Predicting Diabetes Occurrence in the Canadian Population Using Machine Learning Approaches

Nitya Kuruvila & Sashini Kosgoda (Team 8)

Dalla Lana School of Public Health, University of Toronto

CHL 5230: Applied Machine Learning for Health Data

Dr. Zahra Shakeri

December 8, 2023

Introduction

Type 2 diabetes (T2DM) is a chronic condition characterized by insulin resistance, leading to abnormal insulin response in cells (CDC, 2023, p. 2), and poses a serious threat to global health. The disease affects approximately 537 million people worldwide (Magliano et al., 2021), primarily individuals living in low- or middle-income countries (WHO, n.d.). Annually, 1.5 million global deaths are attributed directly to diabetes and this number continues to rise (WHO, n.d.). Diabetes can lead to severe health consequences over time including, damage to the heart, kidneys and blood vessels (WHO, n.d.). The diagnosis rates of diabetes vary significantly across the world, with 87.5% of all undiagnosed cases being in low or middle-income countries (Magliano et al., 2021). This disparity can be attributed to numerous factors including socioeconomic conditions, the local health system, and public awareness among others (Magliano et al., 2021). This paper aims to explore three pivotal questions at the intersection of diabetes diagnosis, biomarker utilization, other medical conditions and the ethical dimensions of machine learning (ML) in health research.

Firstly, this paper examines the relationship between diabetes and other medical conditions. The presence of other medical condition(s) can greatly impact the occurrence, treatment and management of diabetes (Nowakowska et al., 2019). Secondly, this paper aims to use biomarkers to effectively classify individuals with diabetes. The biomarkers explored in this paper are those with lesser-known associations with diabetes and/or those less frequently used for diagnostic purposes (Ghari Arab et al., 2018). The exploration of these first two questions builds on previous work (datathon 3) focused on the same classification goals. Thirdly we aim to shift focus to the ethical dimensions of ML in health research, including diversity, equity, and equality, particularly within diabetes research. Given the global prevalence and severe health impacts that can result from T2DM, it is important to better understand how other conditions (Klein et al., 2022; Petrie et al., 2018; Rhee et al., 2017; Rotella & Mannucci, 2013) and biomarkers (Selvin et al., 2011) may influence a diabetes diagnosis. The insights gained through this analysis can inform and improve new methods for diagnosis and treatment of T2DM, subsequently improving the accessibility and availability of diabetes-focused health care globally. Further, as the utilization of machine learning becomes increasingly prevalent in healthcare (Joo et al., 2023), concerns surrounding diversity, equity, and equality necessitate examination. Addressing how these concepts are incorporated into ML health research is crucial not only for the reliability of predictive models (Houssein & Sayed, 2023) but also to ensure that healthcare advancements are accessible and applicable to diverse populations (Rajpurkar et al.). The global prevalence of this disease highlights the importance of having diversity and equity considerations in ML-driven diabetes research.

Methods

Similar to the previous work, the analysis for the first two questions utilizes a data set from the *Canadian Primary Care Sentinel Surveillance Network* (Keshavjee, 2023). This dataset

consists of 10,000 records, including 5139 diabetic cases and 4861 non-diabetic cases as well as various medical, health and demographic information. The first two questions were explored using long short-term memory networks (LSTM) to classify patients with and without diabetes based on biomarkers and medical conditions. We began our analysis by cleaning and preprocessing the data. This first consisted of dropping variables out of the scope of our analysis. The variables selected for exploration were based on findings from the previous work; this included health conditions shown to have a significant association with diabetes and biomarkers that were determined to be effective for the classification of diabetes. These factors were systolic blood pressure, body mass index, A1c, triglyceride levels, total cholesterol, depression and hypertension. Additionally, sex and age at the exam were included to ensure that equity and diversity were considered during our analysis. Low- and high-density lipoprotein levels were not included in further analysis as these factors are accounted for by the total cholesterol level (Rosenson, 2021). Missing data was identified and then imputed using the median to maintain the distribution of the data. Additionally, categorical variables were replaced with numerical equivalents. Next, we conducted data exploration by examining the distribution of equity-focused variables, age and sex, as well as the distribution of all variables by diabetes class. To prepare our data for modeling we first grouped patients by their patient ID to ensure each patient was accounted for only once during analysis. The data was split into training and testing groups, using 80% for training, and then all numerical variables were normalized to improve the accuracy of our results. Sequences of length 100 were generated and class imbalance was also explored. To develop the LSTM model, we first defined the hyperparameters and set a seed of 100 for reproducibility. Next, we prepared the data, defined the model architecture, and defined the loss and optimizer functions which were cross entropy and Adam respectively. We then trained the model, iterating it over the number of epochs previously defined, validated it and evaluated the model based on the average of the following metrics: precision, recall, F1 score, training accuracy and validation accuracy. Subsequently, we performed a grid search to determine the best hyperparameters. These findings were used to inform the final LSTM model which was evaluated using the aforementioned metrics.

To address the third question, a comprehensive search strategy was implemented, targeting 60 relevant articles on ML diabetes research using the electronic database PubMed. The search terms used included keywords such as “diabetes”, “prediction”, “risk”, “machine learning”, “artificial intelligence”, and combinations of these words. Articles were screened manually, and those focused on diabetes prediction and risk using ML of AI techniques, published after 2000, were included in the analysis. The selected articles' titles, abstracts, and discussions were then compiled into a CSV file for further analysis. To preprocess the data, using natural language processing (NLP) techniques included handling contractions, standardizing text through lowercasing text and, removing special characters, and numbers. Additionally, the text underwent tokenization, removal of English and context-sensitive stop words, Part-Of-Speech

(POS) tagging and lemmatization. To visually represent the processed text, a word cloud was generated, offering a snapshot of the most prevalent terms in the corpus.

Results

Five variables with missing data were identified although none had over 5% missing data and as such, all variables were included for analysis. The distribution of sex in the data set was balanced and the distribution of age demonstrated that most patients were above the age of 50, consistent with the period when individuals are frequently tested for diabetes (Ekoe, 2018). The distribution of all variables by diabetes class demonstrated no significant differences between diabetic and non-diabetic individuals. Preprocessing of the data also determined that the training dataset was not imbalanced. The LSTM model aiming to classify patients based on their diabetes status had an average training accuracy of 52.68%, a validation accuracy of 51.31%, a precision of 0.20, a recall score of 0.35 and F1 score of 0.24.

The evaluation of the previously developed logistic regression model determined AUC and accuracy scores of 0.63 and 0.62 for the training set, and 0.63 and 0.62 for the test set, respectively. The precision scores for individuals with and without diabetes were 0.58 and 0.71 in the training set, and 0.57 and 0.71 in the test set. The recall values were 0.80 and 0.45 in the training set, and 0.79 and 0.46 in the test set for individuals with and without diabetes respectively. The performance metrics of the previous Gaussian Naive Bayes model indicated the precision of predicting diabetic patients is 0.84, recall is 0.78, and F1-score is 0.81. The average AUC and accuracy of the validation set were 0.805 and 0.807 respectively. The AUC and accuracy of the testing set were 0.826 and 0.828 respectively.

The results from the word cloud of the research articles show that the most prevalent words in the analyzed articles are “propose”, “develop”, “high”, “include” and “factor”. No words relating to diversity, equity, and equality are present in the word cloud.

Figure 1. Word Cloud Analysis of Diversity, Equity, and Equality in Current Diabetes Machine Learning Research

Discussion

The evaluation of the LSTM model demonstrates that the model does not indicate being overfit, however, it is not very effective at classifying diabetic vs. non-diabetic individuals based on the aforementioned factors. Given the health consequences of diabetes, it is important that those with diabetes are correctly classified. The evaluation metrics, namely the precision, recall and accuracy score reflect low levels of sensitivity, suggests that the LSTM model is not well suited for diabetes classification. In contrast, the previously developed logistic model, focused on classifying patients based on other health conditions and the Gaussian Naive Bayes model, focused on classifying patients based on biomarkers, performed much better at classifying patients. The evaluation metrics of all models can be seen in *Table 1*. The increased sensitivity and precision of these models would indicate that they are more effective than the LSTM model at correctly classifying those with diabetes based on other medical conditions and biomarkers; the most effective being the Gaussian Naive Bayes models. Some limitations faced throughout our analysis were time and resource constraints that limited the number of models and the complexity of the models we were able to develop.

As evident in the word cloud, current research on diabetics using machine learning does not excessively address diversity, equity, and equality. The word cloud reveals that the thematic focus of diabetics research using ML focuses more on the study design and conducting a study and less attention to making the study more equitable. Our dataset, in particular, only captures age and sex information, neglecting to record other factors such as participants' ethnicity or socioeconomic status; these factors can act as a limitation to the model's generalizability. This oversight highlights a lack of importance assigned to accounting for diversity in the research. Further research in ML, specifically for diabetics, should incorporate the concepts of diversity, equity, and equality into their research studies to allow the studies to be more generalizable to the greater population. However, it is important to highlight that the distribution of both sex and age (of individuals 40 to 90) in the dataset is relatively balanced. Notably, there are fewer individuals under the age of 40 who underwent diabetes testing, aligning with Diabetes Canada's recommendation to screen individuals over the age of 40 for Type 2 diabetes (Ekoe et al., 2018). To address equity in our research, we made sure to include gender and age as factors for consideration. Doing so ensures that our model trains on a diverse population and can determine differences based on these factors.

Individual Contributions

Sashini Kosgoda contributed to working on the Gaussian Naive Bayes model (Q#2) and NLP (Q#3) by creating code, running analysis and writing the report. Nitya Kuruvila contributed to working on the Logistic Regression model (Q#1) and the LSTM model (Q#1 and Q#2) by creating code, running analysis and writing the report. Both contributed to data exploration, writing the introduction and discussion and editing.

GitHub Link: https://github.com/sashini472/Project-Phase-3.git

Presentation: Final Presentation-CHL5230

References

CDC. (2023, April 18). *Type 2 Diabetes*. Centers for Disease Control and Prevention. https://www.cdc.gov/diabetes/basics/type2.html

Ghari Arab, A., Zahedi, M., Kazemi Nejad, V., Sanagoo, A., & Azimi, M. (2018). Correlation between Hemoglobin A1c and Serum Lipid Profile in Type 2 Diabetic Patients Referred to the Diabetes Clinic in Gorgan, Iran. *Journal of Clinical and Basic Research*, *2*(1), 26–31. https://doi.org/10.29252/jcbr.2.1.26

Houssein, E. H., & Sayed, A. (2023). Boosted federated learning based on improved Particle Swarm Optimization for healthcare IoT devices. *Computers in biology and medicine*, *163*, 107195. https://doi.org/10.1016/j.compbiomed.2023.107195

Joo, H., Mathis, M. R., Tam, M., James, C., Han, P., Mangrulkar, R. S., Friedman, C. P., & Vydiswaran, V. G. V. (2023). Applying AI and Guidelines to Assist Medical Students in Recognizing Patients With Heart Failure: Protocol for a Randomized Trial. *JMIR research protocols*, *12*, e49842. https://doi.org/10.2196/49842

Keshavjee, K. (2023, June 20). *PREVENT - Institute of Health Policy, Management and Evaluation*. Institute of Health Policy, Management and Evaluation. https://ihpme.utoronto.ca/initiative/prevent/

Klein, S., Gastaldelli, A., Yki-Järvinen, H., & Scherer, P. E. (2022). Why does obesity cause diabetes? *Cell Metabolism*, *34*(1), 11–20. https://doi.org/10.1016/j.cmet.2021.12.012

Magliano, D. J., Boyko, E. J., & Committee, I. D. A. 10th edition scientific. (2021). Global picture. In *IDF DIABETES ATLAS [Internet]. 10th edition*. International Diabetes Federation. https://www.ncbi.nlm.nih.gov/books/NBK581940/

Nowakowska, M., Zghebi, S. S., Ashcroft, D. M., Buchan, I., Chew-Graham, C., Holt, T., Mallen, C., Van Marwijk, H., Peek, N., Perera-Salazar, R., Reeves, D., Rutter, M. K., Weng, S. F., Qureshi, N., Mamas, M. A., & Kontopantelis, E. (2019). The comorbidity burden of type 2 diabetes mellitus: Patterns, clusters and predictions from a large English primary care cohort. *BMC Medicine*, *17*, 145. https://doi.org/10.1186/s12916-019-1373-y

Petrie, J. R., Guzik, T. J., & Touyz, R. M. (2018). Diabetes, Hypertension, and Cardiovascular Disease: Clinical Insights and Vascular Mechanisms. *The Canadian Journal of Cardiology*, *34*(5), 575–584. https://doi.org/10.1016/j.cjca.2017.12.005

Rajpurkar, Pranav, et al. "AI in Health and Medicine." *Nature Medicine*, vol. 28, no. 1, Jan. 2022, pp. 31–38, www.nature.com/articles/s41591-021-01614-0, https://doi.org/10.1038/s41591-021-01614-0.

Rhee, E.-J., Han, K., Ko, S.-H., Ko, K.-S., & Lee, W.-Y. (2017). Increased risk for diabetes development in subjects with large variation in total cholesterol levels in 2,827,950 Koreans: A nationwide population-based study. *PLOS ONE*, *12*(5), e0176615. https://doi.org/10.1371/journal.pone.0176615

Rosenson, R. S. (2021, July 12). Patient education: High cholesterol and lipids (Beyond the Basics)—UpToDate. UpToDate. https://www.uptodate.com/contents/high-cholesterol-and-lipids-beyond-the-basics/print

Rotella, F., & Mannucci, E. (2013). Depression as a risk factor for diabetes: A meta-analysis of longitudinal studies. *The Journal of Clinical Psychiatry*, *74*(1), 31–37. https://doi.org/10.4088/JCP.12r07922

Selvin, E., Steffes, M. W., Gregg, E., Brancati, F. L., & Coresh, J. (2011). Performance of A1C for the Classification and Prediction of Diabetes. *Diabetes Care*, *34*(1), 84–89. https://doi.org/10.2337/dc10-1235

WHO. (n.d.). Diabetes. World Health Organization. Retrieved October 27, 2023, from https://www.who.int/health-topics/diabetes

Appendix

Model		**Precision**	**Accuracy**	**Recall**	**F1**
Logistic Regression	Non-diabetic	0.71	0.62	0.46	0.56
	Diabetic	0.57	0.62	0.79	0.66
Gaussian naive bayes	Non-diabetic	0.82	0.83	0.87	0.84
	Diabetic	0.84	0.83	0.78	0.81
LSTM	Average	0.20	0.51*validation	0.35	0.24

Table 1. Evaluation metrics of various ML models focused on classifying diabetic vs non-diabetic individuals based on medical conditions and biomarkers.