Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records

Binh P. Nguyen[a,1,*], Hung N. Pham[b,1], Hop Tran[a,1], Nhung Nghiem[c], Quang H. Nguyen[b], Trang T.T. Do[d], Cao Truong Tran[e], Colin R. Simpson[f,g]

[a] School of Mathematics and Statistics, Victoria University of Wellington, Kelburn Parade, Wellington 6140, New Zealand
[b] School of Information and Communication Technology, Hanoi University of Science and Technology, 1 Dai Co Viet Road, Hanoi 100000, Vietnam
[c] Department of Public Health, University of Otago, 23A Mein Street, Wellington 6021, New Zealand
[d] Institute for Infocomm Research, Agency for Science, Technology and Research, 1 Fusionopolis Way, Singapore 138632, Singapore
[e] Faculty of Information Technology, Le Quy Don Technical University, 236 Hoang Quoc Viet Street, Hanoi 100000, Vietnam
[f] Faculty of Health, Victoria University of Wellington, Kelburn Parade, Wellington 6140, New Zealand
[g] Usher Institute, The University of Edinburgh, Edinburgh, EH89AG, United Kingdom

ARTICLE INFO

Article history:
Received 7 March 2019
Revised 17 August 2019
Accepted 27 August 2019

Keywords:
Electronic health records
Incidence
Onset
Prediction
Type 2 diabetes mellitus
Wide and deep learning

ABSTRACT

Objective: Diabetes is responsible for considerable morbidity, healthcare utilisation and mortality in both developed and developing countries. Currently, methods of treating diabetes are inadequate and costly so prevention becomes an important step in reducing the burden of diabetes and its complications. Electronic health records (EHRs) for each individual or a population have become important tools in understanding developing trends of diseases. Using EHRs to predict the onset of diabetes could improve the quality and efficiency of medical care. In this paper, we apply a wide and deep learning model that combines the strength of a generalised linear model with various features and a deep feed-forward neural network to improve the prediction of the onset of type 2 diabetes mellitus (T2DM).
Materials and methods: The proposed method was implemented by training various models into a logistic loss function using a stochastic gradient descent. We applied this model using public hospital record data provided by the Practice Fusion EHRs for the United States population. The dataset consists of de-identified electronic health records for 9948 patients, of which 1904 have been diagnosed with T2DM. Prediction of diabetes in 2012 was based on data obtained from previous years (2009–2011). The imbalance class of the model was handled by Synthetic Minority Oversampling Technique (SMOTE) for each cross-validation training fold to analyse the performance when synthetic examples for the minority class are created. We used SMOTE of 150 and 300 percent, in which 300 percent means that three new synthetic instances are created for each minority class instance. This results in the approximated diabetes:non-diabetes distributions in the training set of 1:2 and 1:1, respectively.
Results: Our final ensemble model not using SMOTE obtained an accuracy of 84.28%, area under the receiver operating characteristic curve (AUC) of 84.13%, sensitivity of 31.17% and specificity of 96.85%. Using SMOTE of 150 and 300 percent did not improve AUC (83.33% and 82.12%, respectively) but increased sensitivity (49.40% and 71.57%, respectively) with a moderate decrease in specificity (90.16% and 76.59%, respectively).
Discussion and conclusions: Our algorithm has further optimised the prediction of diabetes onset using a novel state-of-the-art machine learning algorithm: the wide and deep learning neural network architecture.

1. Introduction

Diabetes is responsible for considerable morbidity, healthcare utilisation and mortality in both developed and developing countries. Globally, in 2017 it was estimated that 425 million people had diabetes – this is predicted to increase to 629 million by the end of 2045 [1]. Type 2 diabetes mellitus (T2DM) is the most common type of diabetes (95%) in the United States (US) [2]. In the US, more than 30 million people had diabetes in 2017 [1]. The high costs of hospital treatment and the high rate of readmission associated with diabetes means that early prevention and effective treatment is crucial [3]. The early prediction of the onset of

[*] Corresponding author.
E-mail address: b.nguyen@vuw.ac.nz (B.P. Nguyen).
[1] These authors contributed equally.

https://doi.org/10.1016/j.cmpb.2019.105055
0169-2607/© 2019 Elsevier B.V. All rights reserved.

diabetes using routinely available data such as electronic health records (EHRs) is therefore important [4].

EHRs are relatively complete electronic systems that have the potential to store information from millions of patients across many healthcare institutions, including patient demographics, medical data (e.g., diagnoses, laboratory tests and medications), clinical notes and so on [5,6]. In the past, EHRs were used by doctors, healthcare practitioners and public health workers to store and extract patients' information for clinical care [7]. The secondary use of EHR data for tool development aims to assist healthcare practitioners and policy makers to initiate or modify interventions, understand disease progress and introduce or improve policies to help prevent disease [8]. Patient information in EHRs is highly varied with dimensions, class imbalanced data (i.e., a heterogeneous sample of diabetic and non-diabetic patients) [4] and missing data [6], making it difficult to develop efficient analytic models using classical statistical analysis methods [9]. The availability of electronic health records (EHRs) along with advances in hardware (Central Processing Units (CPUs) and Graphical Processing Units (GPUs)) and computer algorithms (machine learning and especially its sub-field deep learning) make it possible to predict disease onset with high accuracy. With respect to diabetes, most studies utilising EHRs used and compared the performance of common machine learning algorithms (*k*-Nearest Neighbors, Naive Bayes, Decision Tree, Random Forest, Support Vector Machine, and Logistic Regression) in prediction of diabetes progression [10–17].

Deep learning algorithms have been used in recent years for the prediction of the onset of diseases based on secondary uses of EHRs. With respect to healthcare research, deep learning models can outperform classical machine learning methods which require more manual feature engineering [6]. Moreover, longitudinal event and continuous monitoring characteristics data from EHRs allows training of complex and challenging deep learning models [6]. Compared with statistical models for the prediction of the onset of diabetes using risk factors (logistic regression [18]) and patient mortality using hazard ratios (survival analysis [19]), and classical machine learning (decision tree, random forest and support vector machine [20]), deep learning is capable of automatically learning represented features from input data and subsequently reduces the feature engineering [21]. To attain state-of-the-art performance with less computational resource, a wide and deep learning framework was developed by Google to achieve both memorisation and generalisation [22]. Memorisation is learning a wide set of crossed-product feature transformations representing the correlation between the co-occurrence of a feature pair and the target label. Generalisation is obtained by matching different features that are close to each other in an embedding space generated by a deep feed-forward neural network. In this framework, the wide part accounts for a generalised linear model and the deep element represents a feed-forward neural network. By combining the advantages of both components, this framework is able to use a data structure which is highly varied and complicated. To the best of our understanding, there has been little previous work which has used deep learning approaches to develop risk scores using large healthcare data [23–27].

Miottothe et al. [8] developed a novel unsupervised deep learning algorithm (Deep Patient) to predict the future of patients using 700,000 records from the Mount Sinai EHRs. They used demographic information (age, sex and race), clinical notes as ICD-9 codes, medical prescriptions, procedures and laboratory tests. They designed a multi-layer deep representation neural network optimised with stochastic gradient descent to a local unsupervised criterion. Their model was tested using 76,214 patients comprising 78 diseases. The prediction of T2DM with complications within one year using AUC score was 90.7%. The algorithm was found to improve the prediction of various diseases in EHRs and other tasks such as clinical trial testing and treatment suggestions.

In recent work, Pham et al. [27] introduced a deep dynamic neural network framework (DeepCare) that performed various tasks including assessing patient trajectories and predicting future disease outcomes. The dataset contained more than 12,000 patients between 2002 and 2013 with 7191 patients selected. The dataset was divided into three parts: 67% for parameter estimation, 16.5% for tuning, and 16.5% for testing. The performance of DeepCare using max-pooling on the diabetes dataset was a F-score of nearly 60%.

One of the most important applications of using secondary EHRs is the development of web-based tools or software for the prediction of future outcomes. One of the examples of these online tools and software is QDiabetes^{TM}-2018 [28], an algorithm developed using Cox proportional hazards models by ClinRisk Ltd using information from QResearch database in UK (https://www.qresearch.org). QDiabetes is a risk prediction algorithm which calculates an individual's risk of T2DM for the next 10 years for people aged 25 to 84 years taking account of their individual risk factors (age, sex, ethnicity, clinical values and diagnoses) [29]. This tool is used to predict the risk of developing T2DM and integrated in doctors' computer systems with an average receiver operator curve statistic of 0.85 for women and 0.83 for men.

In summary, compared with classical machine learning models, deep learning can extract useful information from EHRs by learning features related to diabetes outcomes and therefore help in the targeting of people who are likely to develop the disease so that they can change their lifestyles. This information is important for developing tools and software for secondary uses of EHRs. In this study, we applied a wide and deep learning approach to predict the onset of type 2 diabetes mellitus using the Practice Fusion EHR dataset and compared the performance of this approach with a machine learning approach used by Pimentel et al. [17]. The wide and deep learning approach has been increasingly used for clinical risk prediction and classification. It is anticipated that predictive modelling using data from EHRs will drive personalised medicine resulting in improved healthcare quality. This information is important for development of the potential tools to assist health practitioners (doctors/clinicians) with the prognosis of diabetes disease and policy makers with creating suitable interventions to reduce the burden of diabetes.

2. Material and methods

2.1. Data source

We used a publicly available EHR dataset from the United States released by Practice Fusion in 2012 for a data science competition and compared our model performance with another study by Pimentel et al. [17] who applied a random forest with temporal features and feature selection for onset T2DM prediction using this dataset. The dataset consisted of de-identified electronic health records of 9948 patients, with 1904 diagnosed with T2DM over a four-year period (2009–2012). The dataset also included doctor transcripts with diagnoses, laboratory tests and medications. To prevent biases on the prediction of diabetes, direct diabetes-related information in the dataset was removed by Practice Fusion to make the classification challenge more difficult for the competition with some additional modifications. The first modification was to exclude patients that have diagnoses for diabetes complications without a basic diagnosis of T2DM. The second modification was to remove ICD-9 codes (250, 250.0, 250.*0, 250.*2, 357.2, and 362.0*). The third modification was to remove diabetes medications from the medication records. The final modification was to

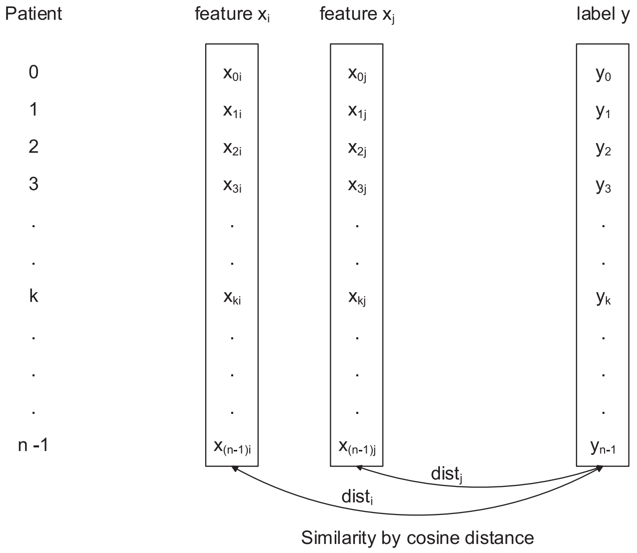


Fig. 1. Calculating the similarity of each feature column x_i and label (target) column y by measuring cosine distance. If the feature column x_i is more related to the column y (cos_i goes to 1) then it is a good feature column to be selected.

Table 1
Special diseases and their ICD-9 codes.

Special diseases	ICD-9 codes
heartDisease	410–414, 420–425, 427, 429, 745, 746
CHF	426
Stroke	430, 431, 433–436, 997.02
sleepApnea	727.23, 780.57
gestDiab	648.8
polyOvary	256.4
frozenShoulder	726.0
Hemochr	275.03
Hepatitis	070.2, 070.3
kidneyFailure	584, 585
Dementia	331, 290, 294, 797
Acanthosis	701.2
Blindness	369
preDiabetes	790.29
sDysfunction	302.7
EssentialHypertension	401, 401.0, 401.1, 401.9
MixedHyperlipidemia	272.2

remove laboratory tests that identified glucose or insulin related tests.

2.2. Feature extraction and selection

The dataset was processed using feature extraction and selection. This process can be used to reduce the dimensions of the dataset by selecting main and important features. We grouped 1312 features into (1) basic and fixed features (age, sex, body mass index (BMI) and blood pressure), (2) adjustable features (diagnostic features based on ICD-9 codes, medication and laboratory tests) with labels encoded into binary vectors corresponding to three embeddings and (3) crossed features by selecting top diagnosis features to cross with top medication features. Embeddings are a mapping of a categorical variable to a vector of continuous numbers which are useful for reducing the dimensionality of categorical variables and meaningfully represent categories in the transformed space. Three types of features (diagnoses, medications and laboratory tests) were label encoded into binary vectors. Each type of features was subsequently mapped into corresponding embeddings using a linear layer of neural network from the deep part of the learning model.

Steps taken for feature extraction and selection are described below:

(1) Outliers of BMI, height and weight variables were cleaned for each patient.

(2) BMI-related features were generated from BMI data for each patient including BMI median, minimum, maximum values, *isOverweight, isObese* and difference between BMI min and BMI max values. *isOverweight* and *isObese* features were determined based on some cut-offs (ranges) of BMI median value of each patient. BMI data was already calculated from height and weight of each patient. Each patient could have more than one BMI data record. This data was used to generate 6 BMI-related features.

(3) Systolic and diastolic blood pressures were calculated to generate blood pressure features (median, min and max values), difference between min and max values of blood pressures, whether a patient had high blood pressure (HBP) in the first, second stage or not. These HBP features (1/0 ~ yes/no) were determined based on a range of threshold values of blood pressures for each patient which were ranked in medical research.

(4) Diagnosis data was analysed to extract ICD-9 codes, excluding the data in 2012, there was a total of 3903 different codes for all the patients. These diagnosis features were encoded with labels as a sparse binary vector (each ICD-9 code was labelled with value 1, otherwise 0) for each patient.

To reduce the dimensions of diagnosis feature vectors for all patients, each column vector corresponding to one column ICD-9 code feature was assigned an important feature score by measuring cosine distance between column feature vector and label column vector (target) (labelled as 1/0 corresponding with yes/no diabetes, respectively). Fig. 1 illustrates the method used to assess the important feature score of the column x_i by a measure that calculates the cosine of the angle between the feature column x_i and column label y. All feature columns associated with diagnosis features were assessed using the important feature score. A histogram of the important feature scores was used to select a threshold value. Feature columns with important feature scores greater than this threshold were selected as important features to create a new set of features. Less important features containing ICD-9 codes were discarded to reduce the dimensions of the diagnosis feature vector.

(5) Medication data was analysed to extract the names of medication using data from 2009 to 2011. There were 2553 types of medication extracted, and similar to diagnosis features, the labels of medication were encoded as a sparse binary vector for each patient. A similar method to diagnosis feature vector was applied to reduce the dimensions of medication feature vector.

(6) Laboratory test data was analysed to extract information on laboratory tests completed for each patient illustrated by a HL7 message. Each HL7 message contains one or more segments in which each segment consists of one or more composites (fields). A total of 334 laboratory tests were reported for all patients. Laboratory test features vector was selected in the same way as diagnosis features (Fig. 1 and step 4) and medication features (step 5).

(7) Special features were created by analysing ICD-9 codes corresponding to some groups of special diseases such as heart disease, cardiomyopathy, kidney failure, blindness and so on. There were total of 17 such special features including risk factors for T2DM (Table 1).

(8) Diagnosis descriptions were grouped by ICD-9 codes to create new properties. There were 19 groups of ICD-9 codes to be used (Table 2).

(9) Crossed-product features were created from diagnosis description features and medication name features (Fig. 2). Diagno-

Table 2
Groups of diagnosis descriptions and their ICD-9 codes.

Diagnosis group	ICD-9 codes
neoplasms	140–149, 200–239
endocrine	240–279
blood	280–289
mental health	290–299, 300–319
nervous	320–359
sense	360–389
circulatory	390–399, 400–459
respiratory	460–499, 500–519
digestive	520–579
genitourinary	580–599, 600–629
pregnancy	630–679
skin	680–699, 700–709
musculoskeletal	710–739
congenital	740–759
perinatal	760–779
symptoms or ill-defined	780–799
injuries	800–899, 900–999
suppl	E, V
infectious	others

sis description features associated with medication name features creates crossed product features that have the potential to better generalise the data. Only top diagnosis description features and top medication name features were used to perform feature crossing. In our setting, 27 diagnosis description features from over 500 observations were used to cross with 33 medication name features over 200 observations. Crossed values were encoded into binary vectors which became diagnosis medication crossed features.

(10) Basic features (age and sex) were extracted for each patient.

The above properties were combined to create 1312 features for each patient (the number of features could be adjustable with selected parameters in the current setting). Among the total of 9948 patients (43% male and 57% female) aged from 21 to 93 years old, 1890 patients (19%) had a diagnosis of diabetes.

2.3. Wide and deep model architecture

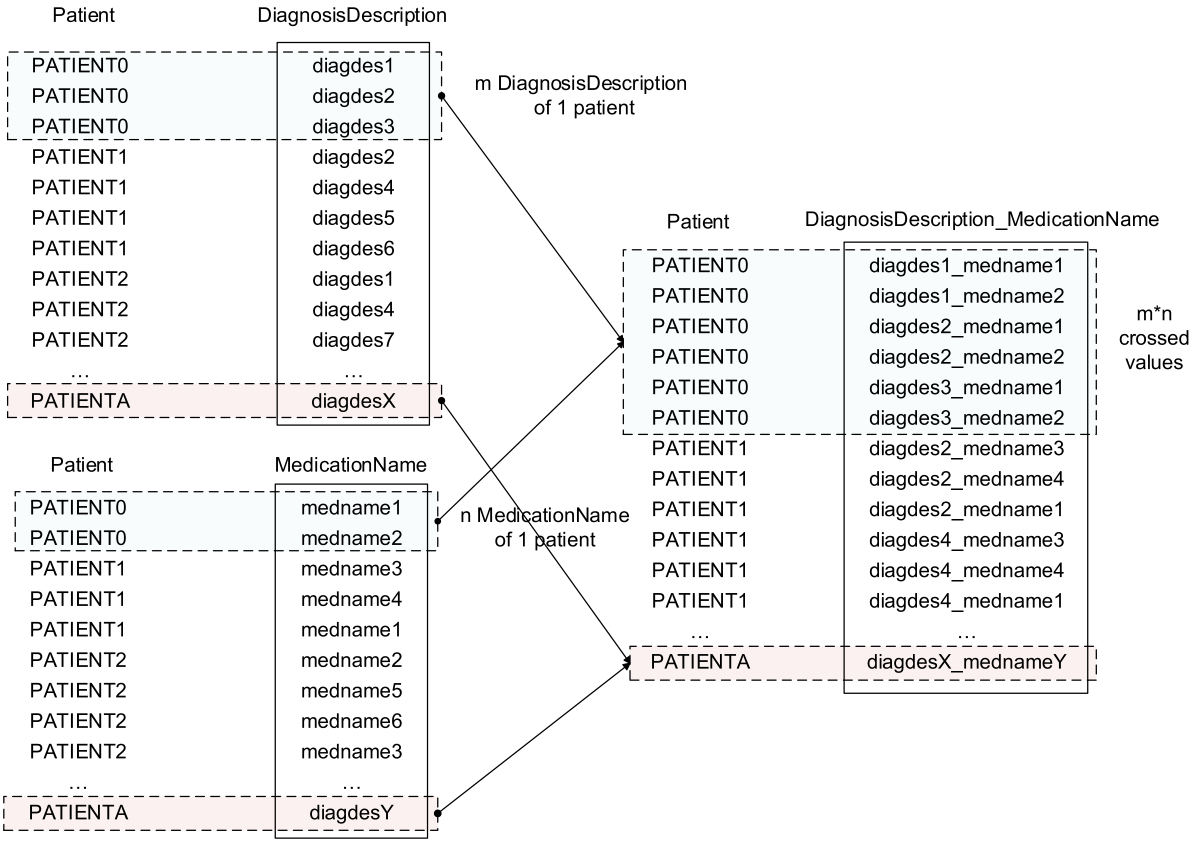
In this study, we developed an algorithm for the prediction of diabetes onset based on the wide and deep learning framework [22]. This framework was used since it was capable of combining the benefits of memorisation and generalisation with less feature engineering, useful for analysing EHR data.

The dataset was divided into a development set (70%) and a testing set (30%), of which the development set (adopting a similar approach to Pimentel et al. [17]) was divided into 10 folds (9 for training and 1 for validation) so that we could compare out results with Pimentel et al. [17]. The workflow for predicting the onset of T2DM using the wide and deep model is illustrated in Fig. 3.

The 1312 patient features were processed by a wide and deep architecture (Fig. 4) which consists of a wide component and a deep component.

The wide component was a generalised linear model used for large-scale regression and classification problems [22]. This component is responsible for memorising feature interactions.

The deep component was a deep neural network that can generalise better to new features using low-dimensional dense embedding. In our framework, this component was composed of two types of layers: embedding and hidden layers. *Embedding layers* included three embeddings corresponding to three groups of features: (1) diagnosis with 151 input features, (2) medication with



Patient	DiagnosisDescription
PATIENT0	diagdes1
PATIENT0	diagdes2
PATIENT0	diagdes3
PATIENT1	diagdes2
PATIENT1	diagdes4
PATIENT1	diagdes5
PATIENT1	diagdes6
PATIENT2	diagdes1
PATIENT2	diagdes4
PATIENT2	diagdes7
...	...
PATIENTA	diagdesX

Patient	MedicationName
PATIENT0	medname1
PATIENT0	medname2
PATIENT1	medname3
PATIENT1	medname4
PATIENT1	medname1
PATIENT2	medname2
PATIENT2	medname5
PATIENT2	medname6
PATIENT2	medname3
...	...
PATIENTA	diagdesY

Patient	DiagnosisDescription_MedicationName
PATIENT0	diagdes1_medname1
PATIENT0	diagdes1_medname2
PATIENT0	diagdes2_medname1
PATIENT0	diagdes2_medname2
PATIENT0	diagdes3_medname1
PATIENT0	diagdes3_medname2
PATIENT1	diagdes2_medname3
PATIENT1	diagdes2_medname4
PATIENT1	diagdes2_medname1
PATIENT1	diagdes4_medname3
PATIENT1	diagdes4_medname4
PATIENT1	diagdes4_medname1
...	...
PATIENTA	diagdesX_mednameY

m DiagnosisDescription of 1 patient
n MedicationName of 1 patient
m*n crossed values

Fig. 2. Generating new crossed features from diagnosis description and medication name using crossed product. Crossed values are label encoded to binary vector of crossed features.

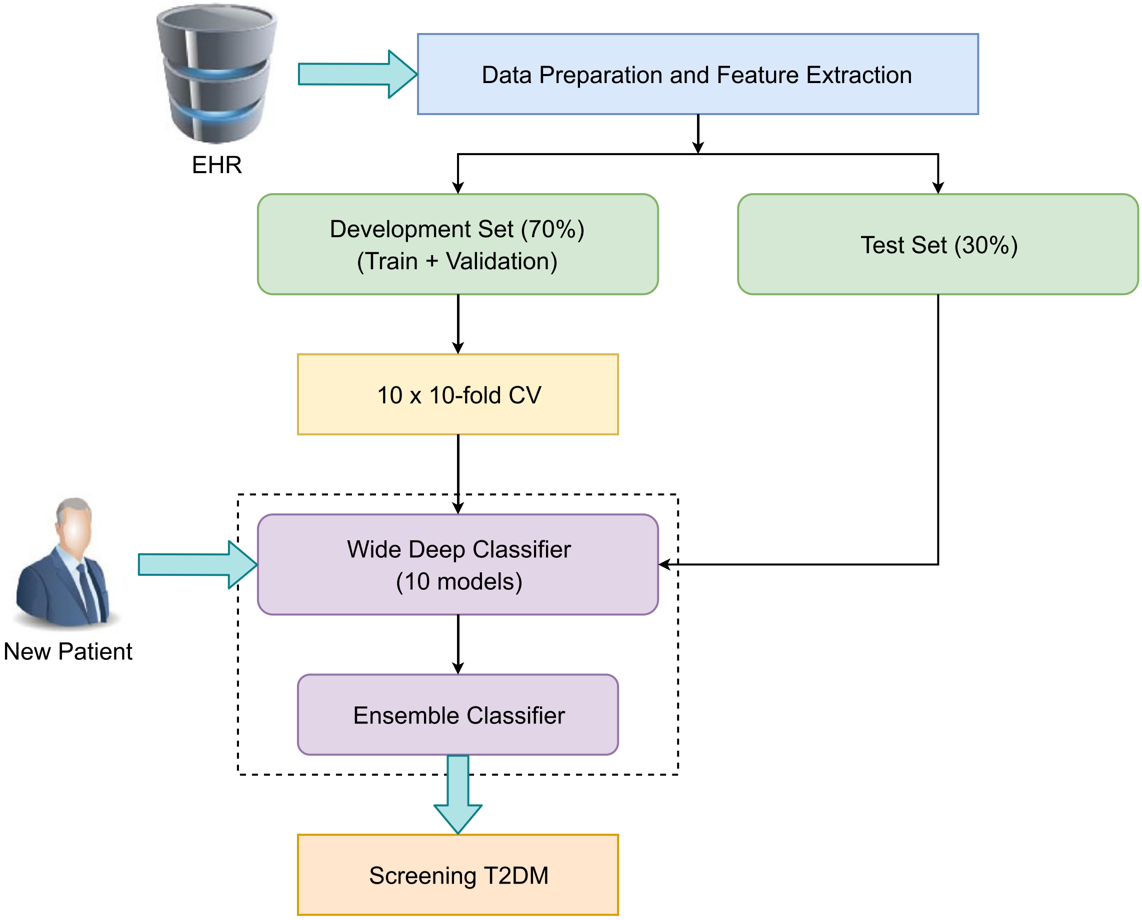


Fig. 3. Summary of the workflow for predicting the onset of diabetes.

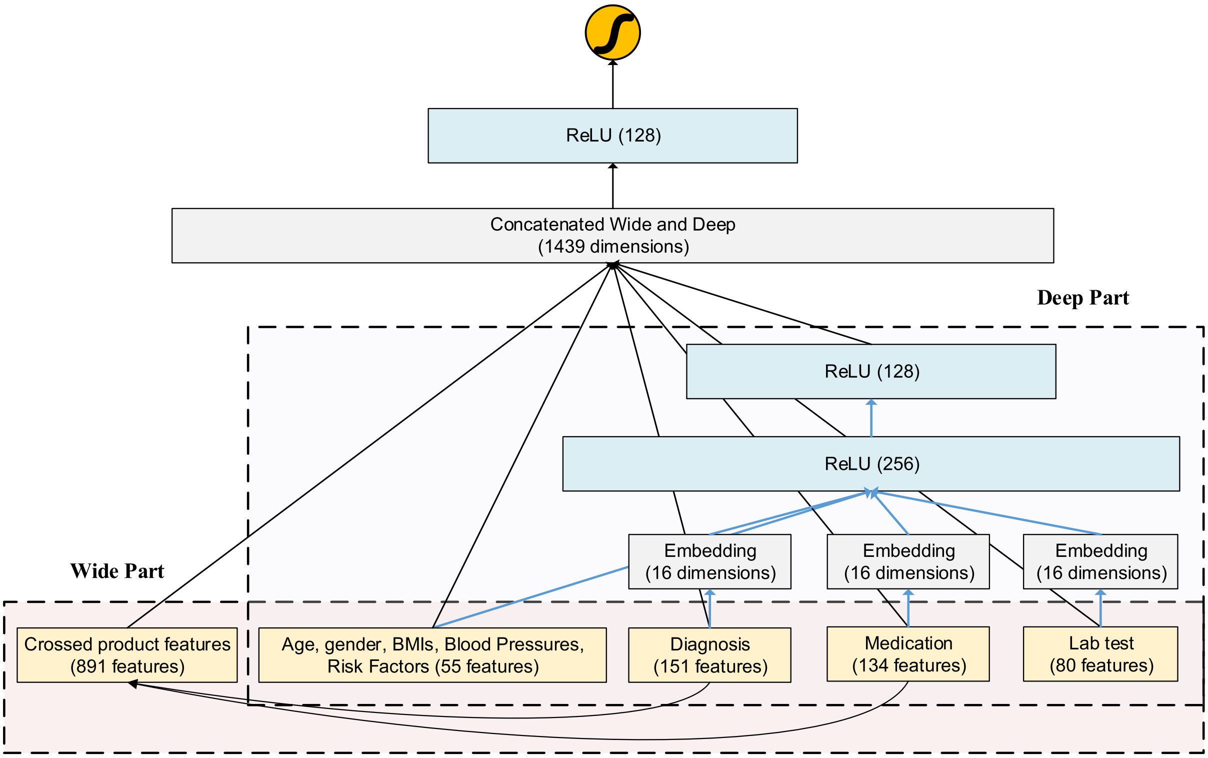


Fig. 4. Wide and deep model structure for predicting the onset of diabetes.

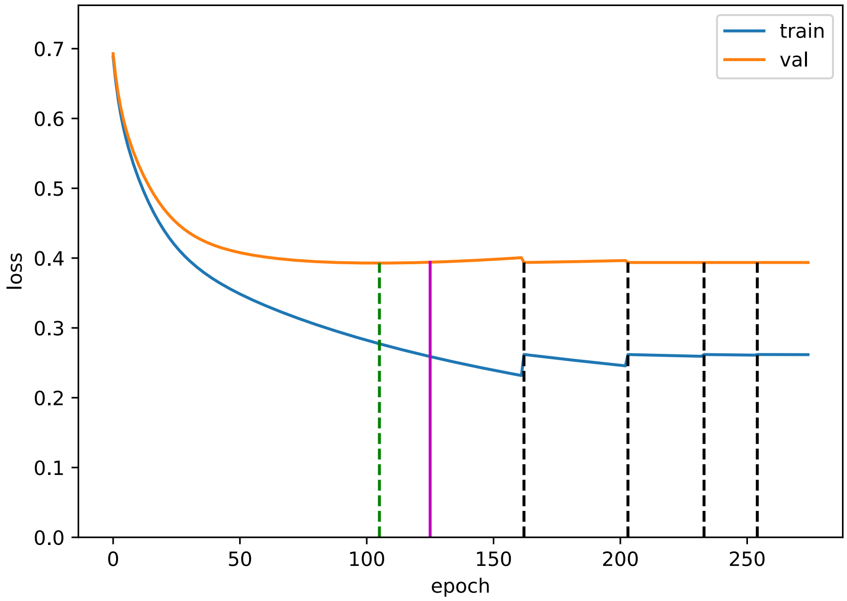


Fig. 5. Training loss and validation loss when training Model 1 in Table 3 with 5 levels of learning rate (see the text for details).

134 input features, and (3) laboratory test with 80 input features. We applied one linear layer for each embedding for learning from a sparse binary vector to a dense 16-dimensional vector. We applied *hidden layers* with two hidden layers of 256 and 128 neurons.

All features were put into the wide part which included the crossed features joined with the output of the deep part in the last layer to form a 1439-dimensional vector. The final output layer of the framework was a linear 128-to-1 layer with sigmoid activation function. The activation function in other layers was the rectified linear unit (ReLU) [30].

2.4. Experimental settings

True positive (TP), true negative (TN), false positive (FP) and false negative (FN) were used to measure the performance of classifiers using the following evaluation metrics:

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

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

Sensitivity was defined as the proportion of subjects with diabetes that were correctly classified as having diabetes. Specificity was defined as the proportion of people without diabetes that were correctly classified. Accuracy was defined as the proportion of all subjects that were correctly classified.

Using 10-fold cross-validation (CV), 10 predictive models corresponding to 10 different sets of training and validation data were built. The stochastic gradient descent (SGD) optimiser and the binary cross entropy loss function [30] were used in our model training. Each model was trained with five levels of learning rates (1e-3, 5e-4, 1e-4, 5e-5 and 1e-5) corresponding to five patience values (40, 40, 30, 20 and 20 epochs). A patience is the number of epochs to wait if the validation loss does not decrease before moving to the next learning rate or stop if the last learning rate has already been used. When moving to the next learning rate, the snapshot corresponding to the current least validation loss was loaded. After stopping, the snapshot corresponding to the best validation loss was used as the final model.

Table 3
Results as percentages obtained from the test set using 10 models from a 10-fold stratified cross validation and the final ensemble model.

Model	AUC	Sensitivity	Specificity	Accuracy
1	83.31	29.59	96.27	83.51
2	82.90	25.56	97.22	83.52
3	83.34	25.56	97.05	83.38
4	83.59	31.52	95.81	83.51
5	82.96	30.64	95.36	82.98
6	82.72	26.44	96.27	82.91
7	82.43	28.37	96.39	83.38
8	83.40	32.74	96.06	83.95
9	82.79	37.30	94.53	83.58
10	83.95	34.32	95.89	84.12
Ensemble	**84.13**	**31.17**	**96.85**	**84.28**

Fig. 5 shows the train loss (in blue) and validation loss (in orange) when training a model (Model 1 in Table 3). Dashed vertical lines denote the epochs at which a new learning rate was used because the corresponding patience value was reached. The magenta vertical line denotes the epoch at which the validation loss is minimum, and the snapshot at that epoch was used as the trained model.

Samples from the testing set were subsequently predicted by 10 optimal models after training. The performance of each model was evaluated on the testing set using the following metrics: accuracy, AUC, sensitivity and specificity. An ensemble model was created by calculating the mean of the output probabilities from the above 10 best models and compared with a threshold (0.5) for diabetes determination. This ensemble model was used as a final predictive model of the onset of T2DM.

As the dataset was imbalanced (only 19% subjects with diabetes), Synthetic Minority Over-sampling Technique (SMOTE) [31] was used in each CV training fold to analyse the performance when synthetic examples for the minority class were created. Similar to Pimentel et al. [17]), we used SMOTE of 150% and 300%, in which 300% means that three new synthetic instances were created for each minority class instance. This results in the approximated diabetes:non-diabetes distributions in the training set of 1:2 and 1:1, respectively.

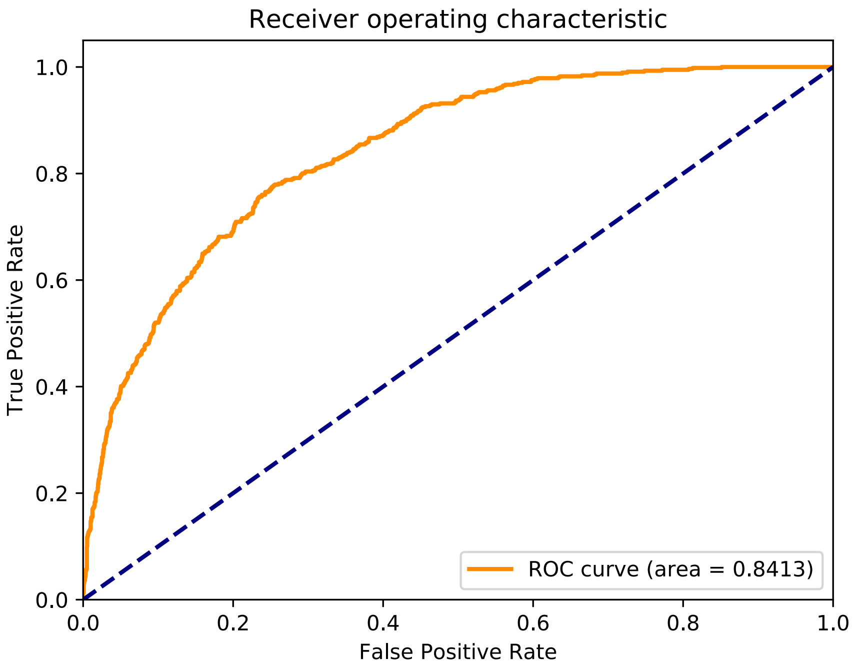


Fig. 6. ROC curve corresponding to the ensemble model in Table 3.

Table 4
Results as percentages obtained from the test set of 10 different data partitions and comparison with results from Pimentel et al. [17].

SMOTE	Model	AUC	Sensitivity	Specificity
0%	Our ensemble model	**84.01**	**29.12**	96.18
	Best model in [17]	83.19	16.07	**99.28**
150%	Our ensemble model	83.33	**49.40**	90.16
	Best model in [17]	**84.22**	29.19	**96.42**
300%	Our ensemble model	82.12	**71.57**	76.59
	Best model in [17]	**84.11**	36.23	**93.77**

3. Results and discussion

Table 3 shows the performance obtained from the test set using 10 models from a 10-fold stratified cross validation and the final ensemble model. The ensemble model produced an AUC of 84.13% (Fig. 6) which is higher than that of each individual model. This means the modeling averaging ensemble is more robust and produces better performance on average than a single model.

We further tested the algorithm using 10-fold cross-validation and the final ensemble model was selected for assessing the performance from the test set of 10 different data partitions (10 sets of 10-fold cross-validation). The model not using SMOTE showed higher AUC score than other models using SMOTE (150% and 300%) (Table 4). Using SMOTE with variation of minority and majority ratio only improved sensitivity but did not improve other performance metrics. The ensemble model without using SMOTE, on average, obtained an AUC score of 84.01%, sensitivity score of 29.12% and specificity of 96.18% (Table 4).

Results on the model comparisons (Table 4) indicated that the ensemble model without SMOTE performed better than the ensemble model using SMOTE (150% and 300%) with higher AUC score (by 0.68% and 1.89%, respectively) and higher specificity (by 6.02% and 19.59%, respectively). However, the models using SMOTE increased sensitivity (24.34% and 42.45%, respectively). These results are in contrary to another study by Pimentel et al. [17] who reported that the performance of their random forest model using SMOTE (150% and 300%) remarkably improved AUC and sensitivity scores. In another study, Alghamdi et al. [4] showed that using SMOTE with ensemble machine learning remarkably improved the model performance for the prediction of incidence of T2DM. The use of an ensemble-based approach with SMOTE has been found to obtain high accuracy of predicting the incidence of diabetes in metropolitan Detroit, Michigan in the US [4].

Compared with the machine learning approach by Pimentel et al. [17] who applied random forest with temporal features and feature selection using the same dataset and experimental settings as ours, the performance on the testing set of our model was higher (AUC score (84.01%) and sensitivity score (29.12%)) than their model (AUC score (83.19%) and sensitivity score (16.07%)) when not using SMOTE. The higher sensitivity score in our model would be a compensation for being able to better predict the proportion of subjects with diabetes that were correctly classified as having diabetes. Our results (Table 4) highlighted some important implications. Both sensitivity and specificity are mostly useful where the target class (positive) is often smaller with a substantial consequence if incorrectly classified. Therefore, the trade-off between sensitivity and specificity are to be carefully taken into consideration to get a good balance. Our ensemble model using SMOTE 300%, when compared to other models, had better sensitivity with a modest reduction in specificity. A prediction model for type 2 diabetes with good sensitivity will reduce the risk of unnecessary interventions and therapy being given to those with low future risk. The trade off of using the SMOTE 300% model however is that a lowering in specificity may result in some people without type 2 diabetes will screen positive and therefore potentially receive unnecessary further investigation. From a clinical perspective, clinicians, when making shared decisions with patients, will be more confident with using a prediction model that has high sensitivity [32].

For a cursory comparison, it is worth mentioning other studies using other machine learning approaches to predict T2DM despite differences in datasets and experimental settings. Mani et al. [10] explored different machine learning algorithms with feature selection to evaluate the risk of T2DM development from six months to one year. They used a de-identified EHR dataset managed by the Vanderbilt University Medical Centre. This dataset included demographic variables (age, sex and race), clinical notes (body mass index (BMI) and diabetes status) and laboratory tests of 2280 patients with 10% diagnosed with T2DM. To perform this predictive modelling task, they used various forms of classifiers (linear, decision tree-based, kernel-based and sampled-based). The population was randomly divided into two groups for model development (50%) and validation (50%). A five-fold nested cross-validation framework was implemented to optimise the parameters of the classification algorithm. The performance of the final model was assessed by taking an averaging of the best *k* models. The highest accuracy was reported with an AUC score greater than 80% for the prognosis of T2DM at 180 days and 365 days.

Razavian et al. [11] applied logistic regression with L1 regularisation for predicting risk factors associated with T2DM between 2009 and 2011 using an electronic claim dataset provided by Independence Blue Cross insurance company in Pennsylvania, the United States. The dataset contained claim information (administrative papers, pharmaceutical records, healthcare utilisation and laboratory tests) of 793,153 cases who matched the selection criteria. The dataset was randomly divided into a training set (67%) and a testing set (33%) using a five-fold cross-validation. The final model improved the prediction accuracy with an AUC score of 80% and was able to predict risk factors associated with the onset of diabetes.

Anderson et al. [12] used machine learning algorithms (multivariate logistic regression and random forest) for exploring the detection and screening of T2DM for the US population using the same dataset used in our study. They compared three separate models: (1) a full model which included medical prescriptions and conventional risk scores, (2) a restricted model similar to (1) but excluding medical notes, and (3) a conventional model containing some conventional risk scores with interactions (BMI, age, sex, smoking and hypertension). For logistic regression, the performance reported as AUC scores were 84.9%, 83.2%, and 75.0%, respectively. For random forest, the AUC scores were 81.3%, 79.6%, and 74.8%, respectively. The inclusion of EHR phenotyping significantly improves the performance of detection and screening of T2DM in this study.

Brisimi et al. [15] developed predictive models for diabetes-related hospitalisations based on EHRs of 40,921 patients from 2001 to 2012 from the largest safety net hospital in New England. Their new joint clustering/classification method achieved an AUC of 89%.

Zou et al. [16] applied machine learning techniques (decision tree, random forest and neural network) to predict diabetes mellitus using hospital physical examination data containing 14 attributes in Luzhou, China. A five-fold cross-validation was used to assess the models. The highest accuracy was reported using random forest with an accuracy greater than 80%.

Compared to other machine learning systems and frequentist statistics presented above, the primary advantage of the wide and deep model is that it incorporates manual feature engineering through the selection of features and the design of the crossed features going into the wide part and auto feature engineering by using deep neural networks in the deep part.

The accuracy of our model was affected by several factors. One of the main challenges in our study is the high dimensions and sparsity of the dataset. As many machine learning algorithms are generally unable to handle insufficient and imbalanced data where the classes are not equally presented, it is not surprising that our approach of using wide and deep learning is severely impacted by the same issue. In our model setting, 27 diagnosis description features and 33 medication features were selected to create crossed-product features but the number of observations in these groups were imbalanced due to a heterogeneous sample of patients having a diagnosis of diabetes and non-diabetes and some features contained missing values or incorrect information. Work by Habibi et al. [33] showed that decision tree could be used to screen T2DM without using laboratory tests. Indeed, in addition to classical regression models, there is some successful research using deep learning for improving the accuracy of diabetes risk prediction [8,34]. However, our work is the first attempt to apply wide and deep learning for the prediction of the onset of diabetes using EHRs. Although our model achieves higher predictive power compared to classical machine learning methods, similar to other deep learning models [8,34], the wide and deep model would not be able to predict some important risk factors incorporated into the model.

4. Conclusions

In this study, we proposed a wide and deep learning neural network architecture for the prediction of the onset of diabetes using a publicly available EHR dataset. Our ensemble model improved AUC and specificity risk scores and substantially improved sensitivity for predicting T2DM onset compared with other machine learning algorithms which used the same dataset and experimental settings [17]. In the future, we will incorporate an auto feature selection method to design the crossed features and select the features for the wide part of the model. Using a more sophisticated embedding method for the deep part may be another way to improve the performance of the model.

Declaration of Competing Interest

All authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

B. P. Nguyen and Q. H. Nguyen gratefully acknowledge the support of NVIDIA Corporation for the donation of the GPUs used for this study.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2019.105055.

References

[1] I.D. Federation, IDF diabetes atlas, eighth ed., 2017, (http://diabetesatlas.org/IDF_Diabetes_Atlas_8e_interactive_EN/ Brussels, Belgium).

[2] R.L. Richesson, S.A. Rusincovitch, D. Wixted, B.C. Batch, M.N. Feinglos, M.L. Miranda, W.E. Hammond, R.M. Califf, S.E. Spratt, A comparison of phenotype definitions for diabetes mellitus, J. Am. Med. Inform. Assoc. 20 (e2) (2013) e319–e326, doi:10.1136/amiajnl-2013-001952.

[3] D.J. Rubin, Correction to: hospital readmission of patients with diabetes, Curr. Diabetes Rep. 18 (21) (2018) 1–9, doi:10.1007/s11892-018-0989-1.

[4] M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, S. Sakr, Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: the Henry Ford exercise testing (FIT) project, PLoS One 12 (7) (2017) 1–15, doi:10.1371/journal.pone.0179805.

[5] J.A. Casey, B.S. Schwartz, W.F. Stewart, N.E. Adler, Using electronic health records for population health research: a review of methods and applications, Ann. Rev. Public Health 37 (2016) 61–81, doi:10.1146/annurev-publhealth-032315-021353.

[6] C. Xiao, E. Choi, J. Sun, Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review, J. Am. Med. Inform. Assoc. 25 (10) (2018) 1419–1428, doi:10.1093/jamia/ocy068.

[7] W.R. Hersh, Adding value to the electronic health record through secondary use of data for quality assurance, research, and surveillance, Am. J. Manag. Care 13 (6) (2007) 277–278.

[8] R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, Deep patient: an unsupervised representation to predict the future of patients from the electronic health records, Sci. Rep. 6 (26094) (2016) 1–10, doi:10.1038/srep26094.

[9] B.A. Goldstein, A.M. Navar, M.J. Pencina, J.P.A. Ioannidis, Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review, J. Am. Med. Inform. Assoc. 24 (1) (2017) 198–208, doi:10.1093/jamia/ocw042.

[10] S. Mani, Y. Chen, T. Elasy, W. Clayton, J. Denny, Type 2 diabetes risk forecasting from EMR data using machine learning, in: Proceedings of the AMIA Annual Symposium, American Medical Informatics Association, 2012, pp. 606–615.

[11] N. Razavian, S. Blecker, A.M. Schmidt, A. Smith-McLallen, S. Nigam, D. Sontag, Population-level prediction of type 2 diabetes from claims data and analysis of risk factors, Big Data 3 (4) (2015) 277–287, doi:10.1089/big.2015.0020.

[12] A.E. Anderson, W.T. Kerra, A. Thames, T. Li, J. Xiao, M.S. Cohen, Electronic health record phenotyping improves detection and screening of type 2 diabetes in the general United States population: a cross-sectional, unselected, retrospective study, J. Biomed. Inform. 60 (2016) 162–168, doi:10.1016/j.jbi.2015.12.006.

[13] J.P. Anderson, J.R. Parikh, D.K. Shenfeld, V. Ivanov, C. Marks, B.W. Church, J.M. Laramie, J. Mardekian, B.A. Piper, R.J. Willke, D.A. Rublee, Reverse engineering and evaluation of prediction models for progression to type 2 diabetes: an application of machine learning using electronic health records, J. Diabetes Sci. Technol. 10 (1) (2016) 6–18, doi:10.1177/1932296815620200.

[14] T. Zheng, W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang, Y. Chen, A machine learning-based framework to identify type 2 diabetes through electronic health records, Int. J. Med. Inform. 97 (1) (2017) 120–127, doi:10.1016/j.ijmedinf.2016.09.014.

[15] T.S. Brisimi, T. Xu, T. Wang, W. Dai, I.C. Paschalidis, Predicting diabetes-related hospitalizations based on electronic health records, Stat. Methods Med. Res. (2018), doi:10.1177/0962280218810911.

[16] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang, Predicting diabetes mellitus with machine learning techniques, Front. Genet. 9 (2018) 515, doi:10.3389/fgene.2018.00515.

[17] A. Pimentel, A.V. Carreiro, R.T. Ribeiro, H. Gamboa, Screening diabetes mellitus 2 based on electronic health records using temporal features, Health Inform. J. 24 (2) (2018) 194–205, doi:10.1177/1460458216663023.

[18] P. Ruscitti, F. Ursini, P. Cipriani, V. Liakouli, F. Carubbi, O. Berardicurti, G. De Sarro, R. Giacomelli, Poor clinical response in rheumatoid arthritis is the main risk factor for diabetes development in the short-term: a 1-year, single-centre, longitudinal study, PLoS One 12 (7) (2017) 1–16, doi:10.1371/journal.pone.0181203.

[19] T. Kümler, G.H. Gislason, L. Køber, C. Torp-Pedersen, Diabetes is an independent predictor of survival 17 years after myocardial infarction: follow-up of the TRACE registry, Cardiovasc. Diabetol. 9 (22) (2010) 1–8, doi:10.1186/1475-2840-9-22.

[20] G.-M. Huang, K.-Y. Huang, T.-Y. Lee, J.T.-Y. Weng, An interpretable rule-based diagnostic classification of diabetic nephropathy among type 2 diabetes patients, BMC Bioinform. 16 (Suppl 1) (2015) S5:1–10, doi:10.1186/1471-2105-16-S1-S5.

[21] S. Purushotham, C. Meng, Z. Che, Y. Liu, Benchmarking deep learning models on large healthcare datasets, J. Biomed. Inform. 83 (2018) 112–134, doi:10.1016/j.jbi.2018.04.007.

[22] H.-T. Cheng, L. Koc, J. Harmsen, T. Shaked, T. Chandra, H. Aradhye, G. Anderson, G. Corrado, W. Chai, M. Ispir, R. Anil, Z. Haque, L. Hong, V. Jain, X. Liu, H. Shah, Wide & deep learning for recommender systems, in: Proceedings of the First Workshop on Deep Learning for Recommender Systems (DLRS 2016), ACM, 2016, pp. 7–10, doi:10.1145/2988450.2988454.

[23] Z. Liang, G. Zhang, J.X. Huang, Q.V. Hu, Deep learning for healthcare decision making with EMRs, in: Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2014), IEEE, 2014, pp. 556–559, doi:10.1109/BIBM.2014.6999219.

[24] T. Tran, T.D. Nguyen, D. Phung, S. Venkatesh, Learning vector representation of medical objects via EMR-driven nonnegative restricted Boltzmann machines (eNRBM), J. Biomed. Inform. 83 (2015) 96–105, doi:10.1016/j.jbi.2015.01.012.

[25] J. Futoma, J. Morris, J. Lucas, A comparison of models for predicting early hospital readmissions, J. Biomed. Inform. 56 (2015) 229–238, doi:10.1016/j.jbi.2015.05.016.

[26] E. Choi, M.T. Bahadori, J. Sun, J. Kulas, A. Schuetz, W. Stewart, RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism, in: D.D. Lee, M. Sugiyama, U.V. Luxburg, I. Guyon, R. Garnett (Eds.), Proceedings of the Advances in Neural Information Processing Systems, volume 29, Curran Associates, Inc., 2016, pp. 3504–3512.

[27] T. Pham, T. Tran, D. Phung, S. Venkatesh, Predicting healthcare trajectories from medical records: a deep learning approach, J. Biomed. Inform. 69 (2017) 218–229, doi:10.1016/j.jbi.2015.05.016.

[28] J. Hippisley-Cox, C. Coupland, Development and validation of QDiabetes-2018 risk prediction algorithm to estimate future risk of type 2 diabetes: cohort study, BMJ 359 (j5019) (2017) 1–18, doi:10.1136/bmj.j5019.

[29] J. Hippisley-Cox, C. Coupland, J. Robson, A. Sheikh, P. Brindle, Predicting risk of type 2 diabetes in england and wales: prospective derivation and validation of QDScore, BMJ 338 (b880) (2009) 1–15, doi:10.1136/bmj.b880.

[30] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, Adaptive Computation and Machine Learning Series, The MIT Press, 2016.

[31] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–â357, doi:10.1613/jair.953.

[32] W.M. Strull, B. Lo, G. Charles, Do patients want to participate in medical decision making? JAMA 252 (21) (1984) 2990–2994, doi:10.1001/jama.1984.03350210038026.

[33] S. Habibi, M. Ahmadi, S. Alizadeh, Type 2 diabetes mellitus screening and risk factors using decision tree: results of data mining, Glob. J. Health Sci. 7 (5) (2015) 304–310, doi:10.5539/gjhs.v7n5p304.

[34] H.N. Mhaskar, S.V. Pereverzyev, M.D. van der Walt, A deep learning approach to diabetic blood glucose prediction, Front. Appl. Math. Stat. 3 (14) (2017) 1–11, doi:10.3389/fams.2017.00014.