IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Implementation of Diabetes Monitoring System Based on EHR Data Analysis

Sachin B S[1], Prema N S[2], Suraj L Kalro[3], Harshitha[4], Ranjitha Raj K N[5]

Student, Dept. of IS&E, Vidyavardhaka College of Engineering, Mysore, India[1, 3, 4, 5]

Associate Professor, Dept. of IS&E, Vidyavardhaka College of Engineering, Mysore, India[2]

Abstract: This paper demonstrates the framework of a system built to monitor health records for diabetes patients and extracting data on a frequent basis to objectively predict the impending health risks using a version of Bayesian network. The system is deployed on a SaaS cloud service to facilitate secure and convenient collection of data from patients using a mobile application and a web console.

Keywords: Diabetes Mellitus, Data mining, Bayesnet, Support vector, patient monitoring, HER.

I. INTRODUCTION

Health care providers have been managing various patient records since the dawn of medical practices. However, during the 21st century, information technology has assisted them in maintaining patient health records electronically. This has eliminated the need to maintain tedious paper records and more importantly it has facilitated the maintenance of long term medical history of a patient making diagnoses much more accurate and precise for the patient's existing medical conditions.

Although Electronic Health Records (EHRs) assist medical practitioners they have not been given independent inference functionalities to specific conditions. The reason being multiple diseases share several indications/symptoms and without intelligible data, it is hard and also a risky proposition to rely on an electronic system. [3] However, there are several diseases that require continued monitoring even though the patient is not available for observation always.

One such condition is Diabetes Mellitus. Diabetes is a widespread health issue prevalent in the world today. The World Health Organisation estimates that 371 million people worldwide suffer from Diabetes currently and the numbers are estimated to rise up to 552 million by 2030. Continuous Medical Assessments and feedback is an absolute necessity for a health issue that has no permanent cure. Such a scenario calls for computing systems that can help achieve data storage and analysis on a single platform using data analytics techniques which allow dynamic analysis and prediction of risk associated with the patient condition.

This provides patients a system to keep track of their health condition and report the same to the concerned medical professionals without physically having to report every time.

This makes the life of patients easier and also greatly reduces the risk of wrong diagnosis due to inconsistent or no health records.

II. REVIEW OF SYSTEM ACCESS AND USER IDENTIFICATION

Managing patient data on a remote access system requires the maintenance of data privacy, blocking unauthorized user access and ensuring personal identification. Hence, the diabetes monitoring system adopts a multi-layered authentication technique. The levels involved are:

- User registration
- Authentic user acceptance/ Administrator scrutiny
- Authorized Login

New users who wish to have access to the system must register themselves on the web console made available through the remote server-hosting system. Their request is then submitted to the system administrator who determines whether the user is the one who purports to be. Upon such confirmation a user is activated/ allowed to access his account to submit details for his biographical page. Also, if the user is found to be unreal his account may be permanently blocked from registration by the system administrator.

The advantage of the described 2-level model lies in Leveraging the ability to separate the authentication process from the EHR application, this research proposes a framework by which authentication of a single EHR system can not only be configured to a single external authentication system but in fact to use any number of authentication systems.

[7] In this model, the authentication event can be performed by any trusted Identity Provider (Administrator). The basic function of an Admin is to be an authoritative source for establishing and maintaining both identities and credentials. An Admin could be a company such as Google, Yahoo!, Microsoft, or MySpace, that offers free services but also tracks relevant identity information. [21] It is important to point out that while all of these Admin can authenticate an individual, it is critical that the identity management system at the local healthcare provider have the ability to map the external

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

Admin's identifier to a user in the local system. For example, many of the free Admin use an email address as the core identifier for users in their systems.

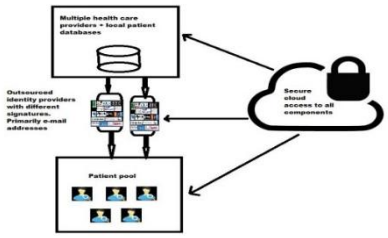

Fig 2.1: Multi-level cloud access mode

III. SYSTEM MODULES AND THEIR INTERACTIONS

The proposed system provides an Interface at the patient end that allows them to upload relevant data such as Blood Glucose levels and their daily physical activities to a remote server by logging in using a private User ID and password.

The data entered by the patients is analysed using Bayesian network algorithm that allows comparisons of data collected to a set of data values that represent threshold values of symptoms that pose possible risks.[22] When such complications are predicted, the system notifies the patient of the need for a visit to the doctor and also notifies the doctor, at his Interface that a patient connected to a particular patient ID needs medical attention.

The system is implemented in object-oriented fashion with multiple modules being integrated to achieve complete system functionality. This allows for easy portability of code to multiple platforms and integration with several outsourced servers.

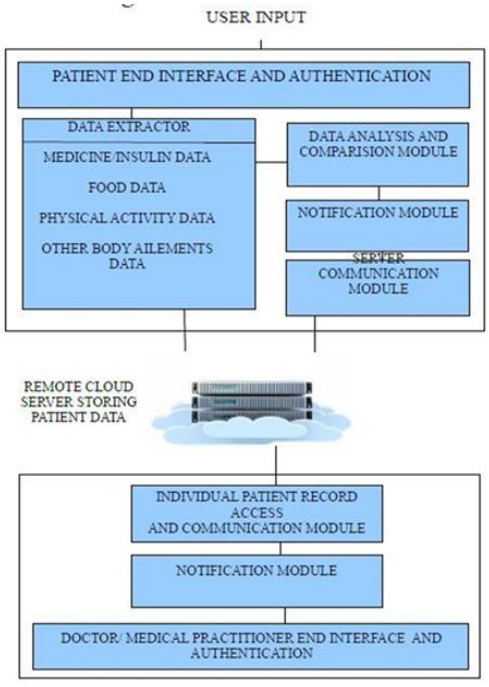

Fig 3.1: System module flow diagram

- **Patient end module:** This module manages all functionalities that are required to be handled for the patient user interface and data collection from patients that are authorized by Admin as described in section 2. A typical scenario includes a patient who is genuine registering himself to request access. The Admin accepts his request and the patient end system assigns him/her a unique patient ID. Once the above activation is done, the patient can start accessing his account. The java server pages allow updating of patient biographical information submitted during registration. This is in turn updated throughout the system with single data server handling all data. Also, the mobile application allows the patient to send specific queries to doctors. Also, the patient glucose and physical activity data is collected.

- **Doctor end module:** This allows medical practitioners as registered stakeholders of the system and allows them to update the detailed health record of the patient, answer patient queries and also obtain notifications on queries from their mobile application. They can view the patient report and once a check-up is done in person, they can add the complete report to the system using the web console.

Similar to the patient end module, a typical scenario includes a doctor who is genuine registering himself to request access. The Admin accepts his request and the patient end system assigns him/her a unique ID. Multiple requests for accounts from the same email address results in the user being blocked access to the system.

- **Knowledge extraction and pre-processing module:** The data obtained using the modules described above entails the extraction of those fields from the database that is necessary for analysis and prediction. To do so, the necessary data is queried as soon as it is obtained and stored in a separate database which behaves like a raw data repository for analysis.

Then the data in this database is further queried to obtain data in necessary format. This is called pre-processing. The continuous attributes are discretized into categorical attributes for analysis. The primary symptoms are compared to parent nodes of the Bayesian network from training data set. The data thus collected is sent to the inference module for calculation of posterior probability.

- **Inference/ Prediction module:** The pre-processed data is input to the Bayesian network classifier that associates attributes with probabilities. Here, the attributes being symptoms for diabetes and also possible high level risks if diabetes (type-1 or type-2) is already present in long term. The pre-processed symptoms being blood glucose level (both fasting and after typical meal) the patient is sent real time messages predicting and warning about his condition.
- **Web services module:** This module connects the two functional units of the system. The web part and the android mobile application. [20] It communicates

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

server message load pages, bring data for authorizing users and store data in appropriate databases for extraction and analysis.

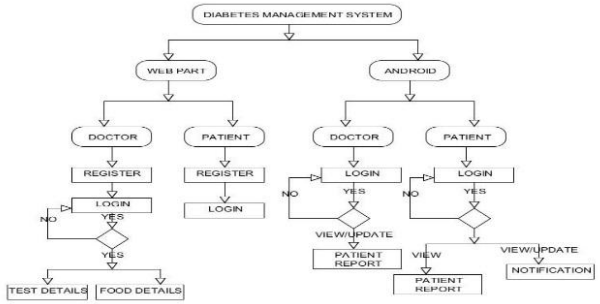

Fig 3.2: System functional units' layout

The complete system is divided into logical units as shown in figure 3.2 and the layout of the modules is elaborated. The 2 main functional divisions being the android and the web units which replicate, in implementation all modules listed above.

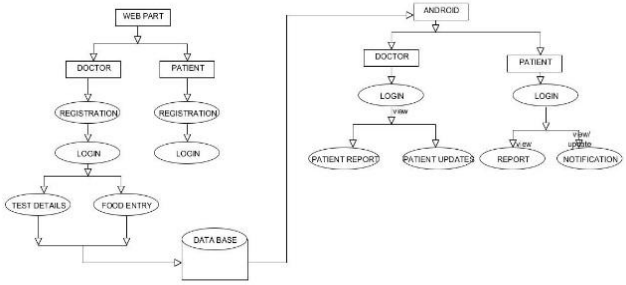

Fig 3.3: Module interaction and data exchange

Web services handle major data exchange between android and the web parts acting as a bridge. It also maps specific IP addresses of local wide area network(WLAN) Internet service providers making the system live and available for remote access across the Inter-web network.

IV. KNOWLEDGE PRE-PROCESSING AND INFERENCE METHOD

A. Variable elimination algorithm for data pre-processing

The point of the variable-elimination algorithm is that it is more bottom-up than top-down: instead of figuring out the probabilities we need to compute and then computing all the other probabilities that each one depends on, we try to compute probabilities and then compute the other terms that depend on them, and repeatedly simplify the expression until we have something that is in terms of only the variable we're looking for.

The variable-elimination algorithm uses things called factors. A factor is basically a conditional probability table, except that the entries are not necessarily probabilities (but they would be if you normalized them). You can think of a factor as a matrix with a dimension for each variable, where Factor [VAL1] [VAL2][...] is (proportional to) a probability such as P(VAR1=VAL1, VAR2=VAL2, ...); or you can think of it as a table with one row for each possible combination of assignments of values to the variables.

The ELIMINATION-ASK function, like ENUMERATION-ASK, takes a variable X and returns a distribution over X, given some evidence e. [5] First it initializes the list of factors; prior to any simplification, this is just the conditional probability tables for each variable given the evidence e. Then, it sums out each variable from the list of factors. The summing-out process takes all the factors that depend on a given variable and replaces them with a single new factor that does not depend on that variable (by summing over all possible values of the variable). By the end of the loop, all the variables have been summed out except the query variable X, so then we can just multiply the factors together and normalize to get the distribution.

THE VARIABLE-ELIMINATION ALGORITHM

function ELIMINATION-ASK(X, **e**, bn) **returns** a distribution over X
inputs: X, the query variable **e**, observed values for some set of variables **E**
bn, a Bayes net
factors ← [for each variable v in bn.VARS, the CPT for v given **e**]
for each var **in** bn**.vars if** var **is not in e and** var **is not** X **do**
relevant-factors ← [all factors that contain var]
factors. remove(relevant-factors)
factors. Append (SUM-OUT(var,POINTWISE-PRODUCT (relevant-factors)))
return NORMALIZE(POINTWISE-PRODUCT(factors))

B. Bayesian network algorithm for probability computation

The Bayesian network used for prediction in the monitoring system is a variation of the algorithm proposed by Russell and Norwig and provides clearer inferences for pre-computed probabilities. A sample network built after variable elimination yields 2 nodes at the base level for diabetes attributes. After several[23] dependency eliminations in pre-processing the network is as shown in the figure below.

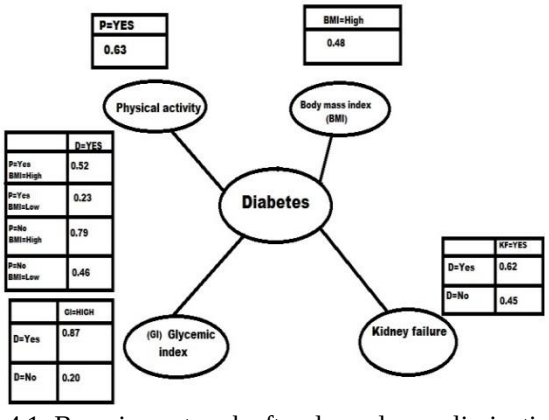

Fig 4.1: Bayesian network after dependency elimination

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 5, May 2016

The enumeration algorithm is a simple, brute-force algorithm for computing the distribution of a variable in a Bayes net. The ENUMERATION-ASK function takes a variable X and returns a distribution over X, given some evidence e. In this case the variable X being the probability of diabetes and e is the data collected from UCI data repository for diabetics which allows for variable dependency determination.

The evidence e is whatever values you already know about the variables in the Bayes net. Evidence simplifies your work because instead of having to consider those variables' whole distributions, you can assign them particular values, so they are no longer variables, they are constants. In the most general case, there is no evidence.

The ENUMERATION-ASK function has to compute a distribution over X, which, because X is a discrete variable, means computing the probability that X takes on each of its possible values (the values in its domain).[5] The algorithm does this simply by looping through all of the possible values, and computing the probability for each one. Note that if there is no evidence, then it is literally just computing the probabilities $P(X=x_i)$ for each xi in X's domain. If there is evidence, then it is computing $P(\mathbf{e}, X=x_i)$ for each xi in X's domain – that is, it is computing the probability that X has the given value (x_i) and the evidence is true – so in that case, we use the he multiplication law of probability, which says that $P(X=x_i \mid \mathbf{e}) = P(\mathbf{e}, X=x_i) / P(\mathbf{e})$, And the fact that P(**e**) is constant: once we have computed $P(\mathbf{e}, X=x_i)$ for all x_i, we can just normalize those values to get the correct distribution $P(X \mid \mathbf{e})$.

THE ENUMERATION ALGORITHM

function ENUMERATION-ASK(X, **e**, bn) **returns** a distribution over X
inputs: X, the query variable **e**, observed values for some set of variables **E**
bn, a Bayes net
Q ← a distribution over X, where **Q**(x_i) is $P(X=x_i)$
for each value x_i that X can have **do**
Q(x_i) ← ENUMERATE-ALL(bn.VARS, $\mathbf{e}_{x_i}$), where $\mathbf{e}_{x_i}$ is the evidence **e** plus the assignment $X=x_i$
return NORMALIZE(**Q**)
function ENUMERATE-ALL (vars, **e**) **returns** a probability (a real number in [0,1])
inputs: vars, a list of all the variables
e, observed values for some set of variables **E**
if EMPTY(vars) **then return** 1.0
Y ← FIRST(vars)
if Y is assigned a value (call it y) in **e then**
return P(Y=y | values assigned to Y's parents in **e**) × ENUMERATE-ALL(REST(vars), **e**)
else
return Σ_{y_i} [P(Y=y_i | values assigned to Y's parents in **e**) × ENUMERATE-ALL(REST(vars), $\mathbf{e}_{y_i}$)], where $\mathbf{e}_{y_i}$ is the evidence **e** plus the assignment $Y=y_i$

The helper function ENUMERATE-ALL is uninformative named, but basically what it is doing is computing (something proportional to) P(**e**), the probability of the evidence **e** (actually, it is only computing it for the variables in **e** that are in vars, so in the top-level call where vars is all the variables in the Bayes net, it is computing it for **e**). The arguments are vars (a list of all the variables we have left to look at) and **e** (a list of all the variables that already have assigned values, along with those values). [18] At each call of this function, we look at the next variable Y from vars (or in the base case, there are no variables left to look at, so we're done). There are two cases. In the first case, Y is already assigned in **e** to some value y, so P(**e**) is just P(Y=y | the rest of **e**) × P(the rest of **e**). In the second case, Y is not assigned, so we have to sum over all possible values yi in Y's domain.

To enumerate the effectiveness of Bayes net classification we use Support Vectors to verify the basic attributes used for Bayesian network in each instance determining the margin of error in the computed probabilities. [9] An instance of the dataset looks like it is shown in figure 4.2 below.

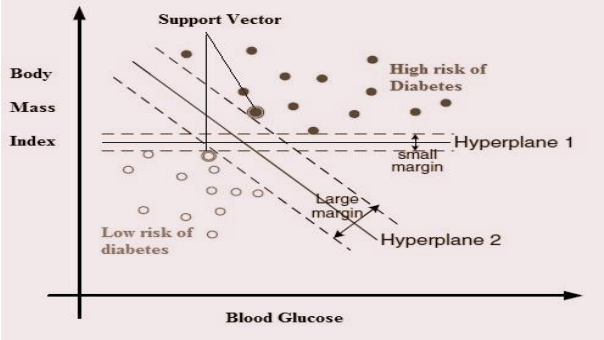

Fig 4.2: Support vector classification

V. SCOPE FOR FUTURE WORK

Diabetes monitoring system implemented as described in above sections in several ways makes the analysis and diagnosis of the disease better. However, providing remote access also means the system might not always be as secure as intended. For instance, an unauthorized user with access to a genuine patient's mobile application can log false values and send irrelevant queries. This causes unreal values to be in for analysis and reduces the overall accuracy of prediction.

Also, the system requires an active Internet connection to access the cloud database. Sometimes, the patient might not have access to Internet while he wants to log values into the system. This problem can be eliminated by buffer repositories on local storage so that values are not lost and are logged to the database once Internet is available.

VI. CONCLUSION

Diabetes is a widespread condition in today's world with multiple complications many of which are fatal. Kerbing such a widespread disorder requires self-monitoring such as glucose testing, recording blood glucose levels, medication use and self-care activities such as dietary

control and regular exercise. Self-monitoring will provide information that is essential to bring glucose levels under control by adjusting the diet, exercise and medication regimen.

[3] Also, recording other health information will help individuals to monitor other complications that may arise.

Diabetes monitoring system provides holistic approach to self-monitoring and up to a large extent achieves its purpose of alleviating the problems involved in managing and keeping the condition in control.

REFERENCES

[1]. T. Bodenheimer, K. Lorig, H. Holman, and K. Grumbach, "Patient self-management of chronic disease in primary care", Journal of American Medical Association (JAMA), pp. 2469-2475, 2002.

[2]. David L. Duke, "Intelligent Diabetes Assistant: Using Machine Learning to Help Managing Diabetes", International Conference on Computer Systems and Application, pp. 913-914, 2008.

[3]. Muhammad Syaifuddin, Kalaiarasi Sonai Muthu Anbananthen, "Framework: Diabetes Management System", IMPACT-2013, pp. 112-116, 2013.

[4]. Brian Coats, Subrata Acharya, "Bridging Electronic Health Record Access to the Cloud", 47th Hawaii International Conference on System Science, pp. 2948-2957, 2014.

[5]. Massachusetts Institute of Technology, courses on Computer science at CSail, retrieved November 2015, "http://courses.csail.mit.edu/6.034s/handouts/spring12/bayesnets-pseudocode.pdf"

[6]. Blumenthal, D., & Tavenner, M. (2010). The "Meaningful Use" Regulation for Electronic Health Records. New England Journal of Medicine, 363, 501-504. doi:10.1056/NEJMp1006114.

[7]. Z. Zheng, "Constructing New Attributes for Decision Tree Learning", Ph.D. Thesis, University of Sydney, Australia, 2002.

[8]. J. Gama, "Combining Classification Algorithms", PhD Thesis, University of Porto, 2000.

[9]. S. Agarwal and G. N. Pandey Divya, "SVM based network for pervasive healthcare monitoring", Proceedings of the First International Conference on Intelligent Interactive Technologies and Multimedia. ACM, 2010.

[10]. M. Singh, P. K. Wadhwa and P. S. Sandhu, "Human Protein Function Prediction using Decision Tree Induction", International Journal of Computer Science and Network Security, vol. 7, no. 4, (2007), pp. 92-98.

[11]. M.M. Alotaibi, R.S.H. Istepanian, A.Sungoor and N. Philip, "An Intelligent Mobile Diabetes Management and Educational System for Saudi Arabia: System Architecture", IEEE conference proceedings, pp. 29-32, 2014.

[12]. L. Piemonte, "5th Edition of the Diabetes Atlas released on World Diabetes Day." Available: 'http://www.idf.org/diabetesatlas/news/fifthedition-release".

[13]. R.S. Istepanian, K. Zitouni. D. Harry, N. Moutosammy, A. Sungoor, B. Tang, et al., "Evaluation of a mobile phone telemonitoring system for glycaemic control in patients with diabetes." J Telemed Telecare, vol. 15, no. 3, pp.125–128. 2009.

[14]. Shovon K. Pramanik, Subrata Pramanik, Bimal K. Pramanik, M. K. Islam Molla and Md. Ekramul Hamid, "Hybrid Classification Algorithm for Knowledge Acquisition of Biomedical Data", International Journal of Advanced Science and Technology Vol. 44, July, 2012, pp. 99-112, 2012.

[15]. C. Apte and S. Weiss, "Data Mining with Decision Trees and Decision Rules", Future Generation Computer Systems, vol. 13, pp. 197-210, 1997.

[16]. Divya Tomar, Sonali Agarwal, "A survey on Data Mining approaches for Healthcare", International Journal of Bio-Science and Bio-Technology Vol.5, No.5, pp. 241-266, 2013.

[17]. D. S. Kumar, G. Sathyadevi and S. Sivanesh, "Decision Support System for Medical Diagnosis Using Data Mining", 2011.

[18]. Silwattananusarn, Tipawan, and KulthidaTuamsuk. "Data Mining and its Applications for Knowledge Management: A Literature Reviewfrom 2007 to 2012." arXiv preprint arXiv:1210.2872, 2012.

[19]. C. H. Jena, C. C. Wang, B. C. Jiangc, Y. H. Chub and M. S. Chen, "Application of classification techniques on development an early-warning systemfor chronic illnesses", Expert Systems with Applications, vol. 39, pp. 8852-8858, 2012.

[20]. M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, M. Zaharia: "Above the Clouds: A Berkeley View of Cloud Computing." Technical Report UCB/EECS-2009-28, University of California at Berkeley, California, U.S.A, February 2009.

[21]. Wikipedia: Megaupload. Available at: http://en.wikipedia.org/wiki/Megaupload. Last visited on: March 2016.

[22]. K. F. R. Liu and C. F. Lu, "BBN-Based Decision Support for Health Risk Analysis", Fifth International Joint Conference on INC, IMS and IDC, 2009.

[23]. J. Yanqing, H. Ying, J. Tran, P. Dews, A. Mansour and R. Michael Massanari, "Mining Infrequent Causal Associations in Electronic Health Databases", 11th IEEE International Conference on Data Mining Workshops, 2011.