# Type-2 Diabetes Mellitus Diagnosis from Time Series Clinical Data using Deep Learning Models

Zakhriya Alhassan[1,2], A. Stephen McGough[3], Riyad Alshammari[4], Tahani Daghstani[4], David Budgen[1], and Noura Al Moubayed[1]

[1] Computer Science, Durham University, Durham, UK
{zakhriya.n.alhassan,noura.al-moubayed,david.budgen}@durham.ac.uk
[2] Computing and Information Technology, University of Jeddah, Jeddah, KSA
[3] School of Computing, Newcastle University, Newcastle upon Tyne, UK
{stephen.mcgough}@newcastle.ac.uk
[4] King Saud Bin Abdulaziz University for Health Sciences, Riyadh, KSA
{alshammariri,daghistanita}@ngha.med.sa

**Abstract.** Clinical data is usually observed and recorded at irregular intervals and includes: evaluations, treatments, vital sign and lab test results. These provide an invaluable source of information to help diagnose and understand medical conditions. In this work, we introduce the largest patient records dataset in diabetes research: King Abdullah International Research Centre Diabetes (KAIMRCD) which includes over 14k patient data. KAIMRCD contains detailed information about the patient's visit and have been labelled against T2DM by clinicians. The data is processed as time series and then investigated using temporal predictive Deep Learning models with the goal of diagnosing Type 2 Diabetes Mellitus (T2DM). Long Short-Term Memory (LSTM) and Gated-Recurrent Unit (GRU) are trained on KAIMRCD and are demonstrated here to outperform classical machine learning approaches in the literature with over 97% accuracy.

**Keywords:** Type 2 Diabetes Mellitus, Deep Learning, Long Short-Term Memory, Gated-Recurrent Unit, King Abdullah International Research Centre Diabetes.

## 1 Introduction

Diabetes is an increasingly growing medical condition worldwide. The estimated number of diabetic patients globally was 415 million in 2015 and is expected to affect one person in 10 by 2040 [6]. The number of people who are borderline diabetic is rapidly increasing. The latest estimates indicate that 35.3% of the adults in the UK are pre-diabetic [17]. Patients suffering from diabetes develop serious and complicated health problems to vital organs such as the kidneys, eyes, as well as the heart. By the end of 2015, there were 5 million deaths caused by diabetes worldwide[6].

There are three types of diabetes: I) Type 1 Diabetes occurs when the body's defence system attacks the pancreas cells, causing it to stop producing the needed

insulin. II) Type 2 Diabetes occurs when the body fails to respond to the insulin produced. III) Gestational Diabetes which happens when hormonal changes during pregnancy make the body resistant to the insulin [18].

Type 2 Diabetes Mellitus (T2DM) is the most common form accounting for 91% to 95% of all cases [6]. It is the main contributor to causes of death from diabetes and its associated cost. Furthermore, T2DM is difficult to diagnose because it does not have clear clinical symptoms. It often stays undetected for a long time as a result of the slow development of its symptoms [1]. Thus, an early diagnosis of T2DM can assist with delaying any long-term complications.

In many hospital systems, patient data, such as vital signs and lab tests, are routinely collected and stored with an associated time stamp which we will refer to as "Clinical Time Series Data". Patient clinical data is usually carried out at irregular times and stored in the hospital record systems. The frequency of taking these measurements is different for each patient, based on the physician's decisions. In addition, patients differ in their visit patterns (e.g., in-patient or emergency visits), therefore the stay length for each patient varies from few hours to days, weeks or even months.

In this study, we use King Abdullah International Research Centre Diabetes (KAIMRCD) dataset. KAIMRCD is a unique dataset of 14,609 patient visits which have been clinically tested against T2DM. It contains the personal details of every patient such as age and gender along with the vital signs and lab test results for every visit. The availability of such large dataset makes it possible to train advance machine learning techniques, e.g. deep learning models to predict T2DM.

The use of Recurrent Neural Networks (RNNs) has recently redefined the standards for several research areas involiving sequential data such as speech recognition, natural language processing and machine translation [8] [11]. Despite their success, RNNs are not usually fit for problems with long temporal dependencies due to the exploding gradients problem [7]. Long Short-Term Memory (LSTM) [9] and Gated-Recurrent Unit (GRU) [5] [3], were specifically developed to model problems that involve both long and short temporal dependencies. Thus, LSTM and GRU have demonstrated the ability to model complex clinical data in variety of medical applications such as diseases diagnosis [13] [14].

The main contributions of this paper are: I) Introducing the largest diabetes patients time series data. II) Applying temporal deep learning models: LSTM and GRU to predict chronic disease, T2DM. III) Integrating non-sequential risk factors into the time series data such as gender and age. IV) Investigating the effect of input size on the performance of the built LSTM and GRU models.

## 2 Related Work

Machine learning has been successfully applied to clinical data and have been demonstrated in tasks such as the prediction of patient progress and length of stay. Disease diagnosis prediction using time series data is a growing field of research for machine learning. Several neural network models have been applied

**Table 1.** Neural Network Models for T2DM Diagnosis

| **Study** | **Dataset** | **No of Features** | **No of Records** | **Data Availability** | **Accuracy** |
|---|---|---|---|---|---|
| Venkatesan et al. [21] | Private Date | 9 | 1800 | No | 91.3% |
| Meng et al. [15] | Private Date | 12 | 1487 | No | 72.59% |
| Temurtas et al. [20] | PPID | 8 | 768 | Yes | 82.37% |
| Motka et al. [16] | | | | | 90.49% |
| Karegowda et al. [10] | | | | | 84.71% |
| Polat et al. [19] | | | | | 89.47% |
| GRU | KAIMRCD | 30 | 14,609 | Upon request | 97.3% |

for T2DM diagnosis prediction, summarised in Table 1. Multi-Layer Perceptron models were applied on various datasets [21, 15, 20]. Motka et al. [16] and Polat et al. [19] used Artificial Neural Fuzzy Inference Systems (ANFIS). Genetic Algorithms (GA) with Back-propagation Neural Network were also applied [10]. It is important to note that the majority of these models were applied to the Pima Indian Diabetes Data (PIDD) [12] and used small datasets that had no temporal information with a small number of features.

To the best of our knowledge, there are no studies that looked at the T2DM diagnosis from a time-series perspective. We are the first to apply deep learning, LSTM and GRU in particular, for classification in T2DM diagnosis as a time series (vital signs or lab test results) data. There are a few recent studies that are related to our work. These studies used RNN models together with general clinical time series datasets for multi-disease (T2DM was not among them) diagnosis classification [13, 14]. However, the time series datasets used in these studies were not specifically collected for the purpose of diabetes diagnosis.

Lipton et al. [13] proposed the first model that applied LSTM on a clinical dataset. The authors used LSTM on a Children's Intensive Care Unit (ICU) dataset to predict multiple diseases diagnosis (such as Asthma, Hypertension and Anemia) using 13 lab test results. The LSTM model was built to classify 128 diseases with competitive accuracy. Another study [4], applied GRU on larger and longitudinal patient data extracted from the general patients clinical records. Similar to Lipton's study, the aim of the study was mainly to predict disease diagnosis. However, The features used in this study are different in type than the ones used in Lipton's study. The authors did not make use of patient's observation records (vital signs or lab test results). Instead, they used previous patient's diagnoses as input to predict future diseases. However, it was not clear how many and what diseases have been examined for evaluating the model.

Both LSTM models as applied in [13] [14], and GRU model as applied in [4], have shown promising results with regard to multi-disease diagnosis. The number of samples for each disease, on which the models were trained, was not reported in either studies.

The work is motivated by the temporal nature of clinical data which would potentially be better modelled by a model that directly models sequential/temporal

data similar to GRU/LSTM. This is particularly relevant given the size of our dataset, KAIMRCD, which considerably larger than any reported in the literature for the diagnosis of T2DM. Our models incorporate not only the clinical vital signs and lab test results, but also non-sequential data such as age and gender, which are important risk factors for T2DM [6].

## 3 Dataset

King Abdullah International Medical Research Center (KAIMRC) is one of the leading institutions in health research in the Middle East. The KAIMRCD[5] dataset was collected by Ministry of National Guard Health Affairs (NGHA) from the main National Guard Hospitals located in three populated regions[6]. It is part of the hospital care service procedures to clinically diagnose visitors against T2DM. The collected data contains records of clinical diagnosis of T2DM from the full visits history of 14,609 patient visits.

KAIMRCD dataset was collected over the period between 2010 and 2015. It contains 41 million time-stamped results for lab tests, such as Blood Urea Nitrogin (BUN), cholesterol (Chol) and Mean Corpuscular Hemoglobin (MCH). It also holds time-stamped data about patient vital signs such as Body Mass Index (BMI) and Hypertension. Other important features are also included, such as visit type (inpatient, outpatient or emergency), discharge type (home, referred to another hospital, patient died), gender, patient's age at the visit, service type (e.g. Cardiology, Neurology, Endocrinology) and stay length [7]. The data is imbalanced with 62% of the patients are diagnosed with diabetes, hence F1 measure is used as an evaluation metric rather than accuracy. Figure 1 shows the distribution of the data projected on a two-dimensional space using t-SNE.

Due to the variety of clinical procedures involved in different patient visits, irregularities in data is expected. The frequency and the order of the clinical procedures varies from one patient to another. Hence the episodes of patient data vary with different sets of measures and their frequencies, pre-processing the data for the purpose of this analysis is critical.

### 3.1 Data Pre-processing

Each patient visit is described by a set of measures. These measures are represented as episodes. An episode contains irregular time-stamped vital signs and lab results. In addition, the non-sequential data (gender and age) is also integrated into the episodes.

Every sequence element consists of 30 features (gender, age and 28 vital signs and lab readings)[7]. The interval between the sequences is one day. There are three types of features, starting with constant features which do not change during a patient's visit, such as age and gender. Frequently changing features are

[5] Access to KAIMRCD dataset can be obtained upon official request to KAIMRC.
[6] Western, Central and Eastern regions of Saudi Arabia.
[7] For space reasons the full list of features can not be listed here

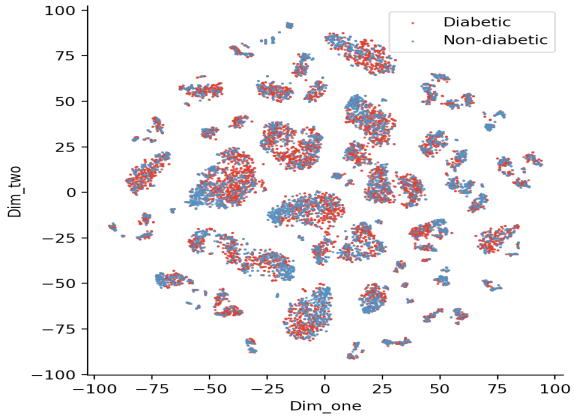


**Fig. 1.** KAIMRCD dataset distribution.

collected on a daily basis, or the average of multiple daily measures, such as vital signs. Finally, the infrequently changing features are collected on an interval of more than a day. As a result, features that may be unavailable for some patients are considered to be missing. The representation of an episode of patient $x$ for our proposed solution is defined as:

$$Episode_x = \begin{Bmatrix} t_1 : R_{11} & R_{12} & ... & R_{1m} \\ t_2 : R_{21} & R_{22} & ... & R_{2m} \\ ... & ... & ... & ... \\ ... & ... & ... & ... \\ t_n : R_{n1} & R_{n2} & ... & R_{nm} \end{Bmatrix}$$

where $R_{ij}$: is the reading values (risk factors) at day $i$ for vital signs, lab test results and the embedded non-sequential values (gender and visit age) $j$. $n$ is the length of the sequence (the input size). $m$ is the number of readings for each sequence.

Patient visit (*Episode*) consists of a sequence length $n$ (based on the length of stay in hospital) at time $t$. If the number of days for a patient's visit is less than $n$, zero padding technique is applied to compensate for the missing sequences. For each sequence there are $m$ reading values ($R$). If $R_{ij}$ is missing then it is assigned the value from the previous day ($R_{ij} = R_{(i-1)j}$). In the case that there was no previous reading, $R_{ij}$ is replaced with zero.

## 4 Methods

Recurrent neural networks, and its variants, have achieved unprecedented accuracy in many domains with sequential data [11]. Unlike other deep learning methods, RNNs have memory cells allowing the previous output to influence the state for the next output, which proved to be a useful feature for sequential data.

Here, we investigate the performance of temporal models: LSTM and GRU, in diagnosing T2DM from time-stamped sequences of patient observations. Given a sequence of observations for a patient $x_t : R_1, R_2, ...R_m$ at time $t$, the activation function of a recurrent hidden unit $h_t$ is:

$$h_t = \nu(U_{x_t} + W_{h_{(t-1)}}), \tag{1}$$

where $\nu$ is a nonlinear function for the sum of the hidden state, $U$, matrix of the current patient's sequences, and W is a matrix of the weight input of the previous sequence.

In the experiments, we use $n$ previous sequences of patient's observations (series) to explore the impact of previous dependencies in influencing the classification decision of T2DM. In practice, RNNs have demonstrated a limited performance when learning from sequences with long-term dependencies [2]. This is mainly caused by limitations in the gradient decent approach, as the gradient tends to either vanish or explode when modelling long dependencies. Hochreiter and Schmidhuber addressed this problem by introducing LSTM [9]. LSTM, uses a sophisticated structure with multiple cell and gated unites (forget and input) to cope with learning from long-term dependencies, described by:

$$f_t = \sigma(W_f.[h_{(t-1)}, x_t] + b_f) \tag{2}$$

$$i_t = \sigma(W_i.[h_{(t-1)}, x_t] + b_i) \tag{3}$$

$$\tilde{C}_t = tanh(W_C.[h_{(t-1)}, x_t] + b_C) \tag{4}$$

$$C_t = f_t \times C_{(t-1)} + i_t \times \tilde{C}_t \tag{5}$$

$$o_t = \sigma(W_o[h_{(t-1)}, x_t] + b_o) \tag{6}$$

$$h_t = o_t \times tanh(C_t), \tag{7}$$

where $f$ represents the forget gate of the cell with a sigmoid activation function $\sigma$ and the weight $W$ and the learned bias $b$ ( Eq. 2). $i$ is the input gate (Eq. 3) which is used in combination with a non-linear(tanh) layer $\tilde{C}$. $\tilde{C}$ is the new value for cell state ( Eq. 4). The update state value $C$ is then the sum of the multiplication of the old state $C_{(t-1)}$ by $f_t$, which decides on what to forget, and the new value $\tilde{C}$ multiplied by the input gate value $i_t$ (Eq. 5). Finally $o$ is the output of the sigmoid gate which is used with the cell state $C$ to produce the final decision (Eq. 6 and Eq. 7) whether the patient $x$ is diabetic or not.

Similar to LSTMs, GRU is used to deal with long-term dependencies. The main difference is that GRU merges the forget and input gates in one unit gate called the update gate. This means that previous memory is kept based on the size of the new dependencies (input). GRUs do not have a protected hidden cell state which gives full access to the corresponding allocated memory content. GRU is formally defined as follows:

$$z_t = \sigma(W_f \times [h_{(t-1)}, x_t]) \tag{8}$$

$$r_t = \sigma(W_r.[h_{(t-1)}, x_t]) \tag{9}$$

$$\tilde{h}_t = tanh(W_C.[r_t \times h_{(t-1)}, x_t]) \tag{10}$$

$$h_t = (1 - z_t) \times h_{(t-1)} + z_t \times \tilde{h}_t, \tag{11}$$

where $z$ and $r$ represent the update gate and the reset gate values. These gates are calculated in a similar way to calculating the input gate and the forget gate of LSTM, except that GRU does not consider adding these values in the formula (Eq. 8) and (Eq. 9). The other difference is that instead of changing the current hidden layer $h$ as in the LSTM method, the input $x$ and the previous layer $h_{(t-1)}$ modify the update gate and the reset gate values in the GRU method. Then the current layer is updated accordingly by $z$ and $r$ (Eq. 11) [4].

## 5 Experimental Setup

Both LSTM and GRU models were implemented, to allow for comparison between their performance in predicting the diagnosis of T2DM. The neural networks of both models have similar architectures. The model contains two LSTM/ GRU layers and two dense layers. The first hidden layer has 128 neurons with a sigmoid activation function,while the second contains 64 neurons with ReLU activation function. The two dense layers also use the ReLU and sigmoid activation functions, with 16 and 1 neurons respectively.

LSTM and GRU are trained using 90% of the data. The remaining 10% is then used for testing. The models use adam optimizer with 0.001 learning rate. The optimisation score function used in both models is root mean squared error. Before preforming the prediction on the test data, the models were trained for 100 epochs. In our experiments, we investigated the performance of each model for six different variations of input sizes (3, 5, 8, 10, 12, and 15). The models are trained and tested using 10-folds cross-validation approach. We report the macro, micro and weighted-averaged F1 scores to compare and evaluate the performance of the classifiers.

**Baseline Models** We compared our results against three commonly used baseline models: Logistic Regression (LR), Support Vector Machine (SVM), and Multi-Layer Perceptron (MLP). These models do not model temporal dynamics in the data, hence the patient visits are assumed independent. Only sequences with fewer missing readings are considered. MLP has similar architecture to LSTM/GRU and uses the same optimiser settings.

## 6 Results

Table 2 shows the performance metrics obtained using LSTM, GRU and baseline models. In table 2, the results show that all of the neural network models, including MLP, with all of the different number of input sizes, achieved better

**Table 2.** Models performance in T2DM diagnosis

| Input size | Model | F1 Weighted | F1 Macro | F1 Micro |
|---|---|---|---|---|
| 1 Sequence* | LR | 0.7790 | 0.7517 | 0.8041 |
| | SVM | 0.7452 | 0.7194 | 0.7576 |
| 3 Sequences | MLP | 0.9409 | 0.9371 | 0.9411 |
| | LSTM | 0.9631 | 0.9649 | 0.9670 |
| | GRU | **0.9706** | **0.9689** | **0.9705** |
| 5 Sequences | MLP | 0.9442 | 0.9406 | 0.9443 |
| | LSTM | 0.9592 | 0.9566 | 0.9596 |
| | GRU | 0.9634 | 0.9612 | 0.9634 |
| 8 Sequences | MLP | 0.9452 | 0.9417 | 0.9451 |
| | LSTM | 0.9565 | 0.9536 | 0.9567 |
| | GRU | 0.9714 | 0.9694 | 0.9715 |
| 10 Sequences | MLP | 0.9508 | 0.9476 | 0.9509 |
| | LSTM | 0.9512 | 0.9485 | 0.9508 |
| | GRU | **0.9729** | **0.9711** | **0.9730** |
| 12 Sequences | MLP | 0.9440 | 0.9403 | 0.9440 |
| | LSTM | 0.9646 | 0.9623 | 0.9646 |
| | GRU | 0.9624 | 0.9598 | 0.9627 |
| 15 Sequences | MLP | 0.9454 | 0.9421 | 0.9451 |
| | LSTM | 0.9669 | 0.9650 | 0.9667 |
| | GRU | 0.9656 | 0.9632 | 0.9657 |

Table 2: shows the performance metrics for LSTM and GRU and baseline classifiers.
* Most complete sequence with fewer missing data among the whole patient's visit.

performance than the models identified in the related work section (Table 1), and the baseline shallow models (LR and SVM).

Both LSTM and GRU outperformed MLP models and achieved promising results using different input sizes (from 3 to 15). GRU with 10 input sequence length is the best performing model with regard to the reported measures (results in bold), but with insignificant difference to GRU with only 3 sequences. Table 2 also shows that GRU models with 3 and 10 sequence length, have better results compared to the same model with larger input size. This is not the same for the LSTM models, which show better results with longer dependencies. Fig 2 shows the performance trend of LSTM, GRU and MLP against the input sizes. Fig 3 demonstrates the models performance results. Fig 3 shows that GRU results are distributed in smaller areas to LSTM, which indicates that GRU approach can have more consistent results when used for predicting T2DM.

### 6.1 Discussion and Conclusion

In this paper, we investigated the use of temporal predictive deep neural network models for the diagnosis of T2DM. The proposed models (LSTM and GRU), using clinical time-stamped data and without intensive feature engineering can achieve very high accuracy with as short as 3 sequences. The models were trained

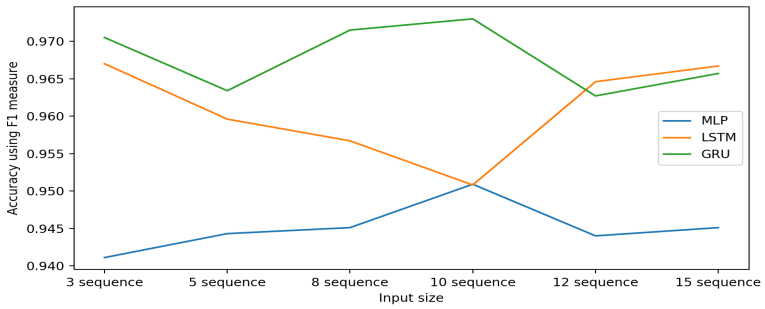


**Fig. 2.** Change of F1 measure with the length of the input size.

and tested with different input sizes using unique and large dataset (KAIMRCD). The results were compared to common baseline classifiers (LR, SVM and MLP) using the same dataset. LSTM and GRU models outperformed the baseline classifiers and achieved 97.3% accuracy. Due to the lack of datasets that are specific to T2DM, replicating this work using different datasets can be difficult.

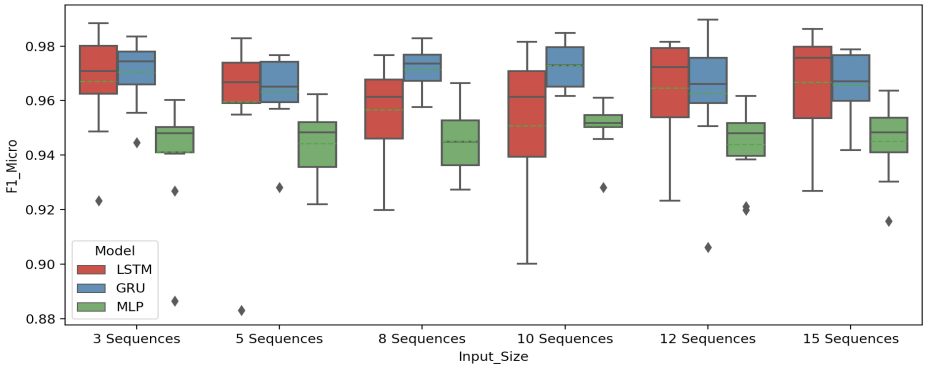


**Fig. 3.** F1 Micro result for LSTM, GRU and MLP Models

The models were able to predict with a high accuracy 97% even with a 3-day length sequence. This is very significant finding as it would reduce the time and associated cost required to perform further tests and delivers early diagnosis. Further work may investigate the impact of applying different techniques for handling the missing data on KAIMRCD data.

## References


1. Beagley, J., Guariguata, L., Weil, C., Motala, A.A.: Global estimates of undiagnosed diabetes in adults. Diabetes research and clinical practice 103(2), 150–160 (2014)
2. Bengio, Y., Simard, P., Frasconi, P.: Learning long-term dependencies with gradient descent is difficult. IEEE transactions on neural networks 5(2), 157–166 (1994)
3. Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078 (2014)
4. Choi, E., Bahadori, M.T., Schuetz, A., Stewart, W.F., Sun, J.: Doctor ai: Predicting clinical events via recurrent neural networks. In: Machine Learning for Healthcare Conference. pp. 301–318 (2016)
5. Chung, J., Gulcehre, C., Cho, K., Bengio, Y.: Empirical evaluation of gated recurrent neural networks on sequence modeling. arXiv preprint arXiv:1412.3555 (2014)
6. Federation, I.D.: Idf diabetes atlas. `http://www.diabetesatlas.org` (2015)
7. Gers, F.A., Schmidhuber, J., Cummins, F.: Learning to forget: Continual prediction with lstm (1999)
8. Goodfellow, I., Bengio, Y., Courville, A.: Deep learning. MIT press (2016)
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)
10. Karegowda, A.G., Manjunath, A., Jayaram, M.: Application of genetic algorithm optimized neural network connection weights for medical diagnosis of pima indians diabetes. International Journal on Soft Computing 2(2), 15–23 (2011)
11. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature 521(7553), 436–444 (2015)
12. Lichman, M.: UCI machine learning repository (2013), `http://archive.ics.uci.edu/ml`
13. Lipton, Z.C., Kale, D.C., Elkan, C., Wetzell, R.: Learning to diagnose with lstm recurrent neural networks. arXiv preprint arXiv:1511.03677 (2015)
14. Lipton, Z.C., Kale, D.C., Wetzel, R.: Modeling missing data in clinical time series with rnns. Machine Learning for Healthcare (2016)
15. Meng, X.H., Huang, Y.X., Rao, D.P., Zhang, Q., Liu, Q.: Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences 29(2), 93–99 (2013)
16. Motka, R., Parmarl, V., Kumar, B., Verma, A.: Diabetes mellitus forecast using different data mining techniques. In: Computer and Communication Technology (ICCCT), 2013 4th International Conference on. pp. 99–103. IEEE (2013)
17. (NHS), U.N.H.S.: `http://www.nhs.uk`
18. Orgnization, W.H.: Global report on diabetes. `http://www.who.int/diabetes/global-report/en/` (2016)
19. Polat, K., Güneş, S.: An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digital Signal Processing 17(4), 702–710 (2007)
20. Temurtas, H., Yumusak, N., Temurtas, F.: A comparative study on diabetes disease diagnosis using neural networks. Expert Systems with applications 36(4), 8610–8615 (2009)
21. Venkatesan, P., Anitha, S.: Application of a radial basis function neural network for diagnosis of diabetes mellitus. Current Science 91(9), 1195–1199 (2006)