A FRAMEWORK FOR ONTOLOGY-BASED DIABETES DIAGNOSIS USING BAYESIAN OPTIMIZATION TECHNIQUE

F.M. OKIKIOLA, O. S. ADEWALE, A.M. MUSTAPHA, A. M. IKOTUN AND O.L. LAWAL

1 Department of Computer Technology, Yaba College of Technology, Lagos.
2 Department of Computer Science, Federal University of Technology, Akure, Ondo State.
3 Department of Computer Science, Federal University of Agriculture, Abeokuta, Nigeria

*Corresponding Author: sade.mercy@yahoo.com Tel: +234

ABSTRACT
Diabetes Management System (DMS) is a computer-based system which aid physicians in properly diagnosing diabetes mellitus disease in patients. The DMS is essential in making individuals who have diabetes aware of their state and type. Existing approaches employed have not been efficient in considering all the diabetes type as well as making full prescription to diabetes patients. In this paper, a framework for an improved Ontology-based Diabetes Management System with a Bayesian optimization technique is presented. This helped in managing the diagnosis of diabetes and the prescription of treatment and drug to patients using the ontology knowledge management. The framework was implemented using Java programming language on Netbeans IDE, Protégé 4.2 and mysql. An extract of the ontology graph and acyclic probability graph was shown. The result showed that the nature of Bayesian network which has to do with statistical calculations based on equations, functions and sample frequencies led to more precise and reliable outcome.

Keywords: Diabetes, Ontology, Bayesian Classifiers, Knowledge Management, Diagnosis Management.

INTRODUCTION
One of the most chronic and serious disease in the present generation is diabetes (El-Sappagh1 and Elmogy, 2016). It is gotten from the American Diabetes Association (ADA), it significantly creates so much economic burden on every country in the world today. It has taken so much of health budget of nations (International Diabetes Federation). It is projected that in decades of years from now, this amount spent would exceed 500 billion of dollars. Over 300 million people all over the world are estimated to be affected with diabetes which could double in 12 years from now (Roche and Wang, 2014). World Health Organization projects that the seventh leading disease that will lead to mortality in decade is diabetes (http://www.who.int/en/). There is a very high rate of increasing number around the world having this disease (Hemopo et al. 2015). About 80% from underdeveloped and developing countries. Many more have this disease without knowing because of the minimal physical effect it is currently having on them. Diabetes is generally of three types

This approach has to do with single or multiple related objects comprising of basic vocabulary items and relationships in a particular area. Employing ontological approach has immense advantages in the medical domain such as theoretical and practical development of cost-effective automation (Liaw et al., 2013), ability to help test and refine relationship existing between disease and supporting clinical findings (Haug et al., 2013) and the ability to share and reuse knowledge easily which will be helpful in this research work.

Although, a great number of health-focused organizations that publish documents relating to detecting and self-care recommendation methods for diabetes patient through ontology or semantic web rule, yet the following problems exist; a number of past research works have been centered on the recommending food, exercise activities which makes the control of the disease really slow (Hempo et al., 2015; Ahmed et al., 2014). The manual power currently employed results in differences in care quality and consumption of lot of man power (Chen et al., 2010). More of research work have been centered on diabetes type II and type I exempting the genital diabetes (Vasant et al., 2015; Zhang et al., 2015; Chalortham et al., 2009). In this paper, a conceptual framework which employs ontology for the management of diabetes diagnosis that will aid the reusability of knowledge in the diagnosis and prescription phase is presented. This will also be considering the three-major diabetes type in it diagnosis approach.

The remaining part of the paper is structured into four sections: the review of researches relating to the research work is done in Section II; the explicit explanation framework and methodology were presented in Section

Knowledge can be divided into two forms: Semi-structured and non-structured knowledge. Promoting the acquisition of knowledge, creation of knowledge and knowledge sharing is needful in the management of diabetes. As regards this, using an ontology approach can fulfill a vital role in knowledge management. This approach has to do with single or multiple related objects comprising of basic vocabulary items and relationships in a particular area. Employing ontological approach has immense advantages in the medical domain such as theoretical and practical development of cost-effective automation (Liaw et al., 2013), ability to help test and refine relationship existing between disease and supporting clinical findings (Haug et al., 2013) and the ability to share and reuse knowledge easily which will be helpful in this research work.
III. Section IV shows the implementation, states the results and discusses the outcomes of the research. Section V gives a conclusion and gives the possibility for future researches that could be conducted.

RELATED WORK

A number of research works related to this work exist in the management of diabetes diagnosis in the provision of medical data privacy, diagnosis, treatment and prevention in literature.

A progressive and full fuzzy and ontology centered case based reasoning framework for managing and utilizing inaccurate knowledge was proposed by El Sappagh et al. (2014). This framework was implemented for addressing the problem of diabetes diagnosis considering the situation of 60 real cases from the electronic health record of the Mansoura University Hospitals, Mansoura, Egypt. The accuracy of the system was gotten to be 97.67% which was really high to show the performance level of the system. Vasant et al., (2015) presented a method of generating ontology through the reviewing and text mining of expert and the resultant ontology of disease-phenotype association through the use of terminologies from mouse and human phenotype ontologies. They considered only type 2 diabetes (T2D). Through their process, a rapid and pragmatic method for information extraction was presented. The text mining approach was considered a very beneficial tool to create ontologies. This proposition had some level of success in its performance with disease-phenotype association in different disease areas. The modifying of input vocabularies for improving recalls was not a consideration in the method. It was discovered form the research that type-2 diabetes phenotype did not have a clear representation in the implementation. Meysam et al. (2016) developed a model that considered the use of neural network for improving how diabetes is being predicted through the use of data relating to clinical and lifestyle characteristics. Memetic algorithms were employed for updating weights and improving the model’s prediction accuracy. The evaluation of the model gave an output with sensitivity (96.2), specificity (95.3), positive predictive value(93.8), negative predictive value (92.4), and ROC (0.958). This gave a foundation to the design of a decision support system to manage and plan the risk of having diabetes as individuals.

Nasib and Pooja (2016) proposed a novel hybrid model for diabetic prediction by using data mining techniques. The main objective of this study is to improve the accuracy rate by significantly reducing the size of the data under analysis at every stage. The PIMA Female Diabetic dataset, extracted from UCI repository were used using HMM and Fuzzy Improved Neural Network. The proposed hybrid model achieved 92% of overall accuracy. The Fuzzy Improved Neural Networks were used for predicting the diabetes disease over the data. The result analysis proved that the prediction accuracy is poor (Naïve Bayes: 76.30%, Neural Networks: 75.13, Support Vector Machine: 77.47, K-Nearest neighbor: 69.79, Decision Tree (J48): 74.21), when the classifiers are implemented separately but when these are amalgamated with each other, produces better results. The proposed model can be applied on gender independent dataset. Further, the accuracy rate of the model can be improved by replacing the missing values of the dataset with the most appropriate value. Thiyagarajan (2016) proposed an effective machine learning algorithm for the classification of type DM patients. This machine learning algorithm used
for classification will find the optimal hyperplane which divides the various classes. By using this machine learning algorithm, the classification accuracy is achieved for classifying the diabetes patients.

Fredrik (2012) investigated simpler linear and combined models of glucose dynamics for short-period prediction, which was developed between the European Union FP7 DIAvvisor project. The models considered are useful in a decision support system, for prompting users about future low and high glucose levels as well as recommend actions that should be taken during their implementation in a control framework. The evaluation of these models was done using 47 patient data records from the first DIAvvisor trial. In Mukesh et al., (2015), a classifier based on Bayesian theorem was proposed for the prediction of a likely diabetes occurrence in a person. The authors made use of a dataset gotten from a hospital, inclusive of both individuals with or without the diseases, after which they were classified. The authors intend with the study to diagnose diabetes by discovering using a decision tree model. The authors found out that a proper prediction model required the gathering of more data for increasing its accuracy. This would be possible through the collection of diabetes datasets from different sources, with the possibility to generate a model from each of the dataset. Mythili (2012) considered a set of sixteen factors and identified the most dominant through the use of regression analysis to diagnose diabetes for achieving better accuracy. The parameters relating to these factors were gotten from the most recognized parameters often employed to predict diabetes through the application of two sets of mathematical models i.e. linear regression and nonlinear regression for the datasets that were given. The level of accuracy for predicting early diabetes disease was improved through the research, hence, helping the patients take timely precaution.

**RESEARCH METHODOLOGY**

This research work proposes a method that will be using ontology as its knowledge base in the diagnosis and prescription of medication for the diabetes mellitus. This system will be addressing this management in view of the three major diabetes types. From Figure 1 the following modules are present which will explain the workability of the system: Patients Information/ Symptoms Entry, Diagnosis module, Diabetes Care Ontology Module, and Knowledge Engineering.

**Patients Information/ Symptoms Entry**

This phase is a graphical user interaction platform that enables the clinical agent to input the information about the patients as well as check symptoms that will be used to diagnose the diabetes type of each patient. Here, the patient’s details such as the name, age category, blood group, date and time of consultation will be inputted to keep record of each patient. This will be built with the use of Java Programming language. An access control platform will also be inculcated into the system to restrict the access of individuals that can utilize the system. This is necessary to ensure that unqualified physician is not granted access to the system. The information of the patient as well as the physician utilizing the system at that time will be sent to the patient database as seen in Figure 1.
Algorithm of the Access Control of the System

INPUT: Username U(i), Password P(i)
OUTPUT: Access Right (Allow Deny)

PROCESS:
1: Start
2: Input U(i)
3: Input P(i)
4: if U(i) is present in patient database
5: if U(i) matches P(i)
6: Return Allow
7: endif
8: else
9: Return Deny
10: endif
11: end if
12: else
13: Return Deny
14: end else
15: End process

Algorithm of the Diagnosis Module Operation

INPUT: Symptoms Indicator, A(i) (Yes, No)
OUTPUT: Symptoms deduced S(i)
PROCESS:
1: Start
2: set yes_count = 0
3: for i = 1 to N of all likely symptoms
4: if (S(i) exist)
5: A(i) = Yes
6: yes_count = yes_count + 1
7: endif
8: else
9: A(i) = No
10: end else
11: end for
12: for j = 1 to yes_count
13: Return S(j)
14: End for
15: End process

Diagnosis Module
This module collects the diagnosis information required (symptoms), and send it to the ontology module to check for the diabetes type that matches them. This will enable the system to return the diabetes type deduced and the connecting prescription/recommendation for each patient. This module will enable the efficient diagnosis for the three-major diabetes type which one or two have not been addressed in existing research works.
Algorithm for Ontology Storage Operation

INPUT: Symptoms (S(i)), Ontology Class (O(j))
OUTPUT: S, Storage

PROCESS:
1: Start
2: for i = 1 to n of all symptoms
3: check S(i) necessary
4: input O(i)
5: if O(i) exist
6: map S(i) to O(i)
7: Input prescription to prescription class
8: End if
9: else
10: return exemption message
11: End else
12: Store S(i) to O(i) to ontology DB
13: End for
14: End process

Figure 1: Structure for the Proposed Framework

Diabetes Care Ontology Module
The diabetes care ontology module is the platform that enables user to store into and query information from the ontology database. The storage of these attributes into classes (type1_symptoms, typ2_symptoms, genital_symptoms) shows a representation of how the diagnosis class is in the database. The prescription class shows what drug and activities are recommended for each diabetes
which could be needed for reference.

**Bayesian Optimization Technique: Mathematical Method**

Due to the need of providing an optimized diagnosis result, an optimization technique is required. The technique considered in this research is the Bayesian classification which is based on the Baye's Theorem. Bayesian classifiers are based on statistical theorems which has the ability of predicting class membership probabilities, for example, the probability of a particular tuple belonging to a class.

**Knowledge Engineering**

This module of the work will be addressing the reasoning and rule aspect of the system. This will help avoid going through a fresh process of re-diagnosing diseases by tracing again through the connection of each classes. This will keep record of previous diagnosis, treatment and recommendation type. This module also allows for the update of the new diagnosis and prescription information into the ontology library. An extract of the ontology graph of the framework is shown in Figure 2.

**Algorithm for Ontology Query Operation**

**INPUT:** Symptoms \( S(i) \)

**OUTPUT:** DBT [Type1, Type2, Genital], Prescription \( P \), Recommendation \( R \)

**PROCESS:**
1. Start
2. for \( i = 1 \) to \( n \) of all symptoms
3. check \( S(i) \) in ontology DB
4. if \( S(i) \) present in diagnosis class
5. check all diagnosis class where \( S(i) \) is present
6. Repeat step 3 to 5
7. for \( j = 2 \) to \( n \) of all symptoms
8. check \( S(j) \) present in all diagnosis class
9. if \( S(i) \) and \( S(j) \) are present
10. Return DBT
11. End if
12. End for
13. Return \( P, R \)
14. End if
15. End for
16. End process

A FRAMEWORK FOR ONTOLOGY-BASED DIABETES DIAGNOSIS...
Directed Acyclic Graph

The different nodes in directed acyclic graph bear a representation of a random variable. These variables could be either discrete or continuous. They likely have the ability of corresponding to the real attribute given. A directed acyclic graph for six Boolean variables of my own perception is shown in Figure 3. The arc in the diagram allows representation of causal knowledge. For example, diabetes mellitus can be caused as a result of a person’s family background of diabetes mellitus, also, if that individual is not addicted to sugar. It is worth noting that the variable Positive XRay does not depend on the

Figure 2: Extract of the proposed framework ontology graph

Bayes’s Theorem

Baye’s Theorem is of two probability types which are posterior probability \( P(H/X) \), prior probability \( P(H) \) where, \( X \) is data tuple and \( H \) is some hypothesis.

According to Baye’s Theorem,

\[
P(H/X) = \frac{P(X/H)P(H)}{P(X)}
\]  

(1)

Bayesian Network

Bayesian Network specifies the collective conditional probability distributions. Bayesian Networks and Probabilistic Network are known as belief network. It gives permission for class conditional independencies to be defined between subsets of variables. It gives the provision of a visual model of causal relationship with which it can be possible to learn. The Bayesian network that is trained is made use of in classifying also. This definition is done through the use of two major components: a directed acyclic graph and a set of conditional probability tables

Directed Acyclic Graph

The different nodes in directed acyclic graph bear a representation of a random variable. These variables could be either discrete or continuous. They likely have the ability of corresponding to the real attribute given. A directed acyclic graph for six Boolean variables of my own perception is shown in Figure 3. The arc in the diagram allows representation of causal knowledge. For example, diabetes mellitus can be caused as a result of a person’s family background of diabetes mellitus, also, if that individual is not addicted to sugar. It is worth noting that the variable Positive XRay does not depend on the
Family History

Sugar

Diabetes Mellitus

ketoacidosis

hyperglycemia

Positive Xray

**Figure 3: Directed acyclic graph for six Boolean variables**

**Set of Conditional probability table representation:**

In Table 1, the conditional probability table for Diabetes Mellitus (DM) variable values is shown with each likely combination of its parent nodes, Family History (FH) and Sugar (S).

<table>
<thead>
<tr>
<th>FH,S</th>
<th>FH,-S</th>
<th>-FH,S</th>
<th>-FH,-S</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM</td>
<td>0.9</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>-DM</td>
<td>0.25</td>
<td>0.51</td>
<td>0.15</td>
</tr>
</tbody>
</table>

**IMPLEMENTATION AND RESULTS**

The Bayesian networks are good displays for the control of uncertainty. A Bayesian network is a directed no circle graph the vertexes of which are accidental variables and each vertex is a conditional distribution based on parent. According to Fig 2, after training the patient dataset, marginal probabilities of symptoms P(si) and disease P(dj) and conditional probabilities of symptoms on all diseases P(si/dj) are calculated by counting frequencies in the data. Given a set
of symptoms \( S = \{ s_i \} \) for a patient, the posterior probability for each diagnosis is calculated as:

\[
pr(e|h) = \frac{pr(e|h)pr(h)}{pr(e|h)pr(h) + pr(e|\neg h)pr(\neg h)}
\]  

(2)

Using Formula above, we can think of an idea for designing the structure of Bayesian network. This formula determines the meaning of a Bayesian network. This value is then divided by the diagnosis and multiplied by 100 to determine the percentage error. Where \( r \) is the total number of disease \( h \) is the disease diagnosis and \( P(e) \) is the prediction. Bayesian Network Structure A Bayesian structure can be made based on diabetes data. Pregnancy, age, DPF (Diabetes Pedigree Function) can be some of the effective factors on the appearance of diabetes. A considerable part of data set is related to two measures of obesity: SKIN (triceps skin fold thickness) and BMI (Body Mass Index) which we assume as a hidden variable in the network. Regarding the fact that skin fold thickness is not a good evidence of diabetes, BMI is considered as obesity value. Both GTT and insulin measurements are used for testing diabetes and cause diabetes. Whether blood pressure is a reason for diabetes or not is a question. Following the experiments, it has been found that blood pressure is not a cause of diabetes. Pregnancy, age and obesity are all reasons for blood pressure. According to the presented analysis, Figure 4 indicates the diagnosis of 15 patients and their diabetes status based on the features stated.

<table>
<thead>
<tr>
<th>id</th>
<th>no_of_pregnancy</th>
<th>glucose_conc</th>
<th>blood_pressure</th>
<th>skin</th>
<th>insulin</th>
<th>body_mass_index</th>
<th>diabetes_pedigree</th>
<th>age_of_patient</th>
<th>diabetes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>76.9</td>
<td>120/178</td>
<td>76.6</td>
<td>45.4</td>
<td>54.1</td>
<td>78</td>
<td>79</td>
<td>Y</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>76</td>
<td>140/125</td>
<td>78.8</td>
<td>45.1</td>
<td>54.3</td>
<td>72</td>
<td>79</td>
<td>Y</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>107.9</td>
<td>120/178</td>
<td>56.9</td>
<td>101.8</td>
<td>75.9</td>
<td>89.3</td>
<td>79</td>
<td>Y</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
<td>3.5</td>
<td>140/125</td>
<td>2.3</td>
<td>122.1</td>
<td>19</td>
<td>8</td>
<td>73</td>
<td>Y</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>4.5</td>
<td>120/125</td>
<td>1.3</td>
<td>104.1</td>
<td>19</td>
<td>9</td>
<td>79</td>
<td>Y</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>3.9</td>
<td>120/178</td>
<td>2.5</td>
<td>147</td>
<td>77</td>
<td>7</td>
<td>33</td>
<td>N</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>4.8</td>
<td>120/125</td>
<td>3</td>
<td>136</td>
<td>25</td>
<td>11</td>
<td>27</td>
<td>Y</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>5</td>
<td>140/125</td>
<td>1.7</td>
<td>124.1</td>
<td>25</td>
<td>13</td>
<td>23</td>
<td>Y</td>
</tr>
<tr>
<td>9</td>
<td>4</td>
<td>160/178</td>
<td>2.3</td>
<td>117</td>
<td>33</td>
<td>12</td>
<td>28</td>
<td>28</td>
<td>Y</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>2.7</td>
<td>117/150</td>
<td>2.4</td>
<td>133</td>
<td>33</td>
<td>12</td>
<td>28</td>
<td>Y</td>
</tr>
<tr>
<td>11</td>
<td>4</td>
<td>4</td>
<td>120/125</td>
<td>3.1</td>
<td>117</td>
<td>29</td>
<td>7</td>
<td>24</td>
<td>N</td>
</tr>
<tr>
<td>12</td>
<td>2</td>
<td>3.1</td>
<td>130/150</td>
<td>3.3</td>
<td>126</td>
<td>25</td>
<td>8</td>
<td>34</td>
<td>N</td>
</tr>
<tr>
<td>13</td>
<td>5</td>
<td>2.7</td>
<td>150/125</td>
<td>3.1</td>
<td>140</td>
<td>27</td>
<td>11</td>
<td>30</td>
<td>Y</td>
</tr>
<tr>
<td>14</td>
<td>4</td>
<td>76.9</td>
<td>120/178</td>
<td>76.9</td>
<td>102.7</td>
<td>100.8</td>
<td>70</td>
<td>30</td>
<td>Y</td>
</tr>
<tr>
<td>15</td>
<td>9</td>
<td>76</td>
<td>140/125</td>
<td>76.8</td>
<td>45.3</td>
<td>54.3</td>
<td>33.1</td>
<td>48</td>
<td>Y</td>
</tr>
</tbody>
</table>

Figure 3: Screenshot of the diagnosis of assumed diabetes patients
It can be seen that there are some who though they are diabetic but really are not. The features in the table aided the appropriate Bayesian probability computation of the patients’ state. Figure 4 and 5 show the diagnosis interface for detecting the diabetes type a patient has based on the symptoms felt. Both the symptoms and Bayesian probability computation are employed in knowing what diabetes type a patient has and prescriptions to be given.
It can be seen from Figure 5 that beyond detecting that diabetes exist in a patient, the diabetes type is also diagnose. This as well gives an outcome of recommendation and drug prescription for treating the disease. This makes this approach of better use than previous approaches.

**CONCLUSION**

In this research, a diabetes mellitus diagnosis ontology framework that will cater for the three-major diabetes type has been presented. A methodology for extracting diabetes concepts based on ontology has been proposed. This purpose of utilizing an ontology is to make it serve as a Case Based Reasoning system as foundation for field knowledge for the facilitation of semantic case retrieval. All the diabetes diagnosis and the laboratory test context are contained in the ontology employed. The framework was demonstrated using protégé and connected to java environment using JENA API. The implementation process showed that the system is easier to use, efficient, effective, and maximally beneficial in the medical domain. The ability for new discovery of symptoms to be included is an indication that the system has high level of flexibility. The use of ontology to such system aids the case reasoning of the system by enabling past diagnosis not to be re-diagnosed again which could lead to a waste of time. We can therefore conclude that, if this approach is used, there will be full diagnosis of the entire diagnosis type with prescription and recommendation with an efficient reasoning ability. Future work can be carried out using more symptom tests based on patients’ unique body physiological and anatomical structure. Also, the framework could be implemented in other life threatening diseases such as cancer.

**REFERENCES**


El-Sappagh, S., El-Masri, S., Elmogy, M., Riad, A. M. 2014. A diabetes diagnostic domain ontology for CBR system from the conceptual model of SNOMED CT. In Engineering and Technology (ICET), 2014 International Conference (pp. 1-7). IEEE.


A FRAMEWORK FOR ONTOLOGY-BASED DIABETES DIAGNOSIS...


(Manuscript received: 14th June, 2018; accepted: 27th August, 2019).