Predictive Diabetic Risk Modeling

David Huang[1], Nick Cheney[2], Susan Kovarik[3], Ellie Mehltretter[4]

QMIND – Queen's AI Hub
Queen's University, Kingston, Ontario K7L 3N6, Canada. Queen's

1 e-mail: 17dh11@queensu.ca
2 e-mail: 16nrc2@queensu.ca
3 e-mail: 17slk6@queensu.ca
4 e-mail: ellie.mehltretter@queensu.ca

Abstract: As diabetes prevalence continues to accelerate globally, methods to better monitor diabetes progression are critical in enabling effective preventative action and reducing the burden on local healthcare systems. Current prognostic models often prioritize between interpretability, by way of stratification on a small set of lab values, and predictive accuracy using deep learning methods on high dimensional data. The objective of this study is to develop a wholistic diabetes risk model that has strong predictive ability and maintains interpretability. Recurrent neural network models were developed on rich EMR data to predict the onset of 10 diabetes-related complications and time series forecasting of six clinically relevant lab tests was used for risk stratification following American Diabetes Association clinical guidelines. We achieved AUC scores greater than 85 for 5 out of ten complication onset models, while lab value forecasting for risk stratification using LSTM and ARIMA models achieved satisfactory RMSE values. In combination, our models provide a comprehensive understanding of the relative risk level for individuals with diabetes.

1. INTRODUCTION

1.1 *Motivation*

The global diabetes burden is expected to increase from 463 million people in 2019 to 578 million people by 2030 with developed countries seeing the greatest increase in prevalence rates [1]. In Canada, diabetes prevalence is expected to increase from 11,232,300 in 2020 to 13.6 million or 32% of all Canadians by 2030. Moreover, the increase in diabetes prevalence presents a significant burden on the health-care system. With the direct cost to the Canadian healthcare system expected to increase from 3.8 billion in 2020 to 4.9 billion by 2030 [2].

Thus, it is critical to develop improved monitoring methods to track the overall health status of those living with diabetes to reduce the diabetes burden on the healthcare system and to ensure preventative action can occur before development of life-threatening complications.

1.2 *Related Works*

Due to the complex and diverse pathophysiology of diabetes, the American Diabetes Association (ADA) recommends individualized treatment and medication plans [3]. As such several studies have focused on personalizing treatment by scoring, or stratifying, the relative health of diabetic patients using clinical test values. These stratification methods allow for better resource allocation, help clinicians better monitor the relative health of their patients and have shown to improve overall diabetes outcomes [4].

More recently, several prognostic machine learning models have been developed alongside the increased adoption of electronic medical records (EMR) systems by healthcare providers. Excellent in finding statistical patterns in rich data, Ljubic et. al. demonstrated the potential for deep learning models trained on EMR data for Alzheimer's onset prediction. To capture the richness of EMR data they trained separate LSTM models on diagnoses, lab tests, and drug domains. The drug and lab test domains produced the best results

with 0.985 and 0.986 AUPRC respectively while the diagnoses domain achieved 0.651 AUPRC [5].

1.3 *Problem Definition*

The complex nature of diabetes and related complications make it difficult to quantify a patient's diabetic risk level. Earlier approaches attempt to quantify risk by stratifying diabetic patients on a small subset of well-controlled clinically relevant lab tests. While stratification by this means is well adopted due to increased interpretability and practicality in a clinical environment, it severely under-utilizes the wealth of data available in today's EMRs.

On the other hand, recent deep learning models have taken advantage of high-dimensional data sources and have shown high prediction accuracy for prognostic disease models. However, due to the 'black box' nature of deep learning models and a subsequent lack of interpretability, deep learning models have not seen wide-spread adoption in a clinical setting.

Our goal is to develop a wholistic risk model for diabetic patients that monitors and predicts their overall diabetic health using diabetes stratification methods while also predicting the onset of diabetes related complications using high accuracy machine learning models.

2. METHODOLOGY

2.1 *Data*

To create a comprehensive risk profile for diabetic patients, we developed two time-series models on EMR data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). The CPCSSN database is comprised of anonymized clinical information from patients presenting with a wide variety of diseases and is split into several domains such as billing information, patient demographics, lab tests and medication [6]. Our first model used the lab and exam domains to forecast the values of 6 clinically relevant lab tests while our second model used billing, demographic, lab, and exam domains to predict the onset of diabetes-related complications. Diagnosis' codes were found in the billing domain and were represented using International Classification of Diseases-Ninth Revision codes (ICD9).

2.2 *Diabetes Stratification*

Following the American Diabetes Association (ADA) clinical guidelines [3], stratification levels were calculated for HbA1C, blood pressure (systolic and diastolic), high- and low-density lipoproteins and triglycerides, and albumin/creatinine ratio lab results. Two threshold values for each lab test determined the relative stratification level, 1 to 3, of patients where a score of 1 represented low values, 2 represented normal values, and 3 represented high values. The scores for HbA1C, blood pressure, lipids, albumin/creatinine ratios and multi-category lipid averages were found and summed to generate a final risk score.

Given the time series nature of EMR data, ARIMA and RNN models were used to forecast the selected lab values. The ARIMA model is a widely used statistical method for analyzing time series data. The model takes three parameters p,d,q where p is the order of the AR (autoregressive) term or lag order, d is the differencing order, and q is the order of the MA (moving average) term respectively. We used the standard parameters of $p = 5$, $d = 1$, $m = 0$ for our analysis.

Outliers from the series were removed by only taking data that fell within the inter-quartile range for each respective feature and 14-day windows were used to produce fixed time-series data. Thus 14-day forecasts were generated for each feature and re-stratification could occur using ADA threshold values. An 80/20 split was used for the LSTM model and each model was trained on an average of 800 samples.

2.3 *Complication Prediction*

In our second approach we sought to independently predict the onset of 10 diabetes-related complications. These complications were angina pectoris, atherosclerosis, ischemic heart disease, depressive disorder, diabetic nephropathy, diabetic neuropathy, diabetic retinopathy, hearing loss, myocardial infarction, and peripheral vascular disease. This was done using both patient diagnosis data and combined lab and exam result data. These sources of information provided two approaches which were developed separately, with the goal of consolidating the results to form a single model with the highest accuracy. In both approaches, two deep learning models were employed consisting of recurrent neural network (RNN) unidirectional LSTM and bidirectional RNN gated recurrent unit (GRU) architectures. The complication

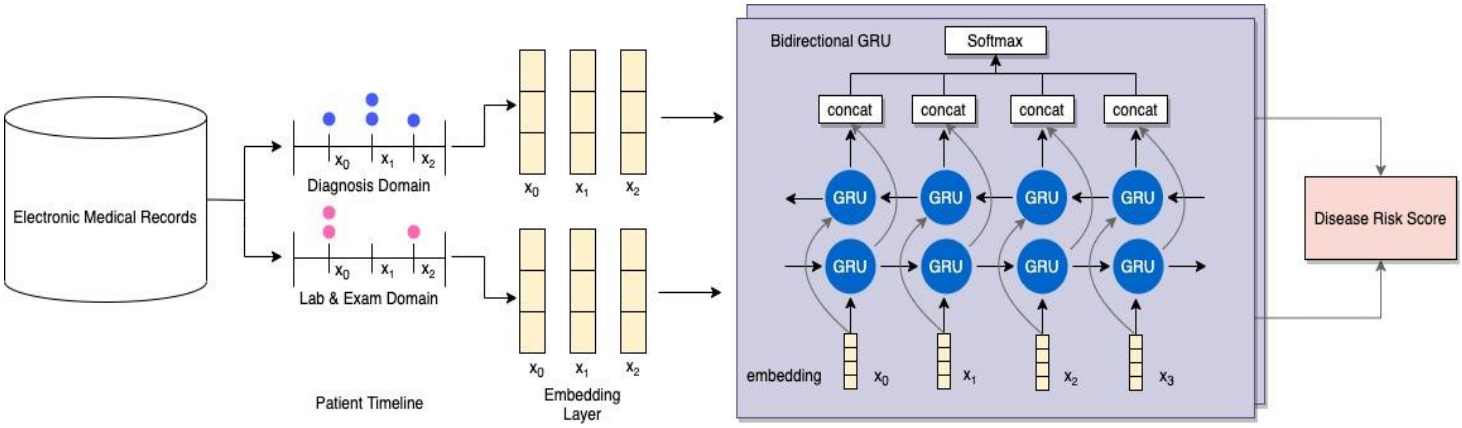

Figure 1: Onset of complication model diagram consisting of: the patient timeline as an input layer, an embedding layer, bidirectional GRU model architecture, and a disease risk score as the output layer.

diagnosis data points were ascertained from the billing table via recorded ICD-9 code ranges unique to each complication. A negative dataset was also created for each positive complication dataset by selecting age and gender matched diabetic patients without complications. Data were filtered to exclude results after the date of complication onset to prevent data leak and patients in both positive and negative datasets excluded results prior to each patient's date of diabetes onset. Additionally, patients were only included if they had at least 4 visits and at most 51 visits.

For the diagnosis domain, each patient's data were transformed into a one-hot encoded matrix of dimensions m x n, where m represented the number of rows or unique dates with at least one result and n = ***3397*** represented the full range of possible ICD-9 codes that presented at least 10 times in the data. Thus, each matrix element e_{ij} had a value of 1 if the patient had a diagnosis with an ICD-9 code *j* on a given date *i*, and a 0 otherwise.

For the lab and exam domains, the test results were filtered to include only 18 selected features, and outliers for each test type were removed. Each patient's results were again represented as one-hot encoded matrices by binning each test using the 33rd and 67th percentile values for each respective feature as thresholds resulting in m x n matrices. Again, m represented the number of unique dates with one or more lab/exam results, and n = 54 represented low, medium, high bins for each feature (18 x 3).

In both approaches, patient's with less than 50 visits were padded with zero vectors until they had 50 rows. Singular value decomposition (SVD) was used to reduce the dimensionality of input matrices, resulting in a final dimensionality of 50 x 50 or 2500 features per patient for the diagnosis domain and 18 x 50 or 900 features per patient for the lab and exam domain. Finally, encoded patient timelines were modeled using LSTM and bidirectional GRU layers. Softmax activation was used to generate onset probabilities. Figure 1 shows the complication model diagram. A 90/10 training/testing split was used and accuracy was evaluated using area under the receiver-operator curve (AUC) metric for each complication and approach.

3. RESULTS AND DISCUSSION

The 5-1-0 ARIMA model produced good results with the systolic and diastolic blood pressure data. For the diastolic blood pressure model, a root mean squared error of 3.620 mmHg was achieved on a range of diastolic blood pressure values between 82.0 mmHg and 71.0 mmHg. For the systolic blood pressure data, a root mean squared error of 6.264 mmHg was achieved on a range of values between 139.0 mmHg and 121.0 mmHg. The LSTM model produced the best results for HbA1C, HDL, LDL, triglycerides, and albumin/creatinine ratio with RMSE values of 5.3 (%), 14.2 (mmol/L), 13.5 (mmol/L), 50.3 (mmol/L), 32.1 respectively.

For our diabetes-complication onset models, it was found that using the bidirectional GRU model architecture yielded higher AUC and accuracy values than LSTM models in all cases, leading us to use it primarily for model evaluation. A model was separately constructed and evaluated for each complication and data source, with the exception of the diabetic retinopathy model using lab and exam data, which lacked a sufficient positive sample size (n < 1000). The results are summarized in Table 2.

The lab and exam data domain produced higher accuracies for seven out of the ten complications than the diagnosis domain, suggesting that this source of data was a better choice for our solution. This finding is also consistent with previous works [5]. We also found that models using lab and exam data with less than 1000 positive patients performed worse with AUC scores less than 0.8. This is not a surprising finding since many deep learning models typically require large training sets for higher accuracies.

Complication	*Diagnosis Data AUC*	*Lab/Exam Data AUC*
Angina pectoris	0.6373	0.9463
Atherosclerosis	0.6932	0.6249
ICHD	0.6941	0.9273
Depressive disorder	0.6714	0.8122
Nephropathy	0.6865	0.9108
Neuropathy	0.6970	0.9221
Retinopathy	0.6824	N/A
Hearing loss	0.7121	0.6659
MI	0.6298	0.8701
PVD	0.5098	0.5201

Table 1: Results from RNN Bidirectional GRU Models trained on the diagnosis and lab and exam domains for complication prediction; ICHD: ischemic chronic heart disease; MI: Myocardial infarction; PVD: Peripheral Vascular Disease.

4. CONCLUSIONS AND FUTURE WORK

As global diabetes prevalence continues to rise, diabetes monitoring and forecasting models are critical for early intervention and prevention of costly complications. In this study we set out to develop a wholistic diabetes risk model that encapsulates overall disease status and diabetes related complication development likelihood. We were able to successfully make a two-week prediction on six clinically relevant lab tests and subsequently calculate stratification levels as a general risk score. We expanded on this approach by developing separate GRU models for complication onset prediction with >0.85 AUC scores for 5 complications.

Without knowledge of the clinical domain, it is difficult to determine which features are of the most importance in determining diabetes risk levels. The features presented in this study, for both models, were found in literature review where data and features often differ from study to study. As such, a combination of features were selected based on frequency and relevancy reported by other researchers. To improve model accuracy, better feature selection can be achieved with the help of a clinical consultant. Specifically, features with high predictive value such as cholesterol or Alkaline Phosphatase in serum were excluded due to a lack of references to these features in the literature.

In addition, combining the diabetes related complication models trained on diagnoses data and lab data has been shown to improve overall model accuracy [5]. An ensemble model is proposed where model inputs would include complication onset probabilities from individual GRU models. Moreover, an ensemble model allows for greater domain usage as patient demographic data such as age, gender, and risk factors can be included as model inputs, providing even greater predictability and interpretability.

REFERENCES

[1] S. Pouya, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9th edition," *Diabetes Research and Clinical Practice,* vol. 157, no. 107843, 2019.

[2] Diabetes Canada, "Report on Diabetes – Driving Change," Ottawa, 2015.

[3] American Diabetes Association, "Clinical Practise Recommendations," *Diabetes Care,* vol. 22, 1999.

[4] J. J. W. S. R. L. M. L. M. S. R. C. G. P. Charles M. Clark, " A Systematic Approach to Risk Stratification and Intervention Within a Managed Care Environment Improves Diabetes Outcomes and Patient Satisfaction," *Diabetes Care,* vol. 24, no. 6, pp. 1079-1086, 2001.

[5] S. R. X. H. C. M. P. S. O. R. N. L. G. Z. O. Branimir Ljubic, "Influence of medical domain knowledge on deep learning for Alzheimer's disease prediction," *Computer methods and Programs in Biomedicine,* vol. 197, 2020.

[6] A. L.-L. K. M. J. A. L. R. M. S. K. R. B. Tyler Williamson, "Primary Health Care Intelligence," Canadian Primary Care Sentinel Surveillance Network, 2013.