effectiveness of detection algorithms. The positive ratio and the false positive ratio are plotted in the AUC. The AUC gauges a method's overall capacity to distinguish between different types of diabetes. In comparison to previous relevant methods, the created strategy reveals a more accurate test with an area of 0.9578 (i.e., its AUC is close to the upper left corner of the plot). This results showed that the benefit of the developed approach outweighs the cost for diabetes prediction. Therefore, the developed approach is a valid method for diabetes prediction.

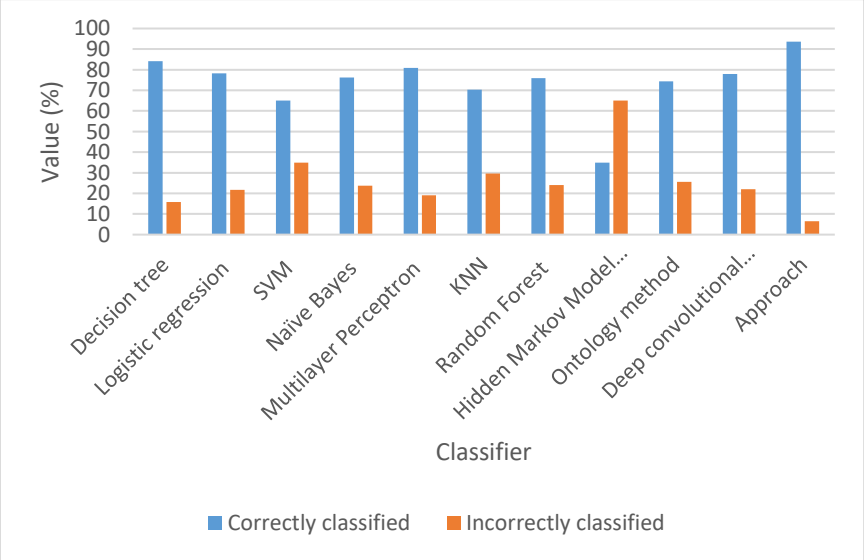

Figure 2: Accuracy and error analysis

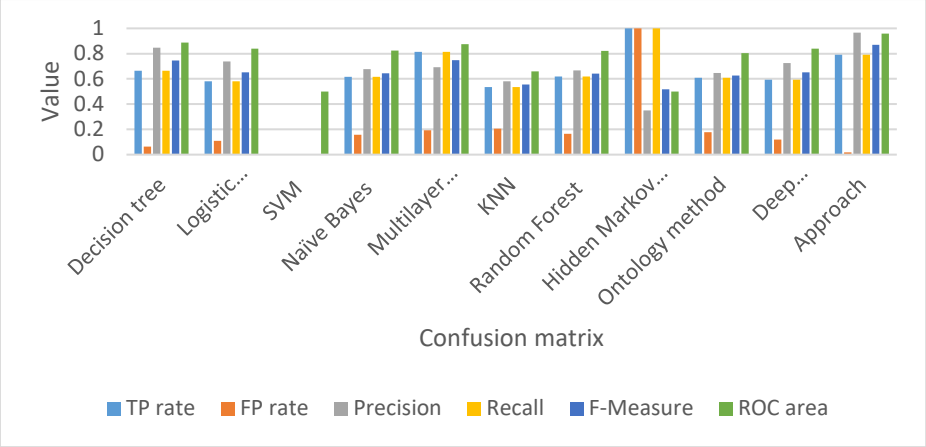

Figure 3: Comparative analysis

CONCLUSION

Using a naive Bayes classifier, the research developed an ontology-based diabetes categorization model. By creating an ontology model based on a decision tree algorithm and a Nave Bayes model, the suggested approach was achieved. The predictive rules for diabetes were automatically generated using a decision tree and then mapped into an ontology using SWRL rules. A Java program was used to extract the SWRL rules from the decision tree model. A testing document was categorized using a Bayesian classifier using some input keywords. The query module algorithm of the study takes as input series of keywords representing diabetes symptoms from the users. For each of the query keyword from the user, the algorithm searches the testing documents to match documents belonging to the keyword category from the domain ontology. If the keyword is similar to a particular attribute from the testing document of the domain ontology, the algorithm gets the document and use the naïve Bayes classification algorithm to automatically generate the output of the user's query. The experimental results of the developed approach for diabetes prediction was tested and validated to be suitable for diabetes classification. The results also showed that the developed approach achieved better classification of diabetes with F-measure of 87% compared to other related methods. Similarly, the developed approach indicates a more perfect AUC test compared to the other related methods. These results inferred that the benefit of the developed approach outweighs the cost for diabetes prediction. This study recommends the adoption of the research method for early detection and prediction of diabetes. The availability of more and robust diabetes datasets should be made available publicly to alleviate the problem of scarce datasets for the evaluation of solutions developed for diabetes predictions. Finally, the medical histories of diabetes patients have produced a vast amount of data, and intelligent techniques can be used to extract this useful information for diabetes prediction.

REFERENCES

Ahlqvist, E., Storm, P., Käräjämäki, A., Martinell, M., Dorkhan, M., Carlsson, A., Vikman, P., Prasad, R. B., Aly, D. M., Almgren, P., Wessman, Y., Shaat, N., Spégel, P., Mulder, H., Lindholm, E., Melander, O., Hansson, O., Malmqvist, U., Lernmark, Å., … Groop, L. (2018). Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The Lancet Diabetes and Endocrinology*. https://doi.org/10.1016/S2213-8587(18)30051-2

Alex, S. A., Nayahi, J. J. V., Shine, H., & Gopirekha, V. (2022). Deep convolutional neural network for diabetes mellitus prediction. *Neural Computing and Applications*. https://doi.org/10.1007/s00521-021-06431-7

Bhutta, Z. A., Salam, R. A., Gomber, A., Lewis-Watts, L., Narang, T., Mbanya, J. C., & Alleyne, G. (2021). A century past the discovery of insulin: global progress and challenges for type 1 diabetes among children and adolescents in low-income and middle-income countries. In *The Lancet*. https://doi.org/10.1016/S0140-6736(21)02247-9

Dremin, V., Marcinkevics, Z., Zherebtsov, E., Popov, A., Grabovskis, A., Kronberga, H., Geldnere, K., Doronin, A., Meglinski, I., & Bykov, A. (2021). Skin Complications of Diabetes Mellitus Revealed by Polarized Hyperspectral Imaging and Machine Learning. *IEEE Transactions on Medical Imaging*. https://doi.org/10.1109/TMI.2021.3049591

El Massari, H., Mhammedi, S., Sabouri, Z., & Gherabi, N. (2022). Ontology-Based Machine Learning to Predict Diabetes Patients. *Lecture Notes in Networks and Systems*. https://doi.org/10.1007/978-3-030-91738-8_40

Hatua, A., Subudhi, B. N., Veerakumar, T., & Ghosh, A. (2021). Early detection of diabetic retinopathy from big data in hadoop framework. *Displays*. https://doi.org/10.1016/j.displa.2021.102061

Kiv, S., Heng, S., Wautelet, Y., Poelmans, S., & Kolp, M. (2022). Using an ontology for systematic practice adoption in agile methods: Expert system and practitioners-based validation. *Expert Systems with Applications*. https://doi.org/10.1016/j.eswa.2022.116520

Komi, M., Li, J., Zhai, Y., & Xianguo, Z. (2017). Application of data mining methods in diabetes prediction. *2017 2nd International Conference on Image, Vision and Computing, ICIVC 2017*. https://doi.org/10.1109/ICIVC.2017.7984706

Krishnamoorthi, R., Joshi, S., Almarzouki, H. Z., Shukla, P. K., Rizwan, A., Kalpana, C., & Tiwari, B. (2022). A Novel Diabetes Healthcare Disease Prediction Framework Using Machine Learning Techniques. *Journal of Healthcare Engineering*. https://doi.org/10.1155/2022/1684017

Kumar, K. G. N., & Christopher, T. (2016). Analysis of liver and diabetes datasets by using unsupervised two-phase neural

network techniques. *Biomedical Research (India)*.

Kushwaha, J. S., Gupta, V. K., Singh, A., & Giri, R. (2022). Significant correlation between taste dysfunction and HbA1C level and blood sugar fasting level in type 2 diabetes mellitus patients in at a tertiary care centre in north India. *Diabetes Epidemiology and Management*, *100092*.

Mandal, N., Grambergs, R., Mondal, K., Basu, S. K., Tahia, F., & Dagogo-Jack, S. (2021). Role of ceramides in the pathogenesis of diabetes mellitus and its complications. In *Journal of Diabetes and its Complications*. https://doi.org/10.1016/j.jdiacomp.2020.107734

Ogurtsova, K., Guariguata, L., Barengo, N. C., Ruiz, P. L. D., Sacre, J. W., Karuranga, S., Sun, H., Boyko, E. J., & Magliano, D. J. (2022). IDF diabetes Atlas: Global estimates of undiagnosed diabetes in adults for 2021. *Diabetes Research and Clinical Practice*. https://doi.org/10.1016/j.diabres.2021.109118

Oza, A., & Bokhare, A. (2022). Diabetes Prediction Using Logistic Regression and K-Nearest Neighbor. *In Congress on Intelligent Systems*, 407–418.

Parveen, S., Patre, P., & Minj, J. (2023). Various Diabetes Detection Techniques a Survey. *Information and Communication Technology for Competitive Strategies (ICTCS 2021)*, 261–269.

PIMA Indian Diabetes Database. (n.d.). https://github.com/npradaschnor/Pima-Indians-Diabetes-Dataset/blob/master/diabetes.csv

Pranata, R., Henrina, J., Raffaello, W. M., Lawrensia, S., & Huang, I. (2021). Diabetes and COVID-19: The past, the present, and the future. In *Metabolism: Clinical and Experimental*. https://doi.org/10.1016/j.metabol.2021.154814

Ranjitha, R., Agalya, V., & Archana, K. (2022). Diabetes Prediction by Artificial Neural Network. *Lecture Notes in Networks and Systems*. https://doi.org/10.1007/978-981-16-5529-6_76

Thakkar, H., Shah, V., Yagnik, H., & Shah, M. (2021). Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis. *Clinical EHealth*. https://doi.org/10.1016/j.ceh.2020.11.001

Vijayan, V. V., & Anjali, C. (2016). Prediction and diagnosis of diabetes mellitus - A machine learning approach. *2015 IEEE Recent Advances in Intelligent Computational Systems, RAICS 2015*. https://doi.org/10.1109/RAICS.2015.7488400

Yun, W., Zhang, X., Li, Z., Liu, H., & Han, M. (2021). Knowledge modeling: A survey of processes and techniques. *International Journal of Intelligent Systems*. https://doi.org/10.1002/int.22357