Journal of the American Medical Informatics Association, 27(9), 2020, 1343–1351
doi: 10.1093/jamia/ocaa120
Advance Access Publication Date: 1 September 2020
Research and Applications

Research and Applications

Predicting complications of diabetes mellitus using advanced machine learning algorithms

Branimir Ljubic, Ameen Abdel Hai, Marija Stanojevic, Wilson Diaz, Daniel Polimac, Martin Pavlovski, and Zoran Obradovic

Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia, Pennsylvania, USA

Corresponding Author: Zoran Obradovic, PhD, Temple University, 1925 N 12th Street, SERC 334, 035-10, Philadelphia, PA 19121, USA; zoran.obradovic@temple.edu

Received 21 February 2020; Revised 17 May 2020; Editorial Decision 19 May 2020; Accepted 28 May 2020

ABSTRACT

Objective: We sought to predict if patients with type 2 diabetes mellitus (DM2) would develop 10 selected complications. Accurate prediction of complications could help with more targeted measures that would prevent or slow down their development.

Materials and Methods: Experiments were conducted on the Healthcare Cost and Utilization Project State Inpatient Databases of California for the period of 2003 to 2011. Recurrent neural network (RNN) long short-term memory (LSTM) and RNN gated recurrent unit (GRU) deep learning methods were designed and compared with random forest and multilayer perceptron traditional models. Prediction accuracy of selected complications were compared on 3 settings corresponding to minimum number of hospitalizations between diabetes diagnosis and the diagnosis of complications.

Results: The diagnosis domain was used for experiments. The best results were achieved with RNN GRU model, followed by RNN LSTM model. The prediction accuracy achieved with RNN GRU model was between 73% (myocardial infarction) and 83% (chronic ischemic heart disease), while accuracy of traditional models was between 66% – 76%.

Discussion: The number of hospitalizations was an important factor for the prediction accuracy. Experiments with 4 hospitalizations achieved significantly better accuracy than with 2 hospitalizations. To achieve improved accuracy deep learning models required training on at least 1000 patients and accuracy significantly dropped if training datasets contained 500 patients. The prediction accuracy of complications decreases over time period. Considering individual complications, the best accuracy was achieved on depressive disorder and chronic ischemic heart disease.

Conclusions: The RNN GRU model was the best choice for electronic medical record type of data, based on the achieved results.

Key words: diabetes mellitus, diabetes mellitus complications, deep learning, machine learning, RNN models

INTRODUCTION

Objective

Type 2 diabetes mellitus (DM2) is a chronic, metabolic disease and affects almost 100 million people all over the world, including over 30 million in the United States.[1–3] In the last 20 years, the number of adults diagnosed with DM2 has more than doubled, and has quickly become one of the most prevalent and costly chronic diseases worldwide.[4,5] Increased levels of glucose in the blood can cause many health complications over time.[6] Management of DM2 requires a multidimensional approach.[7–9] Identification of people at high risk of progression of DM2 enables targeted prevention.[10,11]

Multiple computer science, especially machine learning (ML), applications have been developed to help with DM2 detection, management, and improvement of patients' quality of life.[12] We designed deep and traditional ML models to predict development of complications in patients diagnosed with DM2. Healthcare Cost and Utilization Project (HCUP) electronic medical record (EMR) data for the period of 9 years were used for experiments. They contain diagnosis, procedures and time of patients' visits. We developed models based on a 1-way recurrent neural network long short-term memory (RNN LSTM) and bidirectional RNN gated recurrent units (GRUs) to capture the temporal nature of EMR data. Traditional models such as random forest (RF) and multilayer perceptron (MLP) were used for comparison.

To evaluate prediction performance of different approaches we selected 10 well-described complications of DM2: angina pectoris, atherosclerosis, ischemic chronic heart disease (ICHD), depressive disorder, diabetic nephropathy, diabetic neuropathy, diabetic retinopathy (DR), hearing loss, myocardial infarction (MI), and peripheral vascular disease.

Following were the objectives of our study:

- Predict if these complications will develop along the course of DM2 (in our study within 9 years from DM2 diagnosis).
- Analyze how many hospitalizations between the diagnosis of DM2 and the diagnosis of each of 10 complications were the most optimal for deep learning or traditional models to produce the best prediction accuracy.
- Test if deep learning RNN models are superior to traditional ML models in accuracy of predictions on the EMR heterogeneous temporal data.
- Analyze how the prediction accuracy of complications would change over time period of 9 years.

Timely and accurate prediction of complications could help with implementation of more specific and targeted measures, which would potentially prevent or slow down their development. Consequently, slowing down the development of complications would save significant economic resources needed for their treatment.

BACKGROUND AND SIGNIFICANCE

Patients with DM2 suffer many life-threatening complications, including macrovascular like stroke, coronary artery disease, or microvascular complications: retinopathy, neuropathy, nephropathy, and others. DM2 represents the most common etiology of extremity pain and diabetic neuropathy.[13,14] Diabetic nephropathy continues to be a chronic and devastating complication of DM2.[15] Diabetes and depression occur together frequently.[16] DM2 appears to impair auditory function.[17] A close link exists between DM2 and cardiovascular diseases.[18]

ML methods were proposed (support vector machine, RF, logistic regression, naïve Bayes) to predict diabetic complications.[19] ML was used for forecasting future glucose fluctuations in the blood.[20] Deep learning LSTM neural networks and probabilistic modeling were designed for prediction of diabetes.[21,22] ML models K nearest neighbors, naïve Bayes, support vector machine, decision tree, logistic regression, and RF were also proposed for prediction of onset of diabetes.[23,24] Clinical risk prediction with limited EMR and challenges of deep learning in Medicine were analyzed.[25,26]

MATERIALS AND METHODS

We conducted experiments on hospital discharges data for 9 years (2003-2011) obtained from the HCUP State Inpatient Databases of California database. The studied dataset contains time of hospitalizations (visits) and International Classification of Diseases–Ninth Revision codes of diagnoses and procedures. The HCUP data were preprocessed, and all patients with the diagnosis of DM2 were extracted (1 910 674 patients), using adequate SQL and Python queries. Original data were rearranged to create a table (matrix). Every row represents 1 patient. Each row contained a patient's hospitalizations in the order in which these visits occurred (Figure 1A). Different colors in each row represent different hospitalizations. Within each hospitalization patients had 1 or more diagnoses and sometimes procedures. Because the procedures domain did not produce good accuracy, we performed detailed analyses on the diagnoses domain only.

In the first group of experiments patients who had the index complication diagnosed after at least 2 hospitalizations from the first DM2 diagnosis were extracted as the positive class. The same number of DM2 patients who did not develop the index complication were randomly selected for the negative class, using a population-based sample.[27]

In the second group of experiments, patients were selected for the positive class if the complication appeared after at least 3 visits from the first DM2 diagnosis. In the third group of experiments patients who developed the studied complication after at least 4 hospitalizations from the DM2 diagnosis were selected for the positive class. From each of these 3 datasets we randomly selected matching pairs (by minimum number of hospitalizations) of positive and negative cohorts. Thirty balanced datasets were created, 1 for each of the 10 complications in each of the 3 groups of experimental settings.

The average number of visits per patient was 4.07 ± 5.08. All the hospitalizations starting from the hospital visit in which patients were diagnosed with the complication that we were predicting were excluded from the positive cohort, in order to avoid data leak. After this adjustment, the average number of hospitalizations for patients with a positive label was very similar to the average number of visits for patients with a negative label.

Diseases that appeared rarely or too frequently in the selected datasets do not contribute to the prediction, as they do not have high informative value. All diseases that appeared more than

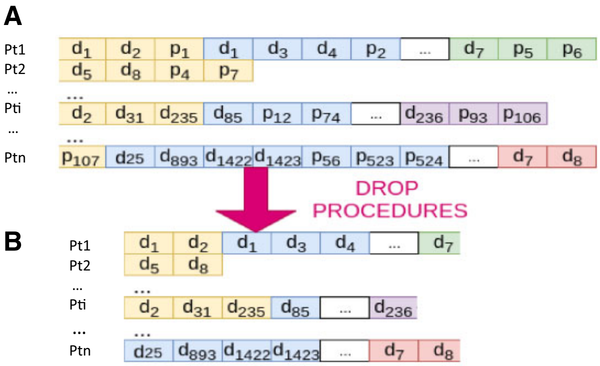
A
Pt1 d_1 d_2 p_1 d_1 d_3 d_4 p_2 ... d_7 p_5 p_6
Pt2 d_5 d_8 p_4 p_7
...
Pti d_2 d_{31} d_{235} d_{85} p_{12} p_{74} ... d_{236} p_{93} p_{106}
...
Ptn p_{107} d25 d_{893} d_{1422} d_{1423} p_{56} p_{523} p_{524} ... d_7 d_8
DROP PROCEDURES
B
Pt1 d_1 d_2 d_1 d_3 d_4 ... d_7
Pt2 d_5 d_8
...
Pti d_2 d_{31} d_{235} d_{85} ... d_{236}
...
Ptn d25 d_{893} d_{1422} d_{1423} ... d_7 d_8

Figure 1. (A) Each row represents 1 patient (Pt). Different colors in each row represent different hospitalizations. Each hospitalization contained 1 or more diagnoses (d) and sometimes procedures (p). (B) Because the procedures domain did not produce good results, we dropped them and performed analyses on the diagnoses domain only.

200 000 or <50 times among all the patients were deleted. After this preprocessing, 1023 International Classification of Diseases–Ninth Revision disease codes were used to represent the patients' hospital visits. We applied singular value decomposition (SVD) to reduce dimensionality of visits.[28] This dimensionality reduction method uses matrix decomposition to transform features and select only features with the highest variance because those are the most informative characteristics.

Input to SVD was a matrix in which rows were all visits in a dataset and columns were all possible disease codes for that dataset. We have used a flat (one-hot encoded) representation of the "diagnoses," and not a flat representation of "time." Each of patients had at least few visits (rows in the matrix). Hospitalizations happened over the time period, with the maximum interval of 9 years. Although time is not specifically used as one of the features, the time component is reflected by the fact that consecutive visits were ordered by their timesteps. Each cell value in the matrix represented if a specific disease was present inside the visit. The value of each cell, therefore, was 0 or 1. Most of the cells had value 0 because only a few diseases appeared in each hospitalization. Output of SVD is a matrix in which rows are patients but columns are 50 features in a transformed space with the highest variance selected by SVD, which captures block correlations between data features.

Further, we created a matrix in which rows represented patients and columns were hospital visits ordered by timestamps. We deleted all patients who had more than 50 hospitalizations to reduce the size of a sparse matrix, but most of the patients had much less than 50 hospitalizations. If a patient had <50 hospitalizations, we padded their row with 0 to ensure that all rows had the same length. Then, we substituted each visit with a 50-feature vector from SVD and zero (nonexisting visits) with a 0-vector of length 50. In other words, new matrix rows are patients containing concatenated feature vectors of that patients' visits (each row has up to 50 hospitalizations * 50 features = 2500). After preprocessing and feature selection, the dimensions of data matrix were (number of patients) × (number of features) and this was the input for all the ML models that we tested in this study (RNN LSTM, RNN GRU, RF, MLP). The input for RNN (and other) models were all hospitalizations of all patients given to the model in chronological order, as the sequence ordered by timestamps for each patient.

Two types of ML models were utilized in this work: deep learning models and traditional models. The proposed deep learning models were 1-way RNN LSTM and bidirectional RNN GRU (Figure 2).

RNN is a neural network where hidden neurons can analyze temporal sequential EMR data.[29] It has the same structure as the basic neural network, but neurons in the same layer are connected, allowing for neurons to learn information from its left neighbor in addition to the current input. Therefore, RNN neurons have 2 sources of inputs, the present and the recent past. Learning process is described with following equations:

$$h^t = relu\left(b + Wh^{t-1} + Ux^t\right) \tag{1}$$

$$\bar{y} = sigmoid\left(b + \sum\nolimits_t Vh^t\right) \tag{2}$$

To calculate value h^t of a hidden neuron t, a nonlinear transformation, ReLU, is applied to weighted W value of its left hidden neuron h^{t-1} and the weighted U value of its input x^t. Prediction is calculated as a sigmoid function of weighted V sum of all hidden neurons with added bias b. Learning is achieved with backpropagation. Because of the long chain through which historical information (in forward direction) and gradient of error (in backward direction) had to be pass, RNN suffers from the vanishing gradient problem, which means that weights do not change and the model is not able to learn. To remedy this, a LSTM was invented, in which simple neurons of RNN are replaced with more complex short-term memory structure. LSTM shares the same weights across layers, which reduces the number of parameters that the network has to deal with. The GRU is another solution for vanishing gradient. It substitutes the simple neuron with a gated unit, which has fewer parameters than the LSTM neuron, because it lacks an output gate.[30]

To model heterogeneous sequential data, we tested bidirectional GRU as a proposed method and compared it to 1-way LSTM. "Keras" Python libraries were used to implement constructed algorithms. We also compared GRU and LSTM to traditional ML algorithms. Our hypothesis was that both deep learning methods (RNN LSTM and RNN GRU) would perform better than traditional ML models (RF and MLP) on medical temporal data like that of HCUP because of their ability to learn from a patient's history. In the proposed model, we used ReLU and sigmoid activations. Also, we added a dropout between hidden and output layers, which randomly selects given percent of connections to cut. This is a well-known regularization technique that helps the model learn general pattern in data.

We used RF as a traditional model because they have been shown as the state-of-the-art model in existing literature on predicting complications of diabetes and MLP because it is a simpler neural network model that does not account for time. Both were implemented with the "Scikit-learn" library in Python. The RF is a classification algorithm which consists of many decision trees.[31] MLP is a network that consists of multiple layers of perceptrons and uses backpropagation learning. It uses a nonlinear activation function, which in addition to multiple layers distinguishes it from a linear perceptron.[32]

We compared the performance of these 4 models in prediction of the occurrence of 10 selected complications of DM2 with each of the 3 settings (2, 3, or 4 hospitalizations after DM2 diagnosis). The problem that we wanted to solve was a binary classification task. The evaluation metric was accuracy (equation 3) computed as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

Further, we tested what would be the minimum number of patients for RNN LSTM and GRU deep learning models to work properly and produce good prediction accuracy, after which the performance of these models starts to decrease. Two groups of experiments considering the minimum required number of patients for deep learning RNN models to perform optimally were conducted on DR data with at least 2 or 4 hospitalizations. We performed experiments with the entire datasets (58 641 patients with 2 and 25 468 patients with 4 hospitalizations) and then with randomly selected 5000, 2000, 1000, and 500 patients in the positive cohort to discover the number of patients in the positive training cohort after which the accuracy starts dropping.

Furthermore, the prediction accuracy was evaluated for different time intervals between DM2 diagnosis and the first diagnosis of studied complications. We evaluated the accuracy of GRU RNN models for the intervals between DM2 and development of DR for: less than a year, 1 year, 2 years, 3 years, 4-5 years, and 6-8 years. Analyses were conducted on 2 experimental models: the complica-

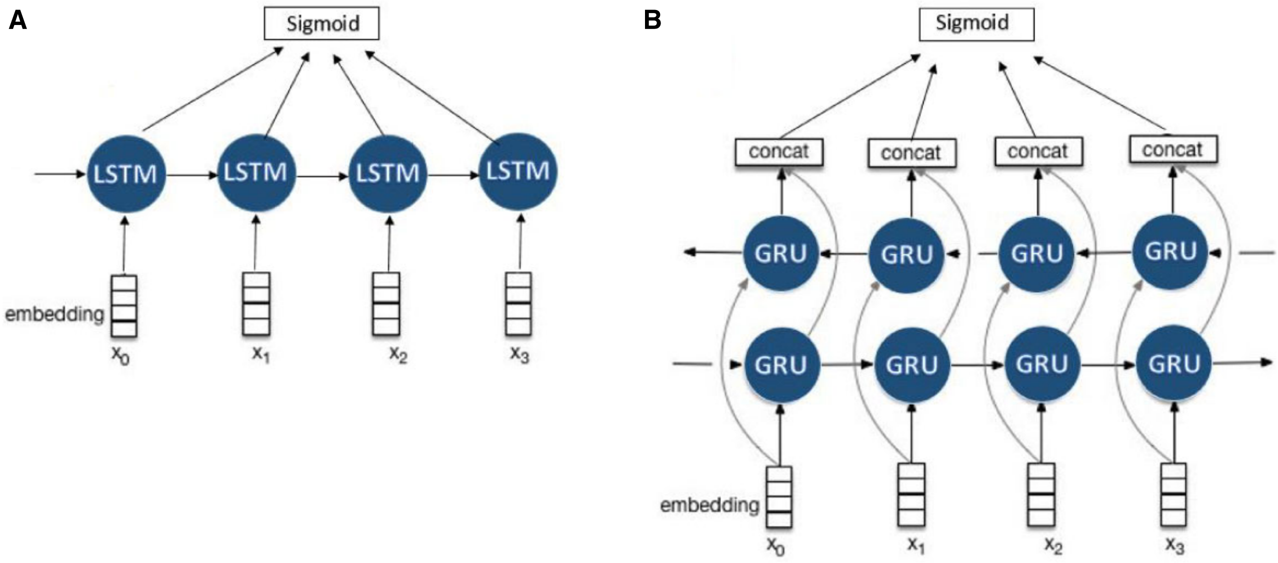

Figure 2. The proposed deep learning models: (A) 1-way recurrent neural network long short-term memory (RNN LSTM) and (B) bidirectional RNN gradient recalled unit (GRU).

tion developed after 2 or 4 hospitalizations from the initial DM2 diagnosis. Because the entire dataset covers the period of 9 years, the maximum interval between DM2 diagnosis and the first diagnosis of DR identified in the dataset was 8 years.

Data in all experimental settings were split into 72% training, 8% validation, and 20% testing. Cross-validation was used to find the best hyperparameters values. For proposed RNN-like models, we varied dropout and number of neurons in GRU/LSTM layer using random search. In the literature, dropout percent is usually between 0% and 50% and the number of units in the GRU/LSTM layer are usually selected among values 32, 64, 128, 256, and 512. For all RNN networks, we used batch size 128, Adam optimizer, and binary cross-entropy loss. We trained 20 epochs for RNN and tested on the epoch that had the best cross-validation accuracy. RF and MLP were trained with specific hyperparameters as well. In RF, maximum height of trees was bounded to 10 and number of trees was 100. MLP had 100 hidden units.

We repeated the same process for 10 complications separately and we repeated all tests for datasets with a filter of the minimum of 2, 3, or 4 hospitalizations before the studied complication was diagnosed. We used a t test at the level of $P = .05$ to check the significance of accuracy results that tested ML models produced in different experimental settings. Finally, probabilities that patients with DM2 will develop each of the studied complications (HCUP data) were calculated. Our codes are available on a public repository (https://github.com/bljubic/diabetes-prediction).

RESULTS

The total number of patients in the HCUP State Inpatient Databases California dataset between 2003 and 2011 as well as the number of patients with DM2 diagnosis are shown in Table 1. We also present the number of patients with at least 4, 3, or 2 hospitalizations after DM2 diagnosis and before an index complication was diagnosed. These datasets were the source of data for positive and negative cohorts for all experiments.

Table 1. Datasets used in experiments and their sizes

Dataset	Patients
HCUP SID California (2003-2011)	11 609 450
Patients in HCUP with diagnosed DM2	1 910 674
Patients with DM2 and 2 hospitalizations	1 295 691
Patients with DM2 and 3 hospitalizations	930 837
Patients with DM2 and 4 hospitalizations	692 397

DM2: type 2 diabetes mellitus; HCUP: Healthcare Cost and Utilization Project; SID: State Inpatient Databases.

Experiments were performed separately for each of 10 complications, and results for RNN deep learning models (bidirectional GRU and 1-way LSTM) as well as traditional models (RF and MLP) are presented in Table 2. The evaluation metric is accuracy on out-of-sample data, and sizes of samples for each type of experiments are shown in the same table.

In Table 3, we present the accuracy, sensitivity, and specificity results for bidirectional GRU RNN models in the 4-visit scenario for all complications of DM2, which was the model that achieved the best prediction accuracy. The results for 2 and 3 visits are omitted because they were consistent with results for 4 hospitalizations.

Different choices of hyperparameters were tested. For bidirectional GRU RNN, the best results were achieved with the dropout parameter value 0.2 and 128 hidden GRU neurons. We tried randomly dropout parameters between 0 and 0.5 and the number of hidden units 32, 64, 128, 256, and 512 in 20 experimental runs for each type of hyperparameters. Accuracy results varied 2% in experiments for parameters selection. The best results for LSTM RNN model were achieved with the dropout parameter value 1.9 and 128 LSTM neurons. We present the average accuracy of 20 runs, including the standard deviation. We used the same set of hyperparameters in experiments with 2, 3, or 4 hospital visits.

Changes in the prediction accuracy of deep learning (RNN) models as well as traditional models when the size of positive training cohorts decreases are presented on the example of DR in Table 4.

Table 2. Presented are results of predicted accuracy that each of the 10 complications of DM2 will develop within a 9-year period after the first DM2 diagnosis using HCUP EMR data (diagnoses domain).

Complication	Patients	Bidirectional GRU	1-way LSTM	RF	MLP
Angina pectoris					
4 visits	19 589	0.796 ± 0.024[a]	0.780 ± 0.016	0.717 ± 0.011	0.743 ± 0.013
3 visits	26 973	0.789 ± 0.012	0.793 ± 0.019[b]	0.722 ± 0.012	0.732 ± 0.008
2 visits	42 459	0.738 ± 0.016[b]	0.738 ± 0.018	0.701 ± 0.013	0.714 ± 0.009
Atherosclerosis					
4 visits	32 914	0.756 ± 0.003[a]	0.750 ± 0.015	0.712 ± 0.007	0.691 ± 0.008
3 visits	44 688	0.750 ± 0.008[b]	0.745 ± 0.012	0.704 ± 0.011	0.671 ± 0.008
2 visits	62 016	0.713 ± 0.011[b]	0.701 ± 0.018	0.689 ± 0.014	0.665 ± 0.012
ICHD					
4 visits	52 959	0.835 ± 0.005[a]	0.828 ± 0.008	0.759 ± 0.009	0.761 ± 0.017
3 visits	81 658	0.814 ± 0.008[b]	0.813 ± 0.007	0.745 ± 0.010	0.763 ± 0.015
2 visits	147 718	0.802 ± 0.010[b]	0.802 ± 0.015	0.744 ± 0.014	0.758 ± 0.015
Depressive disorder					
4 visits	56 343	0.820 ± 0.005[a]	0.812 ± 0.008	0.714 ± 0.011	0.752 ± 0.013
3 visits	78 732	0.802 ± 0.018	0.810 ± 0.004[b]	0.739 ± 0.014	0.741 ± 0.015
2 visits	135 492	0.773 ± 0.021	0.776 ± 0.019[b]	0.722 ± 0.016	0.761 ± 0.016
Hearing impairment					
4 visits	8576	0.734 ± 0.017[b]	0.720 ± 0.021	0.691± 0.019	0.701 ± 0.021
3 visits	12 030	0.743 ± 0.017[a]	0.730 ± 0.020	0.694 ± 0.021	0.704 ± 0.019
2 visits	16 884	0.716 ± 0.019[b]	0.694 ± 0.022	0.680 ± 0.024	0.671 ± 0.023
MI					
4 visits	38 380	0.733 ± 0.011[a]	0.713 ± 0.013	0.691 ± 0.010	0.661 ± 0.016
3 visits	52 896	0.723 ± 0.014[b]	0.701 ± 0.012	0.688 ± 0.013	0.665 ± 0.014
2 visits	92 961	0.711 ± 0.015[b]	0.679 ± 0.013	0.663 ± 0.015	0.662 ± 0.017
Nephropathy					
4 visits	37 982	0.768 ± 0.012[a]	0.750 ± 0.014	0.699 ± 0.014	0.694 ± 0.024
3 visits	52 283	0.766 ± 0.013[b]	0.748 ± 0.013	0.696 ± 0.015	0.689 ± 0.020
2 visits	71 053	0.742 ± 0.008[b]	0.738 ± 0.010	0.695 ± 0.012	0.678 ± 0.015
Neuropathy					
4 visits	49 060	0.746 ± 0.053[a]	0.719 ± 0.073	0.671 ± 0.033	0.668 ± 0.039
3 visits	69 053	0.738 ± 0.043	0.739 ± 0.068[b]	0.664 ± 0.040	0.664 ± 0.046
2 visits	99 825	0.715 ± 0.038[b]	0.712 ± 0.054	0.660 ± 0.035	0.662 ± 0.055
PVD					
4 visits	48 565	0.767 ± 0.002[a]	0.744 ± 0.014	0.695 ± 0.006	0.691 ± 0.014
3 visits	67 686	0.759 ± 0.006[b]	0.743 ± 0.010	0.708 ± 0.009	0.684 ± 0.010
2 visits	93 905	0.738 ± 0.011[b]	0.738 ± 0.014	0.701 ± 0.008	0.680 ± 0.012
Retinopathy					
4 visits	27 796	0.796 ± 0.014[a]	0.782 ± 0.001	0.741 ± 0.011	0.740 ± 0.007
3 visits	36 221	0.752 ± 0.021[b]	0.731 ± 0.013	0.698 ± 0.012	0.700 ± 0.011
2 visits	58 641	0.728 ± 0.019[b]	0.725 ± 0.014	0.696 ± 0.018	0.676 ± 0.012

Values are mean ± SD. This period varies between 1 month and 9 years for individual patients. Results are presented for patients who had at least 2, 3, or 4 visits between the first DM2 diagnosis and before each of 10 complications was diagnosed.

DM2: type 2 diabetes mellitus; EMR: electronic medical record; GRU: gated recurrent unit; HCUP: Healthcare Cost and Utilization Project; LSTM: long short-term memory; MI: myocardial infarction; MLP: multilayer perceptron; RF: random forest; RNN: recurrent neural network.

[a] Best accuracy results for each of 10 complications.

[b] Best accuracy results for each experimental setting.

The performance of both deep learning RNN models deteriorates when the training dataset size decreases, especially when the number of patients in the positive training dataset drops below 1000. The traditional models' performance vary slightly but does not change statistically significantly.

The prediction accuracy (GRU RNN model) of development of DR within the same year when DM2 was diagnosed and after 1, 2, 3, 4-5, and 6-8 years of diagnosis of DM2 are presented in Figure 3. Experiments were completed with data of patients who had at least 2 hospitalizations or at least 4 hospitalizations after DM2 was diagnosed. All other complications have similar trends of the predicted accuracy regarding the time intervals.

Predicted risk probabilities of development of each of 10 studied complications in patients with DM2, according to HCUP data, are presented in Figure 4.

DISCUSSION

In conducted experiments both deep learning algorithms were significantly more accurate than traditional models. Sarwar and colleagues reported accuracies for prediction of diabetes for the following ML models: logistic regression 74% accuracy, SVM 77%, naïve Bayes, decision tree 74%, RF 71% and K nearest neighbors

Table 3. Accuracy, sensitivity, and specificity for bidirectional GRU RNN models in the 4-visit scenario for all 10 complications of DM2

Complication	Accuracy	Sensitivity	Specificity
Angina pectoris	0.796 ± 0.024	0.862±0.019	0.698±0.014
Atherosclerosis	0.756 ± 0.003	0.791±0.012	0.718±0.014
ICHD	0.835 ± 0.005	0.886±0.014	0.787±0.012
Depressive disorder	0.820 ± 0.005	0.848±0.009	0.792±0.010
Hearing impairment	0.734 ± 0.017	0.743±0.016	0.722±0.012
MI	0.733 ± 0.011	0.806±0.021	0.652±0.012
Nephropathy	0.768 ± 0.012	0.826±0.017	0.654±0.021
Neuropathy	0.746 ± 0.053	0.795±0.041	0.701±0.049
PVD	0.767 ± 0.002	0.774±0.005	0.753±0.011
Retinopathy	0.796 ± 0.014	0.799±0.007	0.792±0.018

Values are mean ± SD.

DM2: type 2 diabetes mellitus; GRU: gated recurrent unit; ICHD: ischemic chronic heart disease; MI: myocardial infarction; PVD: peripheral vascular disease; RNN: recurrent neural network.

achieved 77%.[24] Ngufor et al[33] applied tree-based ML algorithms such as RF, gradient-boosted machine, recursive partitioning, conditional inference trees, and a mixed-effect ML (MEml) framework to predict longitudinal change in hemoglobin A1c. Ngufor et al's model assumes that the number of variables which change over the time period is small. In the case of hemoglobin A1c application, there is only 1 continuous variable that changes longitudinally. However, in our experiments, diagnoses are categorical variables (with more than 1000 categories) which change with each visit, making Ngufor et al's method inapplicable. In the last time, according to numerous publications, RNN-based models proved to be superior to traditional models with high-dimensional temporal EMR type of data. Choi et al[34] showed that RNN deep learning models performed better than traditional ML approaches on EMR temporal data. Massaro et al[21] described an application of a deep learning LSTM model as a very good choice for diabetes prediction. Zhang et al[25] applied the MetaPred model and transfer learning using convolutional neural networks and LSTM RNN models in addition to traditional ML models.

Table 4. Experiments conducted on DR datasets with 2 and 4 hospitalizations in order to test changes in accuracy results with the decrease of the training dataset size

Hospitalizations	Patients	Bidirectional GRU	1-way LSTM	RF	MLP
4	27 796	0.796 ± 0.014[a]	0.782 ± 0.001[a]	0.741 ± 0.011[a]	0.740 ± 0.007[a]
4	5000	0.782 ± 0.012	0.776 ± 0.006	0.738 ± 0.010	0.747 ± 0.011
4	2000	0.765 ± 0.011	0.743 ± 0.010	0.752 ± 0.014	0.766 ± 0.012
4	1000	0.769 ± 0.014	0.767 ± 0.002	0.752 ± 0.009	0.742 ± 0.008
4	500	0.745 ± 0.013	0.745 ± 0.008	0.740 ± 0.013	0.750 ± 0.009
2	58 641	0.728 ± 0.019[a]	0.725 ± 0.014[a]	0.696 ± 0.018[a]	0.676 ± 0.012[a]
2	5000	0.715 ± 0.014	0.706 ± 0.015	0.690 ± 0.021	0.662 ± 0.016
2	2000	0.707 ± 0.015	0.707 ± 0.011	0.687 ± 0.019	0.660 ± 0.015
2	1000	0.700 ± 0.019	0.685 ± 0.015	0.662 ± 0.012	0.657 ± 0.009
2	500	0.659 ± 0.018	0.640 ± 0.010	0.650 ± 0.016	0.652 ± 0.011

Values are mean ± SD.

DR: diabetic retinopathy; GRU: gated recurrent unit; LSTM: long short-term memory; MLP: multilayer perceptron; RF: random forest.

[a]Best accuracy result.

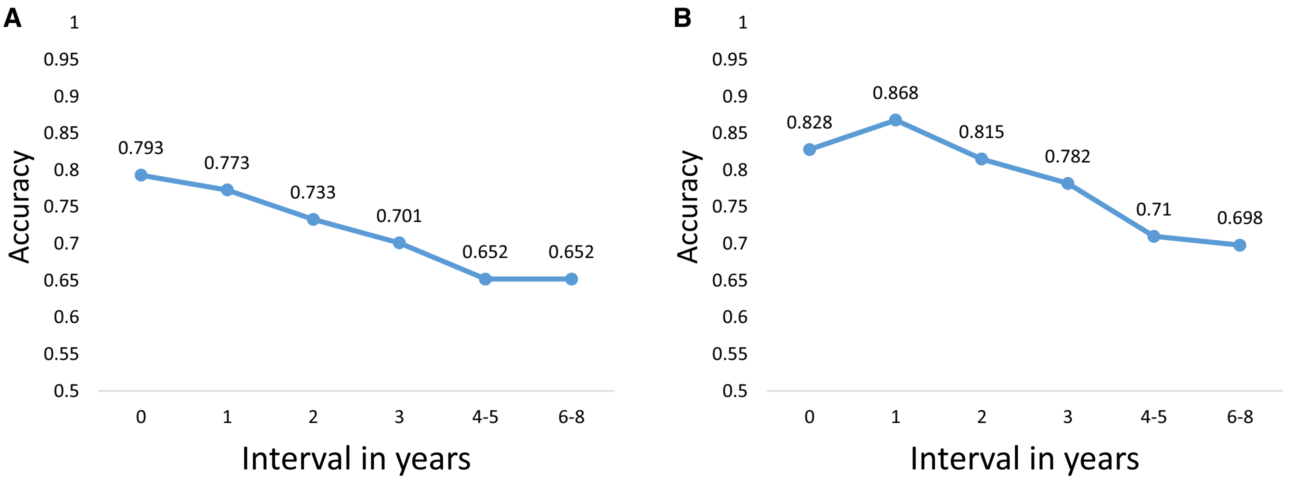

Figure 3. Prediction accuracy (recurrent neural network gradient recalled unit model) that patients with type 2 diabetes mellitus would develop diabetic retinopathy (A) after a minimum of 2 hospitalizations and (B) after at least 4 hospitalizations. The results are presented by intervals when retinopathy developed within 1 year and after 1, 2, 3, 4-5, and 6-8 years from the diagnosis of type 2 diabetes mellitus.

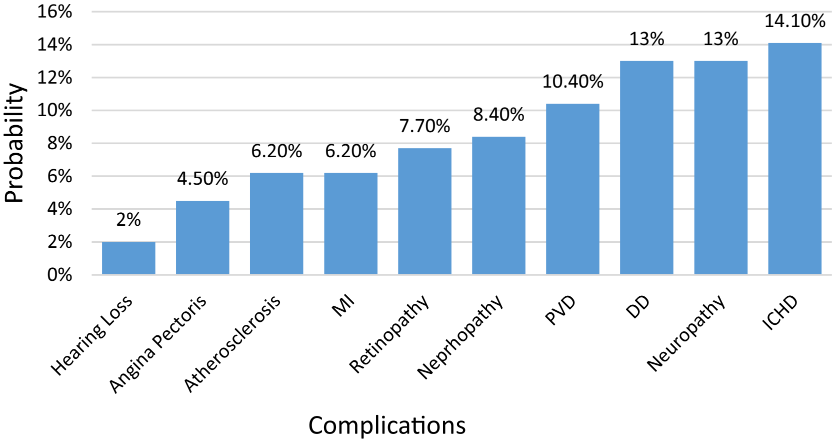

Figure 4. Predicted risk probabilities of development of each of 10 complications in patients with type 2 diabetes mellitus (Healthcare Cost and Utilization Project State Inpatient Databases California data). DD: depressive disorder; ICHD: ischemic chronic heart disease; MI: myocardial infarction; PVD: peripheral vascular disease.

Our study focuses on detection of the most often complications of DM2, using high-dimensional EMR type of data. RNN models, especially GRU, achieved state-of-the-art prediction accuracy (Table 2) by discovering complex temporal relationships inside EMR data. A *t* test ($P = .05$ level) did not show a statistically significant difference between the 2 RNN models. Comparison of 2 traditional ML models shows that both models performed similarly. Our experiments show that an RNN GRU model is the best choice for the high dimensional temporal EMR data. It performed significantly better than traditional models, according to a *t* test at the level of $P = .05$.

Considering the number of hospitalizations filter deep learning models achieved the best results when data include 4 hospitalizations after DM2 diagnosis vs relaying on less hospitalizations. The accuracy of our RNN GRU model with 4 hospitalizations achieved values between 73.3% (MI) and 83.5% (ICHD). Datasets with 2 hospitalizations are 2-3 times larger than datasets with 4 hospitalizations. Our results point toward the number of hospitalizations as more important factor for prediction results than the size of datasets. We did not find significant difference in prediction accuracy if the minimum number of hospitalizations was 3 instead of 4.

Analyzes of the influence of sizes of datasets indicate that about 1000 patients are sufficient in the positive dataset for RNN models. The performance of RNN ML models decreased significantly when the size of datasets decreased to 500 patients. Traditional ML models did not show statistically significant changes in achieved accuracy if the size of datasets decreased to 500. They also did not show significant changes in accuracy with changes in the number of hospitalizations.

Figure 3 shows that prediction accuracy of DR decreases over time. In case of 2 hospitalizations, the accuracy decreases steadily over time because the sizes of datasets are big enough not to affect performance of deep learning models. In experiments with 4 hospitalizations the initial accuracy within 1 year is lower because that dataset was relatively small (523 patients), which is less than the optimal number of patients for the performance of the GRU model. After this initial period, the rest of the curve in Figure 3B is similar to the curve in Figure 3A. Similar changes are noticed with all other complications.

Considering individual complications, RNN models were the most accurate when predicting depressive disorder and ICHD. These 2 diseases had the largest absolute numbers of patients in positive cohorts. The prediction accuracy of ICHD was 83.5%, which was significantly better than the prediction accuracy of hearing loss (the smallest dataset) or MI which were 73.4% and 73.3% consequently. Individual diseases performed differently, and it is difficult to determine whether the size of datasets, comorbidities of individual complications, or perhaps time gaps between different visits influenced prediction results. Analyses of probabilities of development of 10 complications show that ICHD, depressive disorder, and diabetic neuropathy have higher probability of occurrence (13%-14%) than all other complications, including hearing loss with the probability of occurrence of only 2%.

Results of our research with early and accurate prediction of 10 frequent complications of DM2 are important for targeting high-risk patients for monitoring and intervention. They would enable application of timely prevention measures, which will postpone complications, improve quality of life, and increase survival rates. Our methodology could be generally applied to prediction accuracy problems of any other disease or complications of that disease. It could be applied to predict cancer diseases, from EMR type of data, or it can be applied to predict chronic diseases, such as heart or lung diseases or complications of those chronic diseases. It can also be applied to predict acute medical conditions, such as heart attack, stroke, acute kidney failure, or occurrence of infectious diseases (eg, flu, coronavirus, hepatitis).

Further improvement of created RNN models would improve prediction accuracy of DM2 complications and other diseases, which could have significant clinical implications. It could become incorporated into a clinical decision support system and help clinical workers to improve quality of health care. Our GRU RNN model can predict a clinical event (disease, complication) with high accuracy.

By demonstrating that application of RNN deep learning models can make successful prediction of clinical events we hope that our study may contribute to facilitate a wider use of ML in clinical medicine in the form of a clinical decision support system. Also, it could be applied in healthcare emergencies such as the current crisis with COVID-19 (coronavirus disease 2019) to make some important and helpful predictions that will help to public health experts.

CONCLUSION

Deep learning approaches, especially the RNN GRU model, were superior to traditional ML models with temporal EMR medical data. The conducted large-scale experiments suggest that the number of hospitalizations (visits) should be 3 or more in the case of temporal data if deep ML models are applied. The tradeoff between the number of hospitalizations and the size of datasets should be considered, because datasets with 3 visits could be significantly larger than those with 4 visits, which will require more computational resources.

Deep learning models applied on the HCUP data achieved a very good prediction accuracy with 10 selected complications of DM2. Improvements in the accuracy of results might be possible if we had had data from other domains, such as labs or drugs, available.

Our study provides evidence that better understanding and management of DM2 from the aspect of the studied complications is possible when training deep learning models on appropriately preprocessed EMR data. An accurate prediction of the occurrence of complications is important in the planning of targeted measures aimed to slow down or prevent their development.

FUNDING

This research was supported in part by the King Abdullah University of Science and Technology Center Partnership Fund and Pennsylvania Department of Health CURE Health Data Science Research Project.

AUTHOR CONTRIBUTIONS

BL, AAH and MS designed the method and wrote the source code. WD and MP preprocessed and prepared the dataset used in this study. BL, AAH, MS and DP designed the experimental setup and ran all experiments. BL and MP developed the main idea behind the paper and received valuable feedback from ZO. All authors were engaged in the writing of the paper under the supervision of ZO.

CONFLICT OF INTEREST STATEMENT

The authors have no competing interests to declare.

REFERENCES

1. Orasanu G, Plutzky J. The pathologic continuum of diabetic vascular disease. *J Am Coll Cardiol* 2009; 53 (5): S35–42.
2. Deedwania PC. Management of patients with stable angina and type 2 diabetes. *Rev Cardiovasc Med* 2015; 16 (2): 105–13.
3. Duh EJ, Sun JK, Stitt AW. Diabetic retinopathy: current understanding, mechanisms, and treatment strategies. *JCI Insight* 2017; 2 (14): e93751.
4. Centers for Disease Control and Prevention. What is diabetes? https://www.cdc.gov/diabetes/basics/diabetes.html.
5. World Health Organzation. Diabetes. https://www.who.int/health-topics/diabetes.
6. National Institute of Diabetes and Digestive and Kidney Diseases. Diabetes. https://www.niddk.nih.gov/health-information/diabetes.
7. Rodriguez-Gutierrez R, Gonzalez-Gonzalez JG, Zuñiga-Hernandez JA, *et al*. Benefits and harms of intensive glycemic control in patients with type 2 diabetes. *BMJ* 2019; 367: l5887.
8. Garber AJ, Abrahamson MJ, Barzilay JI, *et al*. Consensus statement by the American Association of Clinical Endocrinologists and American College of Endocrinology on the comprehensive type 2 Diabetes management Algorithm - 2018 Executive summary. *Endocr Pract* 2018; 24 (1): 91–120.
9. Qureshi M, Gammoh E, Shakil J, *et al*. Update on management of type 2 diabetes for cardiologists. *Methodist Debakey Cardiovasc J* 2018; 14 (4): 273–80.
10. Bailey CJ, Day C. Treatment of type 2 diabetes: future approaches. *Br Med Bull* 2018; 126 (1): 123–37.
11. Cahn A, Shoshan A, Sagiv T, *et al*. Use of a Machine Learning Algorithm Improves Prediction of Progression to Diabetes. *Diabetes* 2018; 67 (suppl 1): db18-1286-P.
12. Contreras I, Vehi J. Artificial intelligence for diabetes management and decision support: literature review. *J Med Internet Res* 2018; 20 (5): e10775.
13. Bodman MA, Varacallo M. *Diabetic Neuropathy*. Treasure Island, FL: StatPearls Publishing; 2020.
14. Callaghan BC, Cheng H, Stables CL, *et al*. Diabetic neuropathy: clinical manifestations and current treatments. *Lancet Neurol* 2012; 11 (6): 521–34. doi: 10.1016/S1474-4422(12)70065-0.
15. Bouhairie VE, McGill JB. Diabetic kidney disease. *Mo Med* 2016; 113 (5): 390–4.
16. Holt RIG, de Groot M, Golden SH. Diabetes and depression. *Curr Diab Rep* 2014; 14 (6): 491.
17. Konrad-Martin D, Reavis KM, Austin D. Hearing impairment in relation to severity of diabetes in a Veteran cohort. *Ear Hear* 2015; 36 (4): 381–94.
18. Leon BM, Maddox TM. Diabetes and cardiovascular disease: epidemiology, biological mechanisms, treatment recommendations and future research. *World J Diabetes* 2015; 6 (13): 1246–58.
19. Dagliati A, Simone Marini S, Sacchi L, *et al*. Machine learning methods to predict diabetes complications. *J Diabetes Sci Technol* 2018; 12 (2): 295–302.
20. Hayeri A. Predicting future glucose fluctuations using machine learning and wearable sensor data. *Diabetes* 2018; 67 (suppl 1): db18-738-P.
21. Massaro A, Maritati V, Giannone D, *et al*. LSTM DSS automatism and dataset optimization for diabetes prediction. *Appl Sci* 2019; 9 (17): 3532.
22. Perveen S, Shahbaz M, Keshavjee K, *et al*. Prognostic Modeling and Prevention of Diabetes Using Machine Learning Technique. https://doi.org/10.1038/s41598-019-49563-6.
23. Apoorva S, Aditya SK, Snigdha P, *et al*. Prediction of diabetes mellitus type-2 using machine learning. In: Smys S, Tavares J, Balas V, Iliyasu A, eds. *Computational Vision and Bio-Inspired Computing. ICCVBIC 2019. Advances in Intelligent Systems and Computing*, vol 1108. Cham, Switzerland: Springer; 2020: 364–70. https://doi.org/10.1007/978-3-030-37218-7_42.
24. Sarwar MA, Kamal N, Hamid W, *et al*. Prediction of diabetes using machine learning Algorithms in healthcare. In: 2018 24th International Conference on Automation and Computing (ICAC); 2018.
25. Zhang XS, Tang F, Dodge H, *et al*. MetaPred: Meta-Learning for Clinical Risk Prediction with Limited Patient Electronic Health Records. In: proceedings of the 25th ACM SIG KDD International Conference on Knowledge Discovery and Data Mining (KDD); 2019: 2487–95. doi.org/10.1145/3292500.3330779.
26. Wang F, Casalino LP, Khullar D. Deep learning in medicine—promise, progress, and challenges. *JAMA Intern Med* 2019; 179 (3): 293–4.
27. Man CJ. Observational research methods. Research design II: cohort, cross sectional, and case-control studies. *Emerg Med J* 2003; 20: 54–60.
28. Klema V, Laub A. The singular value decomposition: its computation and some applications. *IEEE Trans Automat Contr* 1980; 25 (2): 164–76.
29. Williams RJ, Zipser D. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput* 1989; 1 (2): 270–80.

30. Cho K, van Merrienboer B, Gulcehre C, *et al*. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014); 2014; 1724–34.
31. Breiman L. Random forests. *Machine Learning* 2001; 45 (1): 5–32. doi : 10.1023/A : 1010933404324.
32. Rumelhart DE, Geoffrey EH, Williams RJ. Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL, PDP Research Group, eds. *Parallel Distributed Processing*, *Volume* 1: Explorations in the Microstructure of Cognition: Foundations. Cambridge, MA: MIT Press; 1986: 318–62.
33. Ngufor C, Houten HV, Caffo BS, *et al*. Mixed Effect Machine Learning: a framework for predicting longitudinal change in hemoglobin A1c. *J Biomed Inform* 2019; 89: 56–67.
34. Choi E, Schuetz A, Stewart WF, *et al*. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc* 2017; 24 (2): 361–70.