A Machine Learning Approach for Diabetes Prediction in the Canadian Population

Introduction

In this research project, we investigate the problem of identifying diverse risk factors associated with diabetes within the Canadian population using data sourced from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). Additionally, we aim to explore whether the monitoring and prevention of diabetes can be improved when we model diabetes separately by depression status.

Diabetes, caused by the dysfunction of creating and using insulin in the human body, has become a severe concern around the world. In Canada, around 10% of the population is diagnosed with diabetes, and 30% live with diabetes or prediabetes (1). Due to the high rates of mortality and causes of cardiovascular diseases among diabetes patients, it is critical to thoroughly investigate the disease and provide efficient strategies for its prevention and treatment (2). In prior research, various risk factors associated with diabetes have been investigated, including demographic factors of age, sex, and ethnicity (3,4), biomarkers and clinical vital signs such as triglycerides (TG), systolic blood pressure (sBP), total cholesterol, fasting blood sugar (FBS), BMI (5-7), along with medical history factors like hypertension (HTN), depression, and smoking status (3,7). Noticeably, depression was found to be present more frequently among diabetes patients, and clinical difficulties exist in identifying the associations between these two health outcomes (8).

Thus, with the rise of biotechnology and machine learning techniques in the field of public health, various methods have been employed in diabetes studies, including random forests, decision trees, Naive Bayes, logistic regressions, etc (9,10). In our research, through utilizing another machine learning model, the XGBoost algorithm, we aim to provide knowledge in predicting diabetes and further compare the predicting accuracy with current methods. Also, we aim to generate innovative insights into the different effects of risk factors and explore whether the effects vary among individuals with or without depression. The third research question was to investigate whether using the recurrent neural network (RNN) model with long short-term memory (LSTM) that considers the temporal dynamics could improve diabetes prediction among patients with multiple records. Eventually, we aim to leverage the findings to facilitate Canadian primary care services, specifically in the areas of diabetes diagnosis, prognosis, and providing preventive measures in diabetes progression. Lastly, we employed Natural Language Processing (NLP) on 69 papers at the intersection of machine learning and diabetes prediction from sources like PubMed, Nature, and Springer, which revealed a technical focus on machine learning algorithms, with a gap in addressing diversities, health disparities, and ethical biases.

Methods

The dataset used in this study is the Diabetes Study dataset containing 10,000 observations in 2017, which was generated by randomly sampling patients with and without diabetes from CPCSSN. This dataset contains 8,602 unique de-identified patient visit records; each record contains 43 features on the patient’s basic demographic information, clinical vital signs, blood test results, the dates of these results, use of medication, and medical history. The outcome variable mainly studied in this project is diabetes status. Among the 10,000 patient visit records, there are 4,861 non-diabetic and 5,139 diabetic observations, with 53.36% females and 46.64% males, and the mean age in this dataset is around 63.2 with a range from 18 to 90 years old. Our dataset is relatively balanced in classes between diabetic and non-diabetic patients and covers a wide range of ages.

Based on the literature review, we kept all demographic information (age, sex), clinical vital signs (sBP, BMI), biomarkers (low-density lipoprotein, high-density lipoprotein, HbA1c hemoglobin, TG, FBS, total cholesterol), medical history (depression, hypertension, osteoarthritis, chronic obstructive pulmonary disease), use of medication (hypertension medications, corticosteroids) to predict the outcome diabetes status. For categorical features, we converted diabetes status and sex into binary coding, 0 for non-diabetic, male, and 1 for diabetic, female respectively. For patients' hypertension medication and corticosteroid medication records, we classified the presence of medication into class 1 (use of hypertension or corticosteroid medication) and the null or absence of medication into class 0 (no use of hypertension or corticosteroid medication), respectively. We further looked at the correlation between numerical features and the outcome variable, where A1c and FBS showed a relatively high correlation with diabetes status, while sBP and TG showed little correlation. Furthermore, LDL and total cholesterol, FBS and A1c showed a relatively high correlation with each other.

To fully explore the dataset, we first split the data into 70% training, 15% validation, and 15% test sets, and then normalized the numerical columns in the three sets separately. Also, to impute the missing values, we used Multiple Imputation by Chained Equation (MICE) in the training, validation, and test sets separately. To further select features, we checked multicollinearity between variables by variance inflation factor (VIF) values. The three cholesterol measurements were found to be highly correlated variables with high VIF values, and thus we only kept the variable total cholesterol level to avoid having unstable coefficients or inflated standard errors. Then, we employed and compared multiple models, including K-nearest neighbor (KNN), logistic regression, random forest, XGBoost, Naive Bayes, Linear Discriminant Analysis (LDA), and Quadratic Discriminant Analysis (QDA). For KNN, we used 5-fold cross-validation to determine the most optimal value for K based on overall accuracy. For random forest and XGBoost, we used grid search with 10-fold or stratified 10-fold cross-validation to tune the hyperparameters, respectively. All hyperparameter tuning was performed using the validation set. We also looked at generative and discriminative models, including Naive Bayes, LDA, and QDA. For Naive Bayes, we used Gaussian and Categorical Naive Bayes mixing models to account for the numerical and categorical variables in our model. For LDA and QDA, we only included the numerical features as the model assumes a multivariate Gaussian distribution on the conditional distribution of the features given the class label.

For our second research question, we split the dataset into two subsets, one containing 7,878 observations without depression and the other containing 2,122 observations with depression. Then we split each subset into 70% training, 15% validation and 15% test sets. Normalization and MICE imputation on the three sets separately. We then proceeded to assess the models' performance. Because the goal was to correctly identify as many true diabetic patients as possible, we focused on the recall score of the model on the test set. After selecting the final model using aggregated patient data, we would apply the final model to the two distinct data sets of patients with or without depression.

For the third research question, we first filtered the data set and only included 2,497 patients with 2 or more records, grouped by patient ID, and then split the data set into 80% training set and 20% test sets. Then, we followed by normalization and MICE imputation on the training and test set separately. Because the maximum number of records for each patient ID was 8 in this data set, we created sequences by padding with a maximum length of 8. Then we took a deep learning approach with an RNN-LSTM model and compared its predictive performance between the training and test set. The reason for using the RNN model with LSTM was that LSTM is better at selectively forgetting or retaining information over longer sequences. The model architecture consisted of one LSTM layer followed by two fully connected (FC) layers. Two dropout layers were included to reduce the overfitting of training data, where one was between the LSTM layer and the first FC layer, and the other was between the two FC layers. We applied the Tanh activation function after the first FC layer and the Sigmoid activation function after the second FC layer which resulted in a probability between 0 and 1. During the evaluation of model performance on training and test data, a probability equal to or greater than 0.5 resulted in a prediction of 1 and 0 otherwise. We utilized binary cross entropy loss function with L2 regularization, preventing overfitting and improving the model's generalization capability.

Results

Following model fitting and performance comparison of all models, the XGBoost model had the best overall performance with a recall score of 0.87, a precision score of 0.84, and a 0.86 accuracy for identifying diabetic patients in the test set. Figure 1 (left) shows the corresponding confusion matrix on the test set, where most observations were correctly classified as shown by the high numbers on the diagonal. Table 1 displays the tuned hyperparameters of models, if available. The performance of the XGBoost model also appeared to be comparable on training and test sets, suggesting good generalizability and no evidence for overfitting. Notably, the random forest model delivered competitive performance on the test set, but the f1-score was slightly lower. Another reason for opting for the XGBoost model was the sequential implementations of decision trees, which could result in better overall prediction than the random forest model where multiple trees were constructed independently. As a result, we selected XGBoost as the final model.

For the second research question, we applied two distinct XGBoost models to the two data sets of observations with or without depression, where imputation and hyperparameter tuning were performed separately. For observations without depression, the XGBoost model had a precision score of 0.87 and a recall of 0.90 with an accuracy of 0.88 for identifying diabetes among patients without depression. For observations with depression, the XGBoost model achieved a precision of 0.81, a recall of 0.92, and an accuracy of 0.88 for the identification of diabetic patients with depression.

For the third research question, the training accuracy reached 0.88 while the test accuracy reached 0.84 at the 1000th epoch for the RNN-LSTM model, as shown in the epoch-accuracy plot in Figure 1 (right). The minor difference between the training and test accuracy suggested that the RNN-LSTM model had decent generalizability on unseen patient data. Also, a test accuracy of 0.84 at the last epoch indicated promising performance in predicting diabetes for patients with multiple health records.

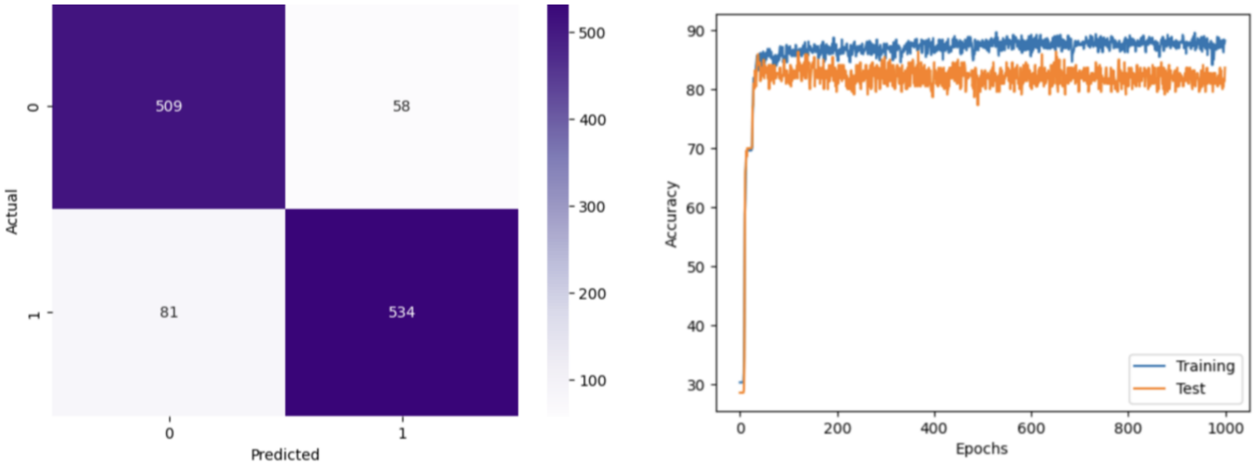

Figure 1. Confusion matrix (left) for the XGBoost model using test set and epoch-accuracy plot (right) for RNN-LSTM model using training and test sets.

Table 1. Hyperparameters for models fitted in this study if hyperparameter tuning was performed.

Model	**Hyperparameter(s)**
KNN (all observations)	k = 40
Random forest (All observations)	Number of estimators: 100 Maximum depth: 20 Maximum number of features: square root of sample size Minimum samples per leaf: 20 Minimum samples to split: 10
XGBoost (All observations)	Number of estimators: 100 Maximum depth: 1 Minimum of samples per leaf: 20 Learning rate: 0.01
XGBoost (Subset of observations with depression)	Number of estimators: 100 Maximum depth: 1 Minimum of samples per leaf: 25 Learning rate: 0.1
XGBoost (Subset of observations without depression)	Number of estimators: 150 Maximum depth: 3 Minimum of samples per leaf: 10 Learning rate: 0.01
RNN-LSTM (Subset of patients with multiple visits)	Hidden size: 3 Epochs: 1000 Learning rate: 0.001, decreased by 10% after 500 epochs Drop-out rates: 0.5 and 0.2 for two drop-out layers, respectively Batch size: 50 L2 regularization parameter lambda: 0.001

Discussion

Diabetes is a complex and widespread metabolic disorder with significant public health implications. Diabetes prediction is thus crucial for potential disease prevention through lifestyle changes and early intervention. Using machine learning techniques like XGBoost to predict diabetes status is crucial because it allows us to go beyond traditional clinical thresholds and include a broader range of risk indicators and problems that can arise in people with varying blood sugar levels. We applied various machine learning models to the diabetes data from the CPCSSN. After comparing model performance with an emphasis on recall score, we selected XGBoost models for their strong predictive capabilities while also reducing the risk of overfitting. Other strengths of the XGBoost models are the ability to handle diverse data types and no requirement for data normalization or scaling. Remarkably, all three XGboost models performed comparably well in predicting diabetes status when using aggregated patient data regardless of depression status, patients with depression, or patients without depression. As a result, when it comes to diabetes prediction, it may not be required to differentiate between patients with or without depression. Alternatively, we applied the RNN model with LSTM to the group of patients with repeated measurements and achieved high accuracy on the test set. The RNN-LSTM model could effectively capture patterns within the sequence of patient health records, identifying potential indicators or trends leading to diabetes onset. Also, the prediction accuracy of RNN-LSTM model was comparable to that of the XGBoost model. Overall, our models may be a useful supplementary tool for clinicians and health care providers to identify diabetes among patients at an earlier stage, provide targeted care pathways and suggestions for those at high risk, and alleviate the prevalence of diabetes by enhancing proper prevention measures and a healthier lifestyle.

To generate a comprehensive view of how diversity, equity, and equality are addressed in ML research, we employed NLP on 69 papers intersecting with machine learning and diabetes prediction, from sources such as PubMed, Nature, and Springer. We first applied data cleaning to the title, abstract, and discussion sections, tokenized the text, removed stop words, and performed lemmatization. Three word clouds were generated based on the processed title, abstract, and discussion sections from the papers. Among titles, "diabetes," "machine learning," and "prediction" frequently appeared. A higher occurrence of words such as "model," "prediction," "use," and "performance" in abstracts suggested a focus on the technical aspects of machine learning. In discussions, words like "model," "use," "study," and "patient" are common, with additional terms such as "population" and "individual" also identified. We thus observed trends of emphasis on evaluating and enhancing the effectiveness and efficiency of machine learning models in ML research, whereas gaps tended to exist relevant to health disparities and ethical biases as very few relevant words, such as "Chinese" and "Ethiopia," can be identified through NLP (11-13). Specifically, potential biases can occur that lead to inadequate diversities and inequities in monitoring healthcare outcomes, such as bias in social determinants of health, with race and sex being primary factors, and vulnerable groups, including ethnic minorities and marginalized communities. Further research can be focused on minority and disadvantaged groups with diabetes, to facilitate public health equity better and construct a machine learning approach at a larger and equally included population scale (12-14).

This study has several limitations. Firstly, patient records may be dependent on each other since there are only 8,602 unique patient IDs. This means that some patients had multiple records in the longitudinal data, but we failed to account for the correlation of the biomarkers within the same patient for the first two research questions. However, we took repeated measurements for some patients into account and created sequences for those patients in the RNN model for the third research question. Also, only sBP and A1c in our data align with the acceptable value ranges given by Dr. Karim Keshavjee. This means that there are outliers that could affect model performance. Also, the selected XGBoost model is an ensemble of multiple trees resulting in aggregated predictions that can be hard to interpret compared to other traditional machine learning models such as KNN, logistic regression, etc. The RNN-LSTM model could also be less interpretable, and it requires the sequential input of multiple patient records that might not be available when the patient only visited once.

For future studies, other common patient features may be collected to improve prediction, including family history of diabetes, smoking status, and waist circumference (2,7,15). In the next steps, we will keep exploring other techniques, such as hidden Markov models and neural networks. Also, we might apply methods to other datasets or merge datasets to obtain more robust prediction results for diabetes status. We might fit models by stratifying demographic groups to develop insights into potential disparities in the identification process of diabetes status. Eventually, we will provide a comprehensive investigation and generate clinically significant suggestions for diabetes diagnosis and prevention.

Individual Contributions: All team members contributed equally to this project.
Code and Presentation: GitHub: https://github.com/Yutong-Lu/CHL5230FinalProject
Presentation slides: Group 5 Final Presentation-CHL5230

Acknowledgement: We want to thank Dr. Karim Keshavjee for providing us with access to this Diabetes Study dataset from CPCSSN under the research project 2015SRSC51. We also want to thank Professor Zahra Shakeri for bridging this opportunity for us.

Data Source: The Diabetes Study dataset used in this study, was sourced from family physicians and other primary care providers. This national database of EMR data was provided by the CPCSSN, a multi-disease EMR-based surveillance system in Canada (http://cpcssn.ca/).

References

1. About diabetes [Internet]. [cited 2023 Oct 29]. Available from: https://www.diabetes.ca/about-diabetes
2. Chen L, Magliano DJ, Balkau B, Colagiuri S, Zimmet PZ, Tonkin AM, et al. Ausdrisk: An Australian type 2 diabetes risk assessment tool based on demographic, lifestyle and simple anthropometric measures. Medical Journal of Australia. 2010;192(4):197–202. doi:10.5694/j.1326-5377.2010.tb03478.x
3. Stern MP, Williams K, Haffner SM. Identification of persons at high risk for type 2 diabetes mellitus: Do we need the oral glucose tolerance test? Annals of Internal Medicine. 2002;136(8):575. doi:10.7326/0003-4819-136-8-200204160-00006
4. Chien K, Cai T, Hsu H, Su T, Chang W, Chen M, et al. A prediction model for type 2 diabetes risk among Chinese people. Diabetologia. 2008;52(3):443–50. doi:10.1007/s00125-008-1232-4
5. Gupta AK, Dahlof B, Dobson J, Sever PS, Wedel H, Poulter NR. Determinants of new-onset diabetes among 19,257 hypertensive patients randomized in the Anglo-scandinavian cardiac outcomes trial–blood pressure lowering arm and the relative influence of antihypertensive medication. Diabetes Care. 2008;31(5):982–8. doi:10.2337/dc07-1768
6. Wilson PW. Prediction of incident diabetes mellitus in middle-aged adults. Archives of Internal Medicine. 2007;167(10):1068. doi:10.1001/archinte.167.10.1068
7. Balkau B, Lange C, Fezeu L, Tichet J, de Lauzon-Guillain B, Czernichow S, et al. Predicting diabetes: Clinical, biological, and genetic approaches. Diabetes Care. 2008;31(10):2056–61. doi:10.2337/dc08-0368
8. Holt RI, de Groot M, Golden SH. Diabetes and depression. Current Diabetes Reports. 2014;14(6). doi:10.1007/s11892-014-0491-3
9. Zou Q, Qu K, Luo Y, Yin D, Ju Y, Tang H. Predicting diabetes mellitus with machine learning techniques. Frontiers in Genetics. 2018;9. doi:10.3389/fgene.2018.00515
10. Perveen S, Shahbaz M, Keshavjee K, Guergachi A. Metabolic syndrome and development of diabetes mellitus: Predictive modeling based on machine learning techniques. IEEE Access. 2019;7:1365–75. doi:10.1109/access.2018.2884249
11. Mhasawade V, Zhao Y, Chunara R. Machine Learning and Algorithmic Fairness in public and Population Health. Nature Machine Intelligence. 2021;3(8):659–66. doi:10.1038/s42256-021-00373-4 \
12. Char DS, Shah NH, Magnus D. Implementing machine learning in health care — addressing ethical challenges. New England Journal of Medicine. 2018;378(11):981–3. doi:10.1056/nejmp1714229
13. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of Algorithmic Fairness Solutions in Health Care Machine Learning. The Lancet Digital Health. 2020;2(5). doi:10.1016/s2589-7500(20)30065-0
14. Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to Advance Health Equity. Annals of Internal Medicine. 2018;169(12):866. doi:10.7326/m18-1990
15. Hippisley-Cox J, Coupland C, Robson J, Sheikh A, Brindle P. Predicting risk of type 2 diabetes in England and Wales: Prospective derivation and validation of qdscore. BMJ. 2009;338(b880). doi:10.1136/bmj.b880