Pregnancy Period Diabetic and Blood Pressure Predictive Analysis using HCNN-LSTM

T. Papitha Christobel, A. Sasi Kumar

Abstract*: Diabetes has transformed into the worldwide diseases and can occur for all age groups irrespective of their gender. Unlike other diseases, Diabetes needs continuous monitoring as it leads to much adverse effect on functioning of human body. Especially, the diabetes that occurs in female during the pregnancy had its impact over the mother along with their infant before its birth. Many studies showed early prediction can prevent and delimit the challenges that were posed by diabetes among pregnant women. Several health care prediction models often suffer from inconsistencies in data and feature selection that reduce the prediction performance. In the present work, we had proposed the novel Health Care Neural Network-Long Short Term Memory (HCNN-LSTM) to predict the Pregnancy Period Diabetic and Blood Pressure. The Pima Indian diabetes dataset was employed construct the proposed prediction model to predict the patient as diabetic and non-diabetic. For the purpose of comparison, the decision tree, random forest and Navies' Bayes algorithm are implemented for classification. From the analysis, it was evident that the proposed HCNN-LSTM showed optimum values on performance metrics than the other classifiers. The proposed work can be expanded considering several features of diabetic prediction in future.

Keywords: Diabetes, Blood Pressure, pregnancy, prediction model, Proposed HCNN-LSTM.

I. INTRODUCTION

The Diabetes is a chronic disease with the potential to cause a worldwide health care crisis. According to International Diabetes Federation 382 million people are living with diabetes across the whole world and it is expected to be doubled By 2035[1].Diabetes is a chronic, complex illness demanding incessant medical care with multifactorial strategies for risk-reduction further than glycemic control [2].Early prediction of diabetes is quite challenging task for medical practitioners as it depend on several developing factors like genetic susceptibility, body weight, food habit and sedentary lifestyle [3]. Diabetes affects human organs such as kidney, eye, heart, nerves, foot [4].

Machine Learning (ML) is a sub area of artificial intelligence which concentrates on the advancement of systems thereby empowering the software applications to get into a self-learning state without being programmed explicitly [5-6]. ML supports the system to identify and understand the input data, so that it can make decisions and predictions based on it [7].ML is about scheming algorithms that permit a computer system to learn. Learning is not essentially includes awareness but learning is defining the statistical constancies or other data patterns [8].The machine learning effect has also observed generally across a wide-range of industries apprehensive with data-intensive problems, like services for consumer, the faults diagnosis is complex [9].

Predictive modeling is a commonly used statistical technique to predict future behavior. Solutions from predictive modeling are in the data-mining form that performs by analyzing current and historical data and making a model to aid in predicting future results. It is the outcome of combining mathematics and data, in which a mapping function was generated between a input fields of data set and a target or response variable [10].

Among the different diabetes, Gestational diabetes mellitus (GDM) is the nightmare of epidemiologists. The condition is well-defined as intolerance of carbohydrate resulting in variable severity hyperglycemia with onset or first acknowledgment during pregnancy [11].Low and high birth weight are likely risk factors for GDM because of their association with insulin resistance. It is assumed that the fetus recompenses for under nourishment in the womb by epigenetically changing the genes expression that included in energy utilization, fat storage, and regulation of appetite [12].

Hence the prediction of GDM is very important at the early stages. However the present prediction models are subjective to many problems like missing data, improper data retrieval and more time to classification of data that affect the prediction accuracy to greater extent. The following contributions are carried out in our work to predict the diabetes during pregnancy along with the blood condition level as.

- The region of improper data is identified in the Electronic Health Record (EHR) and implemented the latent factor model to restructure the misplaced data to form a complete data.
- A novel HCNN-LSTM based multimodal risk prediction (Health Care Neural Network-Long Short Term Memory) algorithm is proposed for EHR data.
- The EHR risk model is developed and through the experiment, the performance of Proposed HCNN-LSTM is found better than state-of-art methods.

Revised Manuscript Received on February 05, 2020.
***** Correspondence Author

T. Papitha Christobel *, PhD Research Scholar, Department of Information Technology, School of Computing Sciences, Vels Institute of Science, Technology and Advanced Studies (VISTAS), Chennai, India. Email: babitha.shaghi@gmail.com

Dr.A. Sasi Kumar, Professor, Department of Information Technology, School of Computing Sciences, Vels Institute of Science,Technology and Advanced Studies(VISTAS), Chennai, India. Email: askmca@yahoo.com

The current work on predicting the diabetes and hypertension during pregnancy is structured in the following order: after the introduction, the related works section with previous researches is explained. Section 3 involves the material and methods that are involved in the work; the next section provides the proposed prediction model. Section 5 encloses the results and its relative discussions and the final section involves the conclusion and future scope.

II. RELATED WORK

Decision support system is recommended through the algorithm of Ada Boost with Decision Stump classifier as a base. Furthermore, Naive Bayes, Support Vector Machine, and Decision Tree have implemented as a base classifiers for accuracy confirmation. The accuracy of Ada Boost with Decision Stump classifier is 80.72%, which is remarkable more than other classifiers[13].Pima Indian dataset was analyzed through different classification techniques like Zero R, Naïve Bayes, random forest, J48, logistic regression, and MLP. Detecting diabetes through WEKA was evaluated on accuracy and performance and showed that MLP is better [14]. The framework with two phases was proposed in which the primary stage highlighted the predicting factors through Generalized Discriminant Analysis followed by the LS-SVM over the diabetes dataset. LS-SVM attained 78.21% grouping precision with 10-overlap whereas the suggested GDA–LS-SVM framework got 82.05% accuracy with 10-crease across approval [15]. A predicting model was proposed to classify treatment plans of type 2 diabetic with three groups including diet, insulin and medication. The JABER ABN ABU ALIZ clinic center dataset was used for generating the model that encloses 318 medical archives. The WEKA tool was used for model generation using J48 classifier and achieved a 70.8%accuracy [16].The two models of neural network one the multilayer and another probabilistic neural network were employed in predicting diabetes. The Pima Indian diabetes dataset with768 samples were segregated as 572 data for training and 192 for testing. The proposed approaches were verified to have better prediction when compared with other preceding methods [17].Another prediction model was developed to predict whether a person progresses diabetes or not through PIMA diabetes dataset. In the suggested method controlled binning technique is applied first then multiple regressions were used to increase the model accuracy. After integrating all methods an accuracy of 77.85% was attained [18].

From the extensive studies over the previous models in predicting the diabetes through the different machine learning algorithm showed that the accuracy, the most important performance parameter has it values in the range of 70-85%. Since the prediction of diabetes among the pregnant women is vital, the accuracy must be of very high order. Thus the model proposed should ensure the accuracy and effectiveness in predicting the diabetes.

III. RESEARCH METHODOLOGY

The supervised machine learning algorithms are those algorithms which needs external assistance. [19]. Common supervised ML algorithms that are implemented in the study are decision tree and the Navies Bayes classifier. The unsupervised learning algorithms learn few features from the data. When a data is presented, it practices the formerly learned features to identify the data class. It is mostly used for feature and clustering reduction. The random forest, an ensemble learning algorithm was employed to train the EHR dataset. Neural networks are a technique form a mathematical representation of inter connected neurons in systems similar to that of brain. The functions are mathematically expressed as nodes, and are interrelated to produce a complex inputs and outputs web based on the sequence; functions types used and connectivity among the neural nodes of networks to dictate the effectiveness in various machine learning problem forms [20]. In the present work we propose the novel HCNN-LSTM (Health Care Neural Network-Long Short Term Memory based multimodal risk prediction) is used to classify the EHR data in the effective manner.

A. Decision Tree

A decision tree is a predictor, h : X → Y, that predicts the feature label associated with an instance x by traveling from a root node of a tree to a leaf. At each node on the root-to-leaf path, the successor feature is chosen on the basis of a splitting of the input space. Usually, the splitting is based on one of the features of x or on a predefined set of splitting rules. Each leaf contains a specific feature label [21].

Algorithm

```
Step 1: procedure (DT,F ) . DT-Training data set and
          F− Subset of Feature
Step 2: if DT Instances label are 1 then Leaf = 1
Step 3: elseDT Instances label are 0 Leaf = 0
Step 4: end
Step 5: if F == ∅ then Leaf = Majority in DT
Step 6: elseb = argmax
Step 7: end
Step 8: if All DT instances have the identical label
          then
Step 9: Leaf = Majority in DT
Step 10: else T1 is the tree returned by ID3
Step 11: T2 is the tree returned by ID3
Step 12: end
```

B. Navies Bayes

The Naive Bayes classifier is a classical demonstration of how generative assumptions and parameter estimations simplify the learning process. Consider the problem of predicting a feature y ∈ {0, 1} on the basis of a vector of features x = (x1, . . . ,xd), where we assume that each xi is in {0, 1}. The probability function P[Y = y|X = x] is estimated that corresponds to P[Y = 1|X = x] for a certain value of x ∈ {0, 1} d. This implies that the instances grows exponentially with the number of features [21].

```
Algorithm

Step 1: procedure (DT, H0, C0, TEM, TAdd ) .
Step 2: C = C0
Step 3: Add new C mixture components to F
Step 4: Remove the initialization instances form DT
Step 5: Assign Instances fractionally in DT
Step 6: Adjust F parameters to fractional assignment
          maximization
Step 7: if logP(H0|F) is the uppermost save F in Fbest
          then
Step 8: UnitillogP(H0|F) fails to develop by radio TEM
over final iteration
Step 9: C = 2×C
Step 10: UnitillogP(H0|F) fails to develop by radio
          TAdd over final iteration
Step 11: Perform M −step and E −step twice more on
             Fbest using instances from both H0 and DT
Step 12: end
```

C. Random Forests

A RF is a classifier containing a decision trees collection in which each tree is built by applying the algorithm with a training set and an additional random vector which is a sampled i.i.d. from some distribution. The prediction of the random forest is obtained by a majority vote over the predictions of the individual trees.

```
Algorithm
Step 1: procedure (DT, F, N).
Step 2: A = ∅
Step 3: for i = 1 to N do
Step 4: D(i) T is bootstrap instance from DT
Step 5: ai is Randomized Tree Learn (D (i) T ,F)
Step 6: A ∈ [ai]
Step 7: If not return A
Step 8: end
Step 9: Begin function Random forest (DTF)
Step 10: Each node f, belong to F subset
Step 11: Segment best feature in F
Step 12: Return learned tree
```

D. Proposed HCNN-LSTM Algorithm

Recurrent neural networks with Long Short-Term Memory have developed as a scalable and effective model for learning several problems associated to sequential data. The RNN are a sub-class of neural networks that were constructed to generate the long-range inherent correlation among data samples. Though the normal NN do not detail the temporal input data order, the RNN eludes this issue by having the time built notion into it. Related to other NN architectures, the RNNs have a hidden layer and update its hidden layer after each time-step process in the input. This confirms that the input sequence temporal structure is valued. The existing RNN may subject to local minimal solutions during the iterations through the layers.

In the current work the novel HCNN-LSTM algorithm with RNN was implemented through one of the more popular activation functions for back-propagation networks is the sigmoid with the novel exponential form, a real function s: R → (0, 1) defined by the expression.

$$s(b) = \frac{1}{1 + \text{expv}} \quad \ldots\ldots(1)$$

The exponential component in the general sigmoid function (expv) is transformed into the cubical interpolation through the equation (2).

$$\text{expv} = \exp((\text{double})\ {-x}); \quad \ldots\ldots(2)$$

The novel sigmoid function for the HCNN-LSTM model is given in equation (3)

$$s(b)= \frac{(4.0 * expv)}{((1 + expv) * (1 + expv)}; \quad \ldots\ldots(3)$$

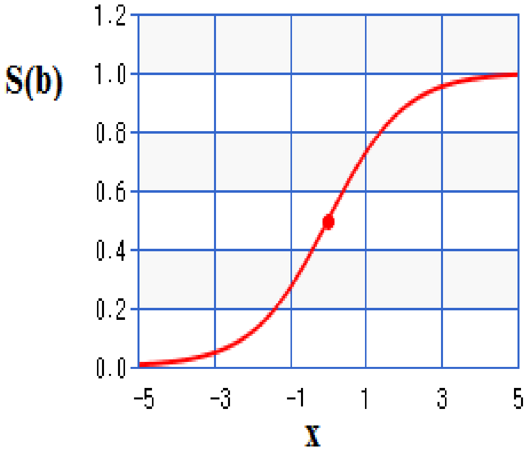

Figure 1 Proposed sigmoid function of HCNN-LSTM

```
Algorithm for Proposed HCNN back propagation
Step 1: procedure()
Step 2: A constant A is employed at the output unit and
teh backwards propogation of network starts
Step 3: Arriving nodal data is added and the outcome
is multiplied with the stored value in the left side of
unit
Step 4: The outcome is diffused to the left side unit
Step 5: The outcome composed at the input unit is the
network function derivative corresponding to x
Step 6: Set RNN parameter, F and shuffle the dataset
DT
Step 7: Set i = 0
Step 8: For each feature belong to A, update F
```

E. Proposed Predictive Model Framework for Diabetic Healthcare

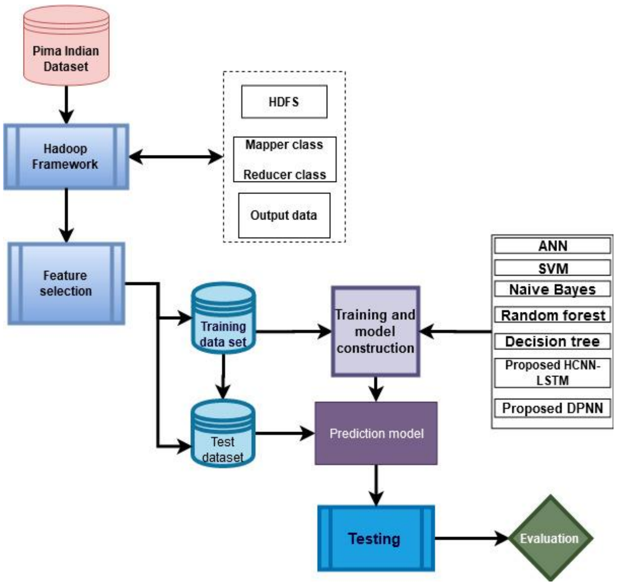

Figure 2: Proposed Predictive model framework

The proposed framework for GDM prediction initially involves the partition of the Pima Indian dataset which as nearly 3.5 GB instances in it. The data set is very large and it is initially subjected to the Hadoop framework that stores the dataset into the HDFS in structured form. The map reduce algorithm is employed to classify the data into mapper class and reducer class. The mapper class tokenizes the data and maps it with the dataset and sorts it. The reducer class transforms the in differentiable data into structured form through merging conditions. The process of feature selection followed through the Boruta wrapper algorithm which was efficient in determining the significant and insignificant feature from dataset. The raw data were then trained over all the four ML algorithms. Finally the test data set were provided to the trained ML algorithm to predict the GDM and blood pressure. The performance metrics are evaluated through the confusion matrix.

IV. RESULT AND DISCUSSION

A. Experimental setup

The results are experimented on CPU with 2.5 GHz intel core i5 processor, 4GB RAM per node and are connected with 1 GB Ethernet; all clusters are configured on ubuntu 11.10 Linux, Hadoop 2.7.2 and source code is compiled under JDK 1.8.0-65. Each slave contains one task tracker and data node. JobTracker and NameNode are executed on master node.

B. Dataset

The dataset with nearly 768 instances on female patient was obtained from the Pima Indian dataset after processing in Hadoop framework. The features to predict the GDM over the dataset was given in table 1. The dataset may have some form of inconsistencies in them; hence the obtained dataset was pre-processed with the latent factor model before loading for analysis. The analysis over the dataset was accomplished through the different performance metrics evaluation like accuracy, specificity, sensitivity, F1 score and area under ROC curve.

Table 1: Attribute for prediction

Attribute No.	Attribute	unit	Variable Type	Range
A1	Age	years	integer	21–81
A2	Pregnancy	number of pregnancy	Integer	0-10
A3	Body mass index	kg/m2	Real	0-67.1
A4	Plasma Glucose	mg/dL	Real	0-199
A5	Diastolic Blood Pressure	mm Hg	Real	0-122
A6	Serum Insulin	mu U/ml	Real	0-846
A7	Triceps skin fold	mm	Real	0-99
A8	Diabetes Pedigree	no unit	Real	0.078-2.42

V. RESULT

The adverse effect of the diabetes during pregnancy can affect the both the mother and the fetus and hence its early prediction prevent those affect. The Pima Indian diabetes dataset was successfully analyzed over the four ML algorithm. For initial learning process, 30% of data were employed and the remaining 70% were used for the purpose of testing. All the Ml techniques predicted the occurrence of diabetes and their performance were discussed in this section based on table 2. In the current analysis the accuracy, sensitivity, specificity along with F1 score and area under curve of ROC are used as the performance metrics.

Table 2: Performance of different Prediction models

S.No	Prediction models	Accuracy	Precision	Recall	F1 score	AUC in ROC
1	Decision tree [22]	0.738	0.735	0.738	0.736	0.751
2	Random Forest [23]	0.747	0.5075	0.694	0.5875	0.806
3	Navies Bayes [22]	0.763	0.759	0.763	0.76	0.819
4	HCNN-LSTM	1	1	1	1	1

The accuracy of the prediction model was the prime parameter as it establishes the correctness of the ML techniques in predicting the occurrence of diabetes among the women during pregnancy. From table 2, it was observed that the proposed novel HCNN-LSTM showed optimum accuracy of 100%, which were distantly ahead of other prediction models. From table 2, the precision of Random forest is very low with the value of 0.5075 with other two prediction models of Decision tress and Navies Bayes attained the value of 0.735 and 0.759 respectively.

Our proposed HCNN-LSTM showed the optimum level of recall with 1.Similar to precision, recall which the measure of negative diabetes cases attained the maximum value in our proposed model and evidently more than the other three techniques as shown in table 2. F1 score is generally the reflection of the precision and recall was found to be in range of 0.76 and 0.587 for Navies Bayes and random forest respectively. Our model accomplished the value of 1, an optimal score whereas the Decision tree achieved the value of 0.736.

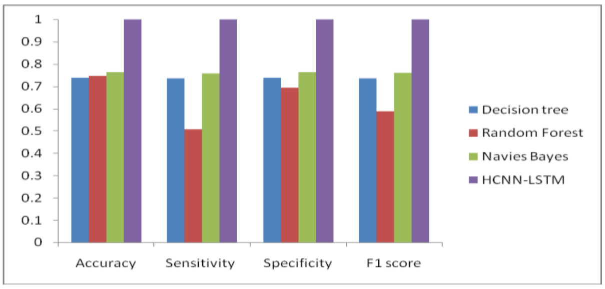

Figure 3: comparison of different ML algorithm in predicting diabetes during pregnancy.

The curve of Receiver operating characteristic (ROC) provide the method to visualize the performance of the prediction model through AUC value. The AUC value of the prediction model is 0.806, 0.751, 0.819 and 1 for random forest, decision tree, Navies Bayes and our novel HCNN-LSTM algorithms respectively. The comparison graphs over various performance metrics were given in figure 3 and 4.

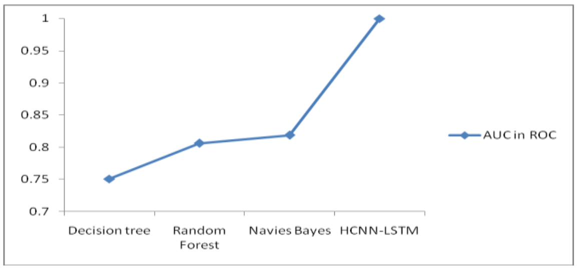

Figure 4: Comparison of AUC in ROC curve for different prediction model

In overall analysis, it was evident that the proposed HCNN-LSTM algorithm was very efficient in predicting the diabetes during pregnancy. Even though the accuracy may not reflect the effectiveness of algorithm in many cases, the other performance metrics also exhibited the appropriateness of the proposed algorithm to be employed in predicting both blood pressure and diabetes among the female patient during pregnancy.

VI. CONCLUSION

The novel HCNN-LSTM prediction model was generated by integrating CNN and LSTM was proposed to predict the diabetes during pregnancy and blood pressure over the Pima Indian diabetes dataset. Additionally, the prediction model with decision tree, random forest and Navies Bayes were developed as the prediction model. The dataset was preprocessed with the latent factor model and segmented into training and testing datasets. The trained algorithms were tested over the various performance metrics. From the testing outcome, it was found that the proposed model provided the optimum outcome on accuracy, F1 score and other parameters and deliberated as the best model for predicting diabetes during pregnancy. The feature for predicting the diabetes can be enhanced in future with different dataset.

REFERENCES

1. Renuka, M., &Shyla, J. (2016). Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus. International Journal of Applied Engineering Research, 11(1), 727–730.
2. Introduction: Standards of Medical Care in Diabetes—2018. (2017). Diabetes Care, 41(Supplement 1), S1–S2. https://doi.org/10.2337/dc18-sint01
3. Kaur, H., &Kumari, V. (2018). Predictive modelling and analytics for diabetes using a machine learning approach. Applied Computing and Informatics.
4. Komi, M., Jun Li, YongxinZhai, &Xianguo Zhang. (2017). Application of data mining methods in diabetes prediction. 2017 2nd International Conference on Image, Vision and Computing (ICIVC). doi:10.1109/icivc.2017.7984706
5. Lidong Wang, Cheryl Ann Alexander, "Machine Learning in Big Data", International Journal of Mathematical, Engineering and Management Sciences Vol. 1, No. 2, 52–61, 2016
6. Dharmarajan, K., and M. A. Dorairangaswamy. "Discovering User Pattern Analysis from Web Log Data using Weblog Expert." Indian Journal of Science and Technology 9.42 (2016).
7. Qiu, J., Wu, Q., Ding, G., Xu, Y., &Feng, S. (2016). A survey of machine learning for big data processing. EURASIP Journal on Advances in Signal Processing, 2016(1), 67.
8. Ayodele, T. O. (2010). Types of machine learning algorithms. In New advances in machine learning. IntechOpen.
9. Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.
10. Kalechofsky, H. (2016). A simple framework for building predictive models. A Little Data Science Business Guide, 1-18.
11. Kahn, Richard. "Follow-up report on the diagnosis of diabetes mellitus: the expert committee on the diagnosis and classifications of diabetes mellitus." Diabetes care 26, no. 11 (2003): 3160.
12. Plows, J., Stanley, J., Baker, P., Reynolds, C., & Vickers, M. (2018). The pathophysiology of gestational diabetes mellitus. International journal of molecular sciences, 19(11), 3342.
13. Vijayan, V. V., & Anjali, C. (2015). Prediction and diagnosis of diabetes mellitus — A machine learning approach. 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS).doi:10.1109/raics.2015.7488400
14. Hina, S., Shaikh, A., &Sattar, S. A. (2017). Analyzing diabetes datasets using data mining. Journal of Basic and Applied Sciences, 13, 466-471.
15. Polat, K., Güneş, S., &Arslan, A. (2008). A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine. Expert systems with applications, 34(1), 482-487.
16. Ahmed, T. M. (2016). Developing a predicted model for diabetes type 2 treatment plans by using data mining. Journal of Theoretical and Applied Information Technology, 90(2), 181.
17. Temurtas, H., Yumusak, N., &Temurtas, F. (2009). A comparative study on diabetes disease diagnosis using neural networks. Expert Systems with applications, 36(4), 8610-8615.
18. Jakhmola, S., &Pradhan, T. (2015, August). A computational approach of data smoothening and prediction of diabetes dataset. In Proceedings of the Third International Symposium on Women in Computing and Informatics (pp. 744-748). ACM.
19. S.B. Kotsiantis, "Supervised Machine Learning: A Review of Classification Techniques", Informatica 31 (2007) 249-268
20. Maheshwari, A., Davendralingam, N., &DeLaurentis, D. A. (2018). A Comparative Study of Machine Learning Techniques for Aviation Applications. In 2018 Aviation Technology, Integration, and Operations Conference (p. 3980).
21. Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.
22. Sisodia, D., &Sisodia, D. S. (2018). Prediction of Diabetes using Classification Algorithms. Procedia Computer Science, 132, 1578–1585. doi:10.1016/j.procs.2018.05.122
23. MahboobAlam, T., Iqbal, M. A., Ali, Y., Wahab, A., Ijaz, S., ImtiazBaig, T., … Abbas, Z. (2019). A model for early prediction of diabetes. Informatics in Medicine Unlocked, 16, 100204. doi:10.1016/j.imu.2019.100204