Data Based Estimation of Near Future Values of Blood Glucose with K-Nearest Neighborhood Algorithm

S.Shanthi, Shyamala Bharath, M.Sujatha

Abstract: ***Diabetes Mellitus is a disease due to the disorder of carbohydrate metabolism. This disease affects the people in various ways in short and long time periods. Research is being carried out in pathological, pharmaceutical, clinical, therapeutic aspects to reduce the impediments of Diabetes. Researchers try to foretell the proximateforthcoming blood glucose values so that the patient could be alerted to take appropriate action with the guidance of medical practitioner. The near future prediction of glucose values is very much needed for a successful artificial pancreas. Since the blood glucose values of human depends on many factors like physiological factors, age, body mass index, glucose metabolism, insulin action etc,, prediction of exact values of blood glucose still remains a tough task. The current research work focuses on the application KNN regression for the forecast of nearby blood glucose values. The KNN algorithm uses the feature similarity to predict any new points on the data sets. The proposed work has been tested with 3 different data sets and the results have been analyzed. Promising results have been obtained which could be extended further with real time analysis.***

Keywords: Artificial intelligence, Continuous glucose monitoring, Diabetes mellitus, Machine learning, KNN algorithm, Prediction.

I. INTRODUCTION

The food we eat is converted to glucose at the end of digestion process. The glucose in the blood stream has to be transformed to energy and will be absorbed by the body cells. Insulin which is secreted from the endocrine pancreas enables this transformation of energy. If there is no insulin secretion or insufficiency in insulin secretion, the glucose stays in the blood itself. This excess glucose levels for prolonged period of time results in the complications of Diabetes mellitus. One has to retain the Blood Glucose levels in the standard range (70 to 120 mg/dL) to avoid the diabetic complications. Lower BG levels (<70 mg/dL) known as Hypoglycemia leads to diabetic coma and seizures. Higher BG levels (>120 mg/dL) known as Hyper glycemia results in Cardio vascular problems, diabetic retinopathy, neuropathy and nephropathy. Type 1 Diabetes (T1D) is due to non secretion of insulin due to the damage of pancreatic beta cells. This occurs in early ages and T1D depend on external insulin for the BG regulation. Type 2 Diabetes (T2D) is due to insufficient insulin secretion and so depends on additional insulin injection or oral medications. Gestational diabetes is during the period of pregnancy[1].

As per the statistics, nearly 429 million adults are living with Diabetes worldwide and this would rise to 645 million in 2045 [2]. Diabetes can be treated and the complications can be avoided by physical activity, diet, regular monitoring and control of BG values in the prescribed levels. The development of Self Monitoring Blood Glucose (SMBG) devices and Continuous Glucose Monitoring Systems (CGMS) assist the diabetic people in regular monitoring and control of BG. SMBG is for random checking of BG levels and CGMS is for the measurement of BG values continuously for a period of 3 to 4 days. The CGM data helps for the analysis of highs and lows of BG values and alert for Hypo glycemia. The CGM data has been used for developing a variety of data driven models for the estimate of adjoining BG values. The forecasting of BG values helps the patient to take appropriate action to prevent the diabetic complications. Research in the prediction of BG values varies from mathematical AR models to the recent Machine learning based methods. The proposed work analyses the effectiveness of applying customized K- Nearest Neighbor (KNN) algorithm for the prediction of BG values. The structure of paper has been organized as follows. Section II narrates the related researchesin the prediction of BG, section III explains the data set and methodology of research work. Section IV analyzes the experimental results and Section V gives the conclusion and further scope of the proposed work.

II. PRIOR WORK

The research in the prediction of BG is being carried out from 1990s in different streams such as regression models, time series analysis, softcomputing methods etc.. Some of the major works have been discussed in this section.Bremer& Gough were the pioneers in the extrapolation of glucose concentration in blood[3]. Based on the recent BG values which are systematic, the near future glucose values could be estimated. In this work, predictionhas been made mathematically with an Auto Regressive Moving Average (ARMA) process, for Prediction Horizon (PH) of 10, 20 and 30 minutes. Palerm et al (2005) had applied a Kalman Filter for the same purpose, and acquired the results with 90% sensitivity and 79% specificity [4].Mougiakakou et al had worked on the modelsof glucose-insulin metabolism of children with T1D[5].The effects of sampling frequency, threshold

Revised Manuscript Received on May 29, 2019.

Dr.S.Shanthi, Professors, Department of ECE, Saveetha School of Engineering, SIMATS, Chennai, Tamilnadu, India
Dr.ShyamalaBharath, Professors, Department of ECE, Saveetha School of Engineering, SIMATS, Chennai, Tamilnadu, India
Dr.M.Sujatha, Professors, Department of ECE, Saveetha School of Engineering, SIMATS, Chennai, Tamilnadu, India

selection and PH on the sensitivity and specificity of foreseeing hypoglycemia have been validated by Palerm&Bequette[6]. A first order polynomial and an Auto Regressive (AR) model had been applied to BG data by Sparacino et al [7]. The performance metrics considered were the Mean Square Error (MSE) and Energy of Second Order Differences (ESOD). This author had alsoused time lag also as a factor, to evaluate the projectingcompetency of the models.

The data driven AR models have also been used by Reifmanet alto capture the associations in glucose time series data. For PH of 30 and 60 minutes, the RMSE is 26 and 36 mg/dLrespectively[8].Quchani&Tahami (2007) compared the performances of a Multi Layer Perceptron (MLP) and Elman neural networks in predicting the BG levels in T1D[9]. Several neural network models with Neuro Solutions software and an electronic diary information have been developed, for the forecast of BG in various PH of 50,75,100,120,150 and 180 minutes[10]. AMean Absolute Difference (MAD) of 43 mg/dL was obtained in the PH of 100 minutes.Eskaf et al (2008) extracted the features from BG time series to obtain knowledge about the food intake and human body activities of the diabetic patient[11]. Then an ANN was trained with these features to predict the future value of BG level in 30 minutes PH.

The predictive data driven models and the frequent BG measurementswere utilized by Gani et al and the results were with clinically acceptable time lags [12]. An user friendly armband has been used to record the physical activity, energy expenditure on a minute-by-minute basis[13]. And glucose concentrations were recorded by using a CGM system. A Gaussian process had been analyzed for the forecast.An Artificial Neural Network (ANN) algorithm had been implemented by Perez-Gandia et al for online glucose prediction [14]. Rollins et al modeled the free-living physical activity data from the armband and relate it to blood glucose[15].Aibinu et al applied windowing to BG data which was normalized to estimate the model (AR and ARMA) parameters with a real valued neural network. The estimated parameters were optimized with genetic algorithm. Then the hybrid ARMA model has been employed to predict the pre-breakfast BG level[16].

A feed forward neural network with specialized transfer functions had been developed by Shanthi et al [17]. The first four statistical moments of a BG time series (Mean, Variance, Skew, and Kurtosis) along with Approximate Entropy (ApEn) were extracted as features and analysis was made on the BG dynamics.

Short-time glucose prediction using past CGM sensor readings and information on carbohydrate intake had been done by Zecchin et al [18]. A neural network (NN) model and a first-order polynomial extrapolation algorithm had been combined to analyze the nonlinear and the linear components of glucose dynamics. The method had been assessed with 20 simulated datasets and on 9 real Abbott Free Style Navigator datasets.

Jensen et alhad used the time of last insulin injection, kurtosis, and skewnessas features and applied linear regression with CGM signal in different time intermissions from data of 10 male subjects[19]. The subjects were made to experience 17 hypoglycemic events by inducing insulin. Non discriminative features were eliminated with SEPCOR and forward selection. The feature amalgamations were used in a Support Vector Machine model and the performance assessed by sample based sensitivity and specificity and event-based sensitivity and number of false-positives.

Researchers have proved that Machine Learning (ML) algorithms could be used for decision support systems and hence works better in diagnosing different diseases [20][21][22]. Based on risk factors, the patients could be classified as Diabetic or Non-Diabetic through Adaboost and Bagging ensemble ML methods in J48 decision tree [23]. Genetic Programming has been applied forworking out Diabetes database from UCI Repository and checking the likelihood for diabetes [24]. ANN has been used to predict diabetes–chronic disease [25]. A system for diabetes diagnosis has been developed with Linear Discriminant Analysis and Morlet Wavelet Support Vector Machine (LDA-MWSVM)[26]. Ant Colony based classification has also been proposed for diabetes diagnosis[27]. A multivariate regression using SVR has been dealt for glucose extrapolation [28]. In multivariate prediction, the problem of subcutaneous glucose estimation in patients with Type 1 Diabetes (T1D) is addressed.

Though we have numerous prediction methodologies, it is difficult to achieve 100% accurate predictions because of vast metabolic biodiversity of people with diabetes and lessfamiliarity on the complex human physiological process of glucose metabolism. One of the ML techniques, K-NN algorithm has been used for a long time as a successful classifier in varied applications. Since KNN has statistical estimation and pattern recognition capabilities, we have applied KNN as regression tool for the prediction of near future BG values. This KNN is a simple algorithm that equips with all the existing data set and foretells the numerical target by similarity features.

III. RESEARCH METHODOLOGY

3.1 Data Set

The proposed method was initially tested with simulated data obtained from Glucosim, a Diabetes simulator developed by Illinois Institute of Technology [29]. This web based simulator gives the 24 hour blood glucose dynamics for every one minute sampling. The second data set has been from the Glucose control project of University of California San Diego. This data set comprises of data set with different sampling frequencies of 10 minutes, 20 minutes etc…[30] . The third data set consists of real time data obtained through CGM devices. These CGM devices track the interstitial fluid glucose dynamics every minute and give the average in every five minutes[31]. For the proposed work all the data have been set uniformly with five minute sampling frequency. The blood glucose dynamics obtained with SMBG readings of a diabetic subject has been shown in figure 1. The BG data from CGMS is given in figure 2.

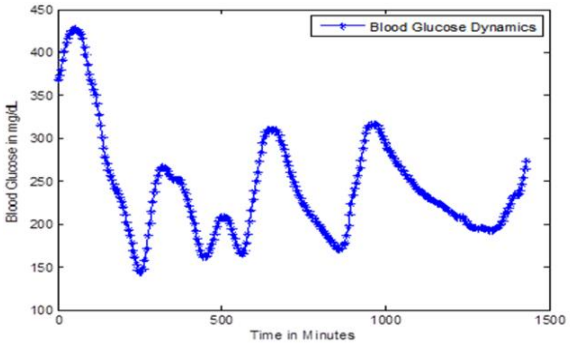

Fig.1 Blood Glucose Dynamics with SMBG

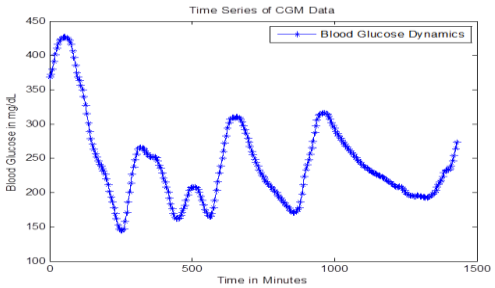

Fig. 2 Blood Glucose Dynamics from CGMS

3.2 KNN Regression

Whenever there is a relationship between dependent variable and one or more independent variables, the regression analysis could be applied with a set of techniques for modeling and analyzing. More predominantly, regression analysis supports to identify how the representative value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are heldstatic. Regression algorithms can be used to estimate the upcoming values in a time-series data.

KNN algorithm is one of the basic but essential algorithms in ML. The KNN algorithm is said to be a non parametric method i.e do not rely on parameterized probability distributions. The input of the algorithm comprises of K- closest training examples in the feature space. Each training example has a numerical value known as property value with the training example. The KNN algorithm employs all the training data sets to predict a property value for the given test sample.

The KNN Algorithm

1. Assign the datasets
2. Assign the initial value of K to the selected number of neighbors
3. For every data sample in the set
   - The distance between the query sample and the current sample has to be computed.
   - The distance and index of the sample is to be added to an ordered collection.
4. From the ordered collection of distances and indices, sort from smallest to largest by the distances.
5. From the sorted collection, the first K entries would be picked up.
6. The selected K entries have to be labeled.
7. The mean of the K labels, which is the regression value, is returned.

The K value depends on bias-variance tradeoff. Smaller values of K gives more flexible fit with low bias but it has a drawback of high variance. Since the prediction is entirely dependent on one observation variable, the variance is high. If K value is high, it provides a smoother less variance output. Since the prediction in the particular region is an average of several points, change in one value has a lesser effect only.

To select the correct value of K, run the algorithm several times with different values of K and choose the K which has minimum prediction error.Lower values of K make the system unstable. Higher values of K, increases the error. Hence an optimum value of K has to be chosen.

IV. EXPERIMENTAL RESULTS

Since the human physiological system is a complex one, the BG dynamics depends on many factors like carbohydrate intake, weight, age, physical activity and insulin action,an expert system which works like a human brain is needed to understand the influence of interdependent factors. KNN algorithm, the most effective ML technique has been customized for tracking BG time series and implemented with Python. The algorithm has been trained with 60% data set and tested with remaining 40% of data. By heuristic method, K value has been chosen as 6 and it was confirmed with many trials of different K values. It was observed that K with value 6 provides minimum Mean Square Error (MSE) values. The algorithm would store all available cases and uses Euclidean distance as a measure of similarity. Then the numerical target is estimated based on this distance function. The distance between any 2 points P1(x1,y1) and P2(x2,y2) is calculated by

$$d = \sqrt{(x2 - x1)^2 + (y2 - y1)^2} \qquad (1)$$

Instead of computing the average of K neighbors, we have taken weighted average of neighbors. Each neighbor is given a weightage of 1/d where d is the distance of the test sample. We modeled the data set as an ordered set {x,y} where x is an ordered set of attribute values i.e the past BG data and y is the BG value to be predicted. The output is an estimated value of y for the given query xt. The experimental procedure has been repeated with BG sample data set and predictions are made on 30 minutes and 60 minutes PH. MSE has been computed between the predicted value and actual BG value. It was observed that KNN regression algorithm is able to predict the BG values better with lesser MSE values in 30 minutes PH than in 60 minutes PH. Sample outputs one each for 30 and 60 minutes PH has been shown below in figure 3 and 4 respectively.

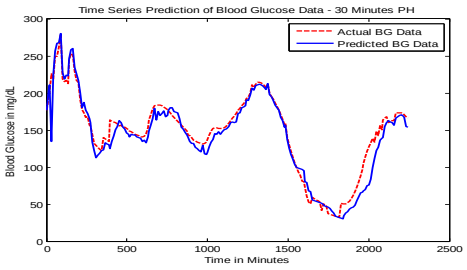

Fig. 3 Prediction Results with KNN Regression for 30 minutes PH

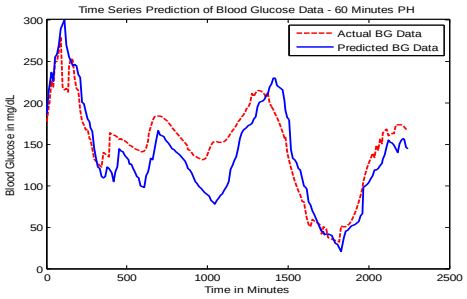

Fig. 4 Prediction Results with KNN Regression for 30 minutes PH

V. CONCLUSION AND FUTURE SCPOE

As this KNN Regression is a non-parametric method, it does not make any assumption on the underlying data distribution. Actually the parameters grow with every training data set. The algorithm is highly unbiased. It is simple and easy to implement. KNN regression applied for the prediction of near future BG values has given promising results which could be further refined with varied data sets and to be tested with diabetes subjects with proper approval procedures. On successful completion of this mission, the proposed work could be used for Artificial Pancreas Project in regulating the BG levels.

REFERENCES

1. https://www.niddk.nih.gov/health-information/diabetes
2. https://www.idf.org/aboutdiabetes/what-is-diabetes/facts-figures.html
3. Bremer T. & Gough D A., "Is blood glucose predictable from previous values? A solicitation for data", Diabetes, 1999, vol. 48, no. 3, pp. 445-451.
4. Palerm C C., Willis J P., Desemone J &Bequette B. W., "Hypoglycemia prediction and detection using optimal estimation", Journal of Diabetes Technology &Therapeutics, 2005, vol. 7, no. 1, pp. 3-14.
5. Mougiakakou S G., Prountzou A., Iliopoulou D., Vazeou A., Bartsocas C S., & Nikita K S., "Neural Network based Glucose-Insulin Metabolism Models for Children with Type 1 Diabetes", Proceedings of IEEE Conf on Eng. Med. Biol.,2006, pp. 3545- 3548.
6. Palerm C C. &Bequette B W, "Hypoglycemia detection and prediction using continuous glucose monitoring-a study on hypoglycemic clamp data", Journal of Diabetes Science and Technology, 2007, vol. 1, no. 5, pp. 624-629.
7. Sparacino G., Zanderigo F., Corazza S., Maran A., Facchinetti A., &Cobelli C., "Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time series", IEEE Transactions on Biomedical Engineering, 2007, vol. 54, no. 5, pp. 931-937.
8. Reifman J., Rajaraman S., Gribok A., & Ward W K., "Predictive monitoring for improved management of glucose levels", Journal of Diabetes Science and Technology, 2007, vol. 1, no. 4, pp. 478-486.
9. Quchani S A., &Tahami E., " Comparison of MLP and Elman Neural Network for blood Glucose Level Prediction in type 1 Diabetics", IFBME Proceedings, 2007, vol. 15, pp. 54-58.
10. Pappada S M., Camero B D., &Rosman P M., "Development of a neural network for prediction of glucose concentration in type 1 diabetes patients", Journal of Diabetes Science and Technology, 2008, vol. 2, no. 5, pp.792-801.
11. Eskaf K., Badawi O., &Ritchings T., "Predicting blood glucose levels in diabetics using feature extraction and Artificial Neural Networks", Proceedings of the 3'rd International Conference on Information and Communication Technologies: From Theory to Applications, 2008, pp. 1-6.
12. Gani A., Gribok A V., Rajaraman S., Ward W K., &Reifman J., "Predicting subcutaneous glucose concentration in humans: data-driven glucose modeling", IEEE Transactions on Biomedical Engineering, 2009, vol. 56, no. 2, pp. 246-254.
13. Valletta J J., Chipperfield A J., & Byrne C D., "Gaussian Process modelling of bloodglucose response to free-living physical activity data in people with type 1 diabetes", Proceedings of the 31st Annual International Conference of the IEEE EMBS, 2009, pp. 4913- 4916.
14. Perez-Gandia C., Facchinetti A., Sparacino G., Cobelli., Gomez E J., Rigla M., de Leiva A & Hernando M E., "Artificial neural network algorithm for on-line glucose prediction from continuous glucose monitoring", Journal of Diabetes Technology & Therapeutics,2010, vol. 12, no. 1, pp. 81-88.
15. Rollins D K., Bhandari N., Kleinedler J., Kotz K., Strohbehn A., Oland L., Murphy M., Andre D., Vyas N., Welk G., &Franke W E., "Free-living inferential modeling of blood glucose level using only noninvasive inputs", Journal of Process Control, 2010, vol. 20, no. 1, pp. 95-107.
16. Aibinu A M., Salami M J E., &AShafie A., "Blood Glucose Level Prediction Using Intelligent Based Modeling Techniques", IEEE Region 10 Conference, 2010, pp.1734 - 1737.
17. Shanthi S., & Kumar D., "Prediction of blood glucose concentration ahead of time with feature based neural network", Malaysian Journal of Computer Science., 2012, vol.23, no.3, pp.136-148.
18. Zecchin C., Facchinetti A., Sparacino G., Nicolao G D., &Cobelli C., "Neural Network Incorporating Meal Information Improves Accuracy of Short-TimePrediction of Glucose Concentration", IEEE Transactions on Biomedical Engineering, 2012, vol. 59, no. 6, pp.1550-1560.
19. Jensen M H., Christensen T F., Tarnow L., Seto E., Johansen M D., &Hejlesen O K., "Hypoglycemia Detection from Continuous Glucose Monitoring Data of Subjects with Type 1 Diabetes", Diabetes Technology & Therapeutics, 2013, vol. 15, no.7.

20. Aishwarya R., Gayathri P., Jaisankar N., "A Method for Classification Using Machine Learning Technique for Diabetes", International Journal of Engineering and Technology (IJET) , 2013, vol.5, pp. 2903-2908.
21. Kavakiotis I., Tsave O., Salifoglou A., Maglaveras N., Vlahavas I., &Chouvarda I., "Machine Learning and Data Mining Methods in Diabetes Research", Computational and Structural Biotechnology Journal, 2017, vol.15, pp.104-116.
22. DhomseKanchan B., M.K.M., "Study of Machine Learning Algorithms for Special Disease Prediction using Principal of Component Analysis", IEEE International Conference on Global Trends in Signal Processing, Information Computing and Communication, 2016, pp.5-10.
23. Perveen S., Shahbaz M., Guergachi A., Keshavjee K., "Performance Analysis of Data Mining Classification Techniques to Predict Diabetes", Procedia Computer Science, 2016, vol. 82, 115-121.
24. Bamnote M.P., G.R., "Design of Classifier for Detection of Diabetes Mellitus Using Genetic Programming", Advances in Intelligent Systems and Computing, 2014, 1, 763-770.
25. Tarik A. P., Rashid S.M.A., Abdullah R.M., "Abstract.: An Intelligent Approach for Diabetes Classification, Prediction and Description", Advances in Intelligent Systems and Computing, 2016, vol. 424, pp. 323-335.
26. DuyguÇalisir, EsinDogantekin, "An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier", Expert Sysems Applications, 2011, vol. 38, issue 7, pp. 8311-8315.
27. Ganji M F., Abadeh M S., "A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis", Expert Sysems Applications, 2011, vol. 38, issue 12, pp. 14650-14659.
28. Eleni I., GeorgaVasilios C., Protopappas, Ardigò D., Michela Marina, IvanaZavaroni, Demosthenes Polyzos, Dimitrios I. Fotiadis, "Multivariate Prediction of Subcutaneous Glucose Concentration in Type 1 Diabetes Patients Based on Support Vector Regression", IEEE J. Biomedical and Health Informatics, 2013, vol. 17, issue 1, pp. 71-81.
29. <http:// 216.47.139.198/glucosim/index.html>, 2006.
30. <http:// glucosecontrol.ucsd.edu>, 2008.
31. Rollins D K., Bhandari N., Kleinedler J., Kotz K., Strohbehn A., Oland L., Murphy M., Andre D., Vyas N., Welk G., Franke W.E.. "Free-living inferential modeling of blood glucose level using only noninvasive inputs", Journal of Process Control, 2010, vol. 20, issue 1, pp. 95-107.