International Journal of Pure and Applied Mathematics
Volume 120 No. 5 2018, 1441-1458
ISSN: 1314-3395 (on-line version)
url: http://www.acadpubl.eu/hub/
Special Issue

http://www.acadpubl.eu/hub/

Intensive Disease Support Measure Based Efficient Disease Prediction Using Artificial Neural Network

Ramya B[1], Dr.K. Arthi[2]

[1]Ph.D Research Scholar, Bharathiar University,

Assistant Professor, Department of Computer Applications,

Dr.N.G.P Arts and Science College (Autonomous), Coimbatore-641046, India.

Email: ramyab@drngpasc.ac.in

[2]Assistant Professor, Department of Computer Application,

Government Arts College (Autonomous), Coimbatore – 641 018.

Email: sek_art@rediffmail.com

Abstract:

The problem of disease prediction has been a challenging task to the researchers and variety of methods has been studied. However, the most approaches suffer to achieve higher performance in disease prediction. To overcome the deficiency, a novel intensive disease support (IDS) based disease prediction with Artificial neural network has been presented. The methods preprocess the medical data set given to remove the noisy data points. The data points which are identified as incomplete have been removed from the data set. Second, the method groups the data points of different disease class. Grouped data points are applied with neural network where each neuron estimates intensive disease support measure for various disease classes. Finally, a cumulative disease influence measure (CDIM) has been estimated. Based on the CDIM, possible disease has been identified. The method produces efficient results on disease prediction and reduces the false ratio.

Index Terms:

Medical Data Set, Disease Prediction, IDS, CDIM, ANN.

Introduction:

The modern society has great influence of various diseases due to the changing life style and other constraints. However, the most diseases have similar symptoms and the medical practitioner struggle to identify what disease being affected by any person. For example, the general fever has been identified by the presence of body pain and higher body temperature. But these symptoms are more common for other diseases like Dengue, malaria and typhoid. By having such overlapping constraints, the medical practitioner could not identify the possible disease in the human. Disease prediction is the process of predicting possible disease based on the presence of set of symptoms. Human errors always present in the disease prediction and cannot achieve higher performance in disease prediction. To solve this issue, some strategic approaches are highly required to support medical practitioners.

The disease prediction can be performed by several ways. According to the symptoms considered and by maintaining the history of various diseased persons, the disease prediction can be performed. When such symptoms set are matching in higher frequency with any disease class, the particular disease can be selected. As discussed above, the body temperature and the pain are appearing in more frequently in different class of diseases, the disease prediction would produce poor results with higher false classification. To support the problem of disease prediction, there are number of approaches available which uses various methods and measure. Still they suffer to achieve higher performance in disease prediction and needs to be improved.

The medical data set being considered has higher role in the problem of disease prediction. The medical data are framed by collecting the traces of patients with various diseases. They can be classified under different class of disease and can be used for the problem of disease prediction. Different medical organizations maintain the data set in different form and to perform disease prediction, they have to merge to a single form which yields an in complete data. To solve this problem, it is necessary to preprocess the data set and identify the noisy incomplete data points. When the data set is prepared, they can be used for disease prediction.

Towards the problem of disease prediction, an efficient approach has been presented in this paper. Any disease has particular set of symptoms in higher ratio which has to be considered. As of the same symptoms has been present in different class, it is necessary to measure the importance of the symptom

on each class. According to this, an Intensive Disease Support (IDS) based approach is presented in this paper. The IDS measures represent the influence of the symptoms on different class of diseases. It has been measured based on the occurrence of the symptom in different class and its importance on each class. When the IDS measure has been computed with each class, the CDIM (cumulative disease influence measure) can be computed to identify the possible disease. The detailed approach is discussed in the next sections.

Related Works:

Different methodologies are presented for the problem of disease prediction. This section review set of approaches for the disease prediction problem.

Application of data mining methods in diabetes prediction [1], works on the risk of high blood pressure in diabetic patients. The method uses various data mining techniques including SVM, ANN, ELM and GMM. Similarly in [2], the author analyzed different techniques of data mining for the prediction of diabetes mellitus. This support the prediction of diabetes in the early stage using PIMA data set.

In [3], a different wavelet transform technique has been adapted for the prediction of coronary artery disease. The heart rate signal has been used for the automatic prediction. Different frequencies of heart signals are applied with wavelet transform which extract specific coefficients. Extracted coefficients are used to perform classification using different models.

In [4], the spherical clustering algorithm has been applied for the clustering of facial feature to perform face expression. The method uses curvelet transform and radial basis functions. This identifies the hidden signals which can be adapted to the problem of disease prediction. In [5], an auto regressive model (ARX) has been proposed which monitors and analyze the diabetic patients towards the prediction. The method collects the patient data through the wireless network and based on the blood pressure, and cholesterol the prediction has been performed.

In [5], an continuous glucose monitoring (CGM) has been presented. The method considers the temporal data by monitoring the changing rate of glucose. Based on the changing rate, the prediction is performed.

In [7], a support vector regression model is adapted for the prediction of diabetic and glucose level. The methods consider different psychological conditions which affect the glucose level. The performance has been compared with different analytical models like GA and ANN.

The glucose concentration in human has been predicted using CGM techniques in [9].different type of diabetes has been considered for the prediction using auto regressive model. In [10], the mobile platform has been used for the prediction of blood sugar. The method uses the support vector regression model and considered the psychological conditions of the patients.

A personalized prediction model for Type 1 Diabetes is presented in [11], which uses real time analysis model. The performance has been compared with the result of auto regression and ANN. In [12], a neuro fuzzy based prediction has been presented for the blood glucose. The method has been applied with the patients affected by Type 1 diabetes. The methods monitor the changes in metabolic behavior. Based on the changing behavior values, the future glucose level has been predicted. Similarly, the future glucose level has been predicted with subject specific recursive linear model. The method uses the time series data to uses the variability in data and consider the relation between them.

Expectation maximization (EM) algorithm is presented for the predictions of diabetes in [14], which consider the delay in the glucose level. The method uses ARX model to monitor the changing dynamics of glucose. The algorithm has been evaluated for its performance based on the data collected from Chinese hospitals. In [15], prediction of diabetes has been performed through the mobile devices which has been used to perform data collection, where the prediction is performed using the neural network. In [16], multi variant regression models has been used for the glucose prediction which consider the energy loss due to physical activity, meal derived glucose, insulin concentration and glucose profile.

In [17], a Non-linear Dynamic Model has been used to support short term prediction. The method used the Fixed Budget Quantized Kernel Least Mean Square (QKLMS-FB) algorithm to design the model and to perform prediction. In [18], a statistical method has been presented to identify the pathway for the cancer. The author defines a set of pathway for the disease class which uses the Sparse Probabilistic Principal Component Analysis (SPPCA). In [19], a gene subset selection algorithm has been presented for cancer classification. The method uses the Tabu search algorithm for classification and to measure the similarity.

In [20], the author presented a iterative influence measure based medical data classification algorithm which estimates the influence measure on multiple levels to identify the target class.

All the above discussed methods suffer to achive higher performance in disease prediction.

CDIM-ANN Based Disease Prediction:

The proposed cumulative disease influence measure with artificial neural network based disease prediction algorithm, reads the input data set and applies preprocessing technique to remove the noisy data points. The noise removed data set has been split into number of disease classes. The classified data set has been used to train the neural network. At the testing phase, the input symptom set has been given as input and the neuron at the output layer produces set of cumulative disease influence measure. Based on the CDIM value a single disease has been selected. The detailed approach is discussed in the next section.

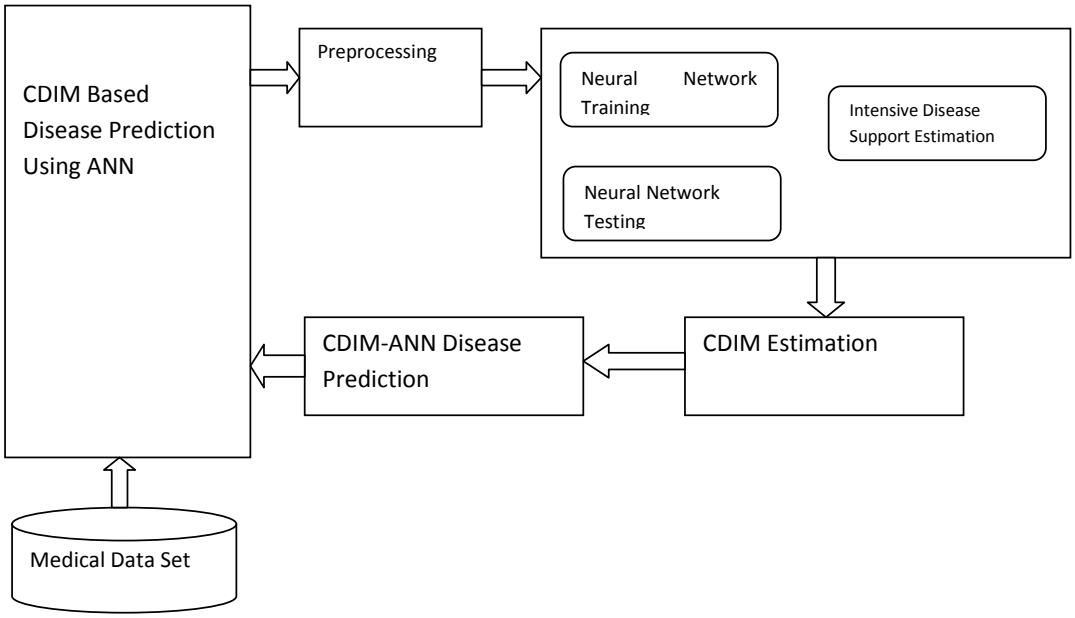


Figure 1: Architecture of Proposed CDIM-ANN Disease Prediction

The Figure 1, present the architecture of the proposed CDIM ANN based disease prediction algorithm and shows various components involved.

Preprocessing:

In this stage, the input medical data set has been prepared for processing. First, the data point of the data set has been identified. Second, the list of attributes present in the data set has been identified. Third, each data point has been verified for the presence of all the dimensions present in the dimension list. If any of the data point has been identified as missing dimension and any of the data attribute has no value, such data points are considered as noisy data points and removed from the list. Fourth, the list of disease class has been identified and the data points are grouped under different class names. The noise removed data set has been used to estimate different measures on disease prediction.

Algorithm:

Input: Medical data set M

Output: Disease class set Dcs

Start

Read input data set M.

Identify list of data points Dpl = $\sum_{i=1}^{size(M)} M(i) \nexists Dpl$

Identify list of attributes Atl = $\sum_{i=1}^{size(Dpl)} \sum_{j=1}^{size(i)} Dpl(i)(j) \nexists Atl$

For each data point di

If $\sum_{i=1}^{size(Atl)} Di \nexists Atl(i)$ then

Remove data point Di

Else

Add data point to Ps = $\sum(Di \in Ps) \cup di$

End

End

Identify list of disease class Dc = $\sum_{i=1}^{size(Dpl)} Dpl.Disease \exists Dc$

For each disease class Dc_i

Identify list of data points Dcs(Dci) = $\sum_{i=1}^{size(Ps)} Ps(i).Disease == Dci$

End

Stop

The above discussed algorithm preprocesses the input data set and groups the data points of different disease class.

ANN Training/Testing:

The preprocessed data set has been used to train the neural network. For each disease class, a single neuron has been initialized with list of data points of the class identified in the previous stage. When the input data point is given, each neuron estimates the intensive disease support measure for the input data point towards the particular disease class data points. Similarly, there will be N number of IDS measures are estimated according to the number of neurons generated. In the output layer, the neurons produce single value which has been used to estimate the cumulative disease influence measure to perform disease prediction.

Algorithm:

Input: Disease class set Dcs, Test Sample Ts.

Output: Neural network NN, IDS.

Start

Read DCS.

If Training then

Initialize no of neurons NoN = size(Dcs)

For each disease class Dc

Neuron(Dc) = $\sum_{i=1}^{size(Dcs(Dc)} Dcs(dc(i))$

End

Add to NN.

Else

For each neuron N

IDS = Estimate intensive disease support (NN,Ts)

End

End

Stop

The above pseudo code shows how the network has been trained or tested with input sample, which in turn returns the set of IDS values and neural network being trained.

Intensive Disease Support Estimation:

In this stage, the neuron receives the input test sample. According to the test sample given, the method estimates the similarity of the values of each dimension in training values. For each data point available in the training set, the method estimates the feature similarity by estimating the distance between the values of each dimension. It has been measured for all the data points. Finally, according to the similarity value, a single IDS measure has been estimated.

Algorithm:

Input: Test Sample Ts, Neural Network NN

Output: IDS

Start

Read input sample Ts.

Read NN.

For each training sample Trs

For each dimension Di

Compute feature similarity Fs = $\frac{\sum_{i=1}^{size\,(Trs)} Dist\left(Ts(Di),Trs(Di)\right)<Th>}{size\,(Ts)}$

End

End

Compute IDS = $\frac{\sum_{i=1}^{size\,(Trs)} Trs(i).Fs>Th}{size\,(Trs)}$

Stop

The above discussed algorithm estimates the intensive disease support measure for the input test sample towards a specific disease class. Estimated IDS measure has been used to perform disease prediction.

CDIM Estimation:

The cumulative disease influence measure represent the influence of the test sample in particular disease. According to the given test sample, the method estimates the IDS measure for different disease class. Based on the IDS measure being estimated, the method estimates the CDIM measure for different classes. Estimated CDIM measure has been used to perform disease prediction. Finally, the disease class with higher CDIM value has been selected as target disease class.

Algorithm:

Input: Neural Network NN, Test Sample Ts

OutSput: CDIM

Start

Read NN, Ts.

For each Disease class Dc

DiDs = Estimate IDS

ODiDS= Estimate IDS for other class.

Estimate CDIM = DiDs×ODIDS

End

Stop.

The above discussed algorithm estimates the CDIM measure for specific class given and has been used to perform disease prediction.

CDIM-ANN Disease Prediction:

The disease prediction has been performed using the artificial neural network which has been trained initially. At the testing phase, the method computes the CDIM measure for different disease class which in turn computes the intensive disease support measure for all the classes. Finally Based on the CDIM measure obtained, a single target class has been selected.

Algorithm:

Input: Data Set Md, Test sample Ts.

Output: Disease D

Start

Read Md, Ts.

Disease class set Dcs = Preprocessing(Md)

For each disease class Dci

Estimate CDIM measure.

End

Choose the disease class with higher CDIM value.

Stop

The above discussed algorithm shows how the disease prediction has been performed.

Result and Discussion:

The CDIM-ANN algorithm has been designed and developed using Matlab. The performance of the CDIM-ANN algorithm has been evaluated using different data sets. The result produced by the proposed method has been compared with other approaches.

Parameter	Value
Data Set Name	PIMA, UCI, Vanteer
No of Dimensions	9,14, 12
No of Data Points	768,275,240
No of classes	5

Table 1: Data set for evaluation

The data sets considered for the performance evaluation of different methods is presented in Table 1. The PIMA data set is obtained from the Indian institutions which has the features of insulin, age, no of pregnancy, BMI and so on. The UCI data has been collected from UCI repository. Similarly the vanteer data set has details above various users which contain much information.

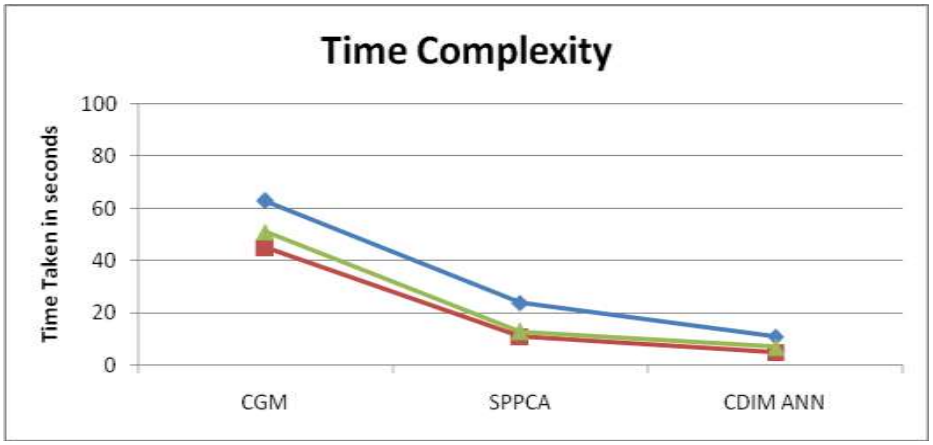
Time Complexity:

The time complexity of classification and prediction has been measured. It has been estimated based on the time value taken for prediction with X number of samples. It has been approximated for different number of samples.

Method Name	PIMA	UCI	Vanteer
CGM	63	45	51
SPPCA	24	11	13
CDIM ANN	11	5	5

Table 2: Comparison on classification time complexity

The performance on time complexity has been measured and presented in Table 2. The result clearly points that the proposed approach has reduced the time complexity than other methods.



Graph 1: Performance on time complexity

The performance of the methods on time complexity has been measured and compared. The comparison result is presented in Graph 1, which justifies that the proposed CDIM-ANN algorithm produces less time complexity than other approaches.

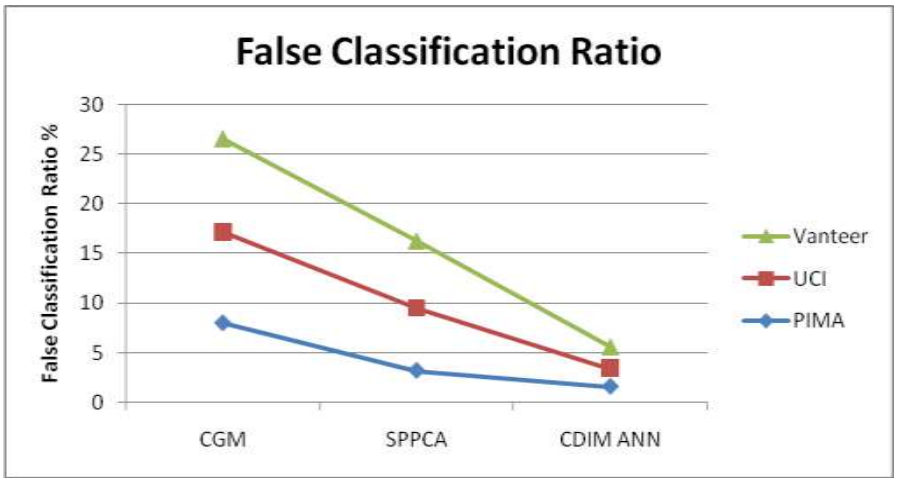
False Classification Ratio:

As the method includes the higher number of features, the false classification ratio will get reduced. Also the naïve Bayes and svm algorithms does not consider the range values and because of including the fuzzy rules here, the false classification ratio will be reduced and the same implies on the increase of prediction accuracy.

Method Name	PIMA	UCI	Vanteer
CGM	8	9.2	9.4
SPPCA	3.2	6.3	6.8
CDIM ANN	1.6	1.8	2.2

Table 3: Evaluation on false ratio

The method has been evaluated for its false classification. The result obtained has been compared with the false ratio of other methods and presented in Table 3. However, the proposed algorithm has reduced the false ratio than previous methods.



Graph 2: Comparative study on false ratio

The comparative study on false classification ratio of different methods has been performed. The result of comparative study has been presented in Graph 2. It clearly depict that the proposed algorithm has produced less false classification ratio.

Prediction accuracy:

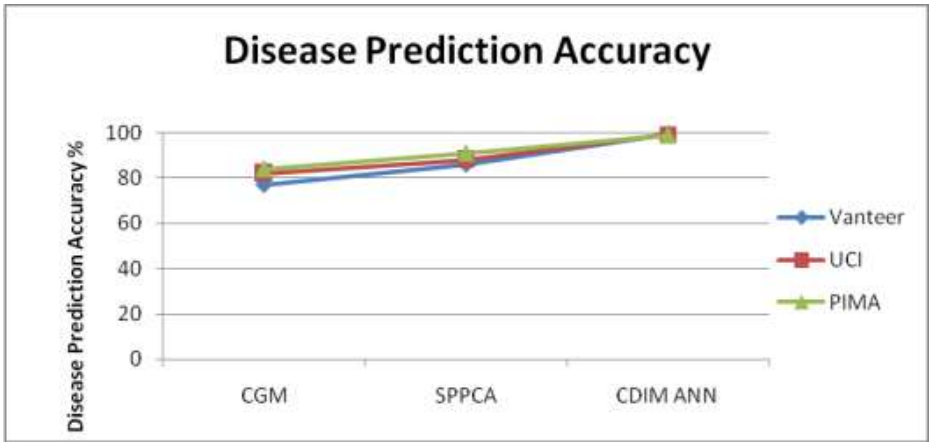
The prediction accuracy has been increased, by including different characteristic features. The lifestyle feature has been included in the base algorithm but does not include the physical characteristics like hours of exercise and pancreas with eye conditions. Because the symptom of disease

will be first displayed in the eyeballs and pancreas only. The proposed algorithm include these features so that the accuracy will be increased

Method Name	Vanteer	UCI	PIMA
CGM	77	82	84
SPPCA	86	88	91
CDIM ANN	99.5	99.3	99

Table 3: Comparative study on prediction accuracy

The proposed method has been evaluated for its performance in prediction accuracy. The result of prediction accuracy has been presented in Table 3.



Graph 3: Comparative result on disease prediction accuracy

Different methods have been considered for performance analysis. Their efficiency in disease prediction has been measured and compared with different methods. However, the proposed CDIM-ANN approach has produced higher prediction accuracy than any other method considered.

Conclusion:

In this paper an efficient CDIM ANN based disease prediction algorithm has been presented. The methods preprocess the input data set to perform noise removal and group the data point of different classes. The preprocessed data set has been used to train the neural network. At the testing phase, the

method estimates IDS and CDIM measures for each disease class. Finally the disease class with higher CDIM value has been selected as target class. The method produces higher efficiency in map reduce, classification and prediction accuracy.

References:

1. Messan Komi ; Jun Li ; Yongxin Zhai ; Xianguo Zhang,Application of data mining methods in diabetes prediction, IEEE Conference on Image, Vision and Computing (ICIVC), 2017 .

2. Devi, M. Renuka, J. Maria Shyla, "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus", International Journal of Applied Engineering Research, vol. 11, no. 1, pp. 727-730, 2016.

3. Giri Donna et al., "Automated diagnosis of coronary artery disease affected patients using LDA PCA ICA and discrete wavelet transform", Knowledge-Based Systems, vol. 37, pp. 274-282, 2013.

4. Ayşegül Uçar, Yakup Demir, Cüneyt Güzeliş, "A new facial expression recognition based on curvelet transform and online sequential extreme learning machine initialized with spherical clustering", Neural Computing and Applications, vol. 27, no. 1, pp. 131-142, 2016.

5. Geshwaree Huzooree ; Kavi Kumar Khedo ; Noorjehan Joonas ,Glucose prediction data analytics for diabetic patients monitoring, IEEE Conference on Next Generation Computing Applications (NextComp), 2017.

6. E. Aboufadel, R. Castellano, D. Olson, "Quantification of the variability of continuous glucose monitoring data", Algorithms, vol. 4, no. 1, pp. 16-27, 2011

7. R. Bunescu, N. Struble, C. Marling, J. Shubrook, F. Schwartz, "Blood Glucose Level Prediction Using Physiological Models and Support Vector Regression", 2013 12th Int. Conf. Mach. Learn. Appl., vol. 1, pp. 135-140, 2013.

8. R. H. E. Botwey, E. Daskalaki, P. Diem, S. G. Mougiakakou, "Multi-model data fusion to improve an early warning system for hypo-/hyperglycemic events", Conf. Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. IEEE Eng. Med. Biol. Soc. Annu. Conf., vol. 2014, pp. 4843-4846, 2014.

9. A. Gani, A. V. Gribok, Y. Lu, W. K. Ward, R. A. Vigersky, J. Reifman, "Universal Models For Predicting Glucose Concentration In Humans", WO Pat. WO/2010/, vol. 14, no. 1, pp. 157-165, 2011.

10. M. P. Reymann, E. Dorschky, B. H. Groh, C. Martindale, P. Blank, B. M. Eskofier, *Blood Glucose Level Prediction based on Support Vector Regression using Mobile Platforms*, pp. 2990-2993, 2016.

11. E. Daskalaki, A. Prountzou, P. Diem, S. G. Mougiakakou, "Real-Time Adaptive Models for the Personalized Prediction of Glycemic Profile in Type 1 Diabetes Patients", Diabetes Technol. Ther., vol. 14, no. 2, pp. 168-174, 2012.

12. K. Zarkogianni, K. Mitsis, A. Fioravanti, K. S. Nikita, "Neuro-Fuzzy based Glucose Prediction Model for Patients with Type 1 Diabetes Mellitus", Ieee, pp. 252-255, 2014.

13. M. Eren-Oruklu, A. Cinar, L. Quinn, D. Smith, "Estimation of Future Glucose Concentrations with Subject-Specific Recursive Linear Models", Diabetes Technol. Ther., vol. 11, no. 4, pp. 243-253, 2009.

14. Z. Fanmao, W. Y ouqing, Dynamic model with time varying delay for type 1 diabetes mellitus identified by using expectation maximization algorithm, pp. 9376-9381, 2016.

15. C. Bayraktar, H. Giimiis, O. Karan, C. Bayraktar, H. Gumiiskaya, B. Karlik, "Diagnosing diabetes using neural networks on small mobile devices", Expert Syst. Appl., vol. 39, no. 1, pp. 54-60, 2012.

16. E.I. Georga, V. C. Protopappas, D. Ardigo, M. Marina, I. Zavaroni, D. Polyzos, D. I. Fotiadis, "Multivariate Prediction of Subcutaneous Glucose Concentration in Type 1 Diabetes Patients Based on Support Vector Regression", Biomed. Heal. Informatics IEEE J., vol. 17, no. 1, pp. 71-81, 2013

17. E. I. Georga, J. C. Principe, D. Polyzos, D. I. Fotiadis, S. Member, Non-linear Dynamic Modeling of Glucose in Type 1 Diabetes with Kernel Adaptive Filters, no. i, pp. 5897-5900, 2016.

18. Shuichi Kawano et al., “Identifying Gene Pathways Associated with Cancer Characteristics viaSparse Statistical Methods” IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 9, no. 4, pp. 966-972, 2012.

19. Jiexun Li, Hua Su, Hsinchun Chen, and Bernard W. Futscher," Optimal Search-Based Gene Subset Selection forGene Array Cancer Classification", IEEE Transactions on Information Technology in Biomedicine, vol. 11, no. 4, pp 398-405, 2007.

20. K.Ananthajothi and M.Subramaniam, Multi level incremental influence measure based classification of medical data for improved classification, Springer, Cluster Computing, 2018.