Predicting Diabetes in Canadian Adults Using Machine Learning

Kayla Esser[1*], Monica Duong[1*], Khalil Kain[1*], Son Tran[1*], Aryan Sadeghi[2], Aziz Guergachi[2,3], Karim Keshavjee[2], Mohammad Noaeen[1], and Zahra Shakeri[2]

***Abstract*— Rising diabetes rates have led to increased healthcare costs and health complications. An estimated half of diabetes cases remain undiagnosed. Early and accurate diagnosis is crucial to mitigate disease progression and associated risks. This study addresses the challenge of predicting diabetes prevalence in Canadian adults by employing machine learning (ML) techniques to primary care data. We leveraged the Canadian Primary Care Sentinel Surveillance Network (CPCSSN), Canada's premier multi-disease electronic medical record surveillance system, and developed and tuned seven ML classification models to predict the likelihood of diabetes. The models were tested and validated, focusing on clinical patient characteristics influential in predicting diabetes. We found XGBoost performed best out of all the models, with an AUC of 92%. The most important features contributing to model prediction were HbA1c, LDL, and hypertension medication. Our research aims to aid healthcare professionals in early diagnosis and to identify key characteristics for targeted interventions. This study contributes to an understanding of how ML can enhance public health planning and reduce healthcare system burdens.**

I. Introduction

Diabetes, a chronic metabolic disease characterized by hyperglycemia, is one of the largest global health emergencies of the 21st century[1], [2]. Diabetes rates continue to rise worldwide, as do the number of people experiencing acute and chronic complications. High blood glucose is estimated to be the third highest risk factor for premature mortality globally [3]. Canada has also seen rising rates of diabetes. As of 2022, 8.9% of the population had been diagnosed with diabetes, and prevalence has increased by an average of 3.3% per year[4], [5]. Diabetes is the leading cause of blindness, non-traumatic amputation, and end-stage renal disease in Canadian adults [3]. The cost of diabetes in Canada was estimated at $15.36 billion in 2022 [6]. However, it is also estimated that half of diabetes cases go undiagnosed, meaning the true population burden is higher than reported [7]. Early and accurate diagnosis and treatment of diabetes are imperative to prevent further disease progression and complications such as diabetic retinopathy, cardiovascular events and mortality [1].

*These authors contributed equally to this work and share first authorship.

[1]Kayla Esser, Monica Duong, Khalil Kain, Son Tran, and Mohammad Noaeen are with the Dalla Lana School of Public Health, University of Toronto, Canada. kayla.esser@mail.utoronto.ca

[2]Aryan Sadeghi, Aziz Guergachi, Karim Keshavjee and Zahra Shakeri are with the Institute of Health Policy, Management and Evaluation, Dalla Lana School of Public Health, University of Toronto, Canada. zahra.shakeri@utoronto.ca

[3]Aziz Guergachi is with Ted Rogers School of Information Technology Management, Toronto Metropolitan University, Toronto, Canada; and Department of Mathematics and Statistics, York University, Toronto, Canada.

Machine learning (ML) involves the application of computer algorithms to create a model of sample data to make predictions or decisions [8]. These models can be improved through testing, parameter tuning and validation. Deep learning is a subcategory of ML in which a neural network uses a hierarchical architecture to adapt to features of the dataset and learn from the data to improve predictive abilities. ML models have been used in the prediction, classification, and management of diabetes [1], [8]–[17]. Models that can predict diabetes may be useful for aiding clinician diagnosis of diabetes, as well as highlighting which features may be meaningful to target for early intervention. Much of the literature on ML diabetes prediction has focused on the US and other populations, therefore a model developed and validated on Canadian patient data may be more informative for the Canadian context, given differences in the health care system [14].

This study uses cohort data from a Canada-wide multi-disease electronic medical record surveillance system to answer the following research questions (RQs):

RQ1: Can we predict diabetes prevalence accurately, by comparing the performance of seven ML models: logistic regression (LR), decision tree (DT), random forest (RF), XGBoost (XGB), support vector machine (SVM), combined Naive Bayes (CNB), and an artificial neural network (ANN)?

RQ2: Which clinical patient characteristics are most influential in the best-performing ML model's prediction of diabetes prevalence?

These research questions have significant implications for public health planning and interventions. Our goal is to aid physicians in accurate diagnosing of diabetes, which can lead to earlier treatment and intervention, and ultimately reduce disease burden and health care system costs. Additionally, a systematic review of predictive machine learning diabetes models found a large variety in both the number and type of features included in models, indicating a lack of consensus [1]. Our study may clarify the utility of clinical and biomarker data available in the health records of many patients.

II. Methods

A. Data collection and preparation

This study used data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN), Canada's first multi-disease electronic medical record surveillance system [18]. The CPCSSN is a network that spans eight provinces and one territory and has over two million patients and 1500 primary care clinicians [18]. CPCSSN extracts de-identified

electronic medical records from primary care practices in Canada and applies cleaning, coding, and standardization algorithms to transform the data for use in quality improvement, surveillance, and research [18]. Diabetes cases are defined as type 2 diabetes mellitus, controlled or uncontrolled, excluding gestational diabetes, chemically induced diabetes, neonatal diabetes, hyperglycemia, prediabetes, or similar states or conditions [19]. The dataset for this study was generated by CPCSSN and shared with the research team. This was done by pulling patients aged 18 and up with a blood pressure reading and joining records that were the closest in time for other measurements (e.g., ±1 year). Patients on insulin were removed. The final data set is a random subset of 10,000 observations from the original CPCSSN dataset who had a blood pressure reading at their primary care providers' offices between 2004 and 2014.

Data exploration included examining the extent of missing data in each variable, the variables' distribution and central tendencies, and observing continuous variables' correlations for potential collinearity. Missing values were addressed through multiple imputations by chained equations (MICE) for variables with under 10% missing (this cutoff was chosen to minimize introduced bias) [20]. We performed data standardization. In the case of two collinear variables, we selected one based on the strength of the association with the outcome in the literature. The final predictor variables used in all models were age, sex, systolic blood pressure (sBP), body mass index (BMI), low-density lipoprotein (LDL), high-density lipoprotein (HDL), HbA1c, triglycerides (TG), depression, hypertension, osteoarthritis, chronic obstructive pulmonary disease (COPD), use of hypertension medications and use of corticosteroids. For patients with multiple appointments, only data from the last visit was included. The final dataset contained 8,602 observations.

B. Development of predictive models

We first held out a random 15% of the data for validation. Then, we randomly split the remaining data into 70% training and 30% testing. We used the training data to train the seven models (LR, SVM, DT, RF, XGBoost, CNB and ANN). Subsequently, we used the testing data to evaluate how well the models perform on unseen data (via confusion matrices and utility functions for accuracy, precision, recall, F1 score, and area under the receiver operating characteristic (ROC) curve (AUC)). A 5-fold cross-validation with grid-search was used to tune model hyperparameters to obtain the set of optimal hyperparameters that yielded the highest AUC. We tested multiple ANN models with varying configurations (number of hidden layers and type of regularization) to compare performance and reported the results for the ANN with the highest accuracy. We selected the best model based on an overall assessment of performance metrics and validated it on the held-out dataset to approximate external validation. We also conducted a feature importance analysis of the best model using SHAP (SHapley Additive exPlanations).

To ensure replicability and facilitate further research, the source code of all the presented machine learning models is available on GitHub[1].

TABLE I: Clinical patient characteristics. Data are mean [SD] or counts (%). P-values generated from t-test.

Variable	Diabetic	Non-diabetic	p-value
n (%)	4142 (48)	4460 (52)	–
Age at exam (years)	65 [12]	61 [14]	< 0.001
Gender (males)	2104 (51)	1885 (42)	< 0.001
sBP (mmHg)	131.2 [16.8]	129.0 [17.0]	< 0.001
Body Mass Index	31.6 [7.0]	29.1 [6.5]	< 0.001
LDL (mmol/L)	2.2 [0.9]	2.9 [0.9]	< 0.001
HDL (mmol/L)	1.2 [0.4]	1.4 [0.4]	< 0.001
HbA1c (mg/dL)	6.8 [1.])	5.7 [0.4]	< 0.001
Triglycerides (mmol/L)	1.7 [1.1]	1.4 [0.9]	< 0.001
Depression	828 (20)	981 (22)	0.007
Hypertension	2941 (71)	2368 (53)	< 0.001
Osteoarthritis	1364 (33)	1264 (28)	< 0.001
COPD	437 (11)	387 (28)	0.003
Hypertension Med	3422 (83)	2429 (54)	< 0.001
Corticosteroids	1182 (29)	1221 (27)	0.231

Missing data: n (D:ND): sBP 4 (2:2), LDL 52 (33:19), HDL 68 (20:48), TG 51 (17:34).

TABLE II: Classifiers and their tuned hyperparameters

Model	Hyperparameters	Parameters	Selections
LR	C	[0.01, 0.1, 1, 10]	1
	penalty	['l1', 'l2']	'l1'
	solver	['liblinear', 'saga']	'liblinear'
DT	max_depth	[5, 10, 15, 20]	5
	min_samples_leaf	[5, 10, 15, 20]	10
RF	n_estimators	[10, 50, 100, 200]	200
	max_depth	[5, 10, 15]	15
	min_samples_leaf	[15, 20, 25]	15
XGBoost	n_estimators	[50, 100, 200]	100
	learning_rate	[0.01, 0.1, 0.2]	0.1
	max_depth	[3, 5, 7]	3
	gamma	[0, 0.1, 0.2]	0.2
	reg_lambda	[0, 1, 5]	5
	reg_alpha	[0, 1, 5]	1
SVM	kernel	['rbf']	'rbf'
	C	[0.01, 0.1, 1, 10]	1
	gamma	[0.01, 1, 10]	0.01
	class_weight	[None, 'balanced']	'balanced'
CNB	GNB: smoothing	np.logspace(0, -9, 10)	0.001
	BNB: alpha	np.logspace(0, -9, 10)	1
ANN	layers	[1, 2]	2
	hidden_units	[64, 128, 256]	128
	lambda	[0.001, 0.01, 0.1]	0.01
	epochs	[500, 1000]	500
	learning_rate	[0.0001,0.001,0.01,0.1]	0.001
	LR_scheduling	[0.9]	0.9
	batch_size	[32, 64, 128]	64
	regularization	['dropout', 'l2']	'l2'

III. Results

A total of 4,412 (48%) diabetic and 4,460 (52%) non-diabetic patients were included in this dataset for a total of n = 8,872 unique adult patients. The patients in the diabetic dataset were 51% male, with a mean age of 65 years (standard deviation [SD] 12 years), and those in the non-diabetic dataset were 49% with a mean age of 61 (SD 14) (see Table I for participant demographics). The majority of diabetic individuals (83%) were prescribed hypertension medication, compared to nearly half (54%) in non-diabetic individuals. Patients with diabetes presented a higher average concentration of HbA1c (mg/dL) levels.

[1]https://github.com/andysontran/2024_IEEE_EMBC_Diabetes-I

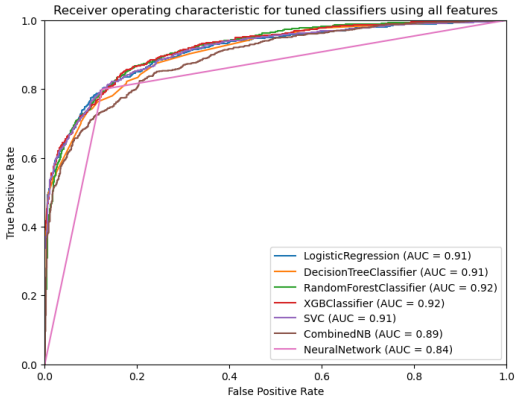

Fig. 1: ROC curves for tuned classifiers and neural network models on the test dataset.

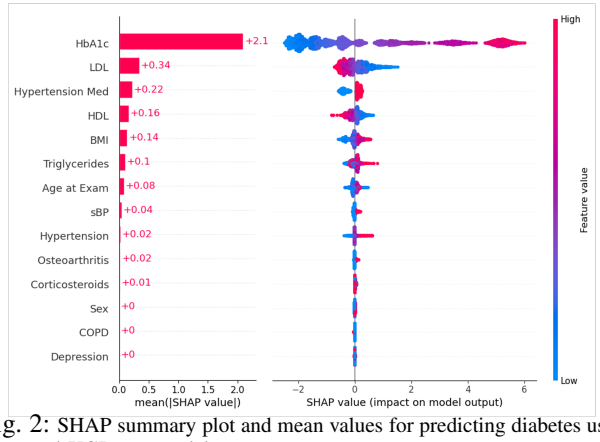

Fig. 2: SHAP summary plot and mean values for predicting diabetes using the tuned XGBoost model.

The optimal set of hyperparameters for each model is summarized in Table II. From the seven predictive models tested, we deemed the tuned XGBoost classifier as the best-performing model based on its high AUC (92%) (Figure 1; Table III). The XGBoost model was then tested on the held-out validation set. This yielded the following performance metrics in non-diabetic and diabetic patients: 80% precision (ND), 87% recall (ND), F1 score of 83% (ND), 87% precision (D), 79% recall (D), F1 score 83% (D), 83% accuracy, AUC 90%, and a macro-average of 83%. SHAP analysis of the XGBoost model revealed the five most influential features were HbA1c, LDL, hypertension medication, HDL, and BMI (mean SHAP values 2.10, 0.34, 0.22, 0.16, and 0.14, respectively) (Figure 2).

IV. Discussion

The main goal of this study was to determine the best model to predict diabetes prevalence by comparing the performance of seven ML models and identifying the most influential clinical patient characteristics in the model's predictions. The use of seven ML model architectures allowed for a comprehensive comparison of methods to evaluate predictive accuracy and recall, while limiting overfitting through the use of cross-validation. We observed consistent predictive performance of the tuned XGBoost classifier across the training, testing, and validation datasets. Furthermore, visual inspection of the ROC curves suggests that the XGBoost classifier has a smoother shape. This indicates less instability in its predictions and thus less overfitting. We identified XGBoost as the best-performing model based on these characteristics. The high AUC indicated the model performs well at distinguishing positive and negative diabetes classes, and the validation results mean the model can generalize well to real-world clinical settings compared to the other classifiers. These findings align with existing literature on diabetes prediction using ML. One study compared SVM, K-Nearest Neighbours (KNN), NB, DT, LR, and XGBoost algorithms for diabetes prediction using physical examination data and found XGBoost had the highest accuracy (81%), similar to our model's accuracy of 83% [21]. In another study by Li et al. (2020), the authors also used CPCSSN data to compare the performance of LR, Gradient Boosting Method (GBM), DT, and RF models in predicting diabetes, using parameters age, BMI, TG, FBS, sBP, HDL, and LDL [14]. They found GBM and LR outperformed RF and DT, with 85% and 84% AUC and 72% and 73% sensitivity, respectively [14]. Our XGBoost model has higher AUC and sensitivity (recall) scores than the other study's GBM model, as XGBoost is a more regularized form of GBM which contributes to improved model generalization capabilities [22]. Our models also included additional clinical features, which could contribute to improved performance.

The SHAP analysis the most important features for prediction. For example, a mean SHAP value of 2.10 for HbA1c implies that this feature contributes positively to the model's prediction across multiple instances (Figure 2). This may have useful clinical applications as these features could be prioritized for clinical monitoring of patients at risk of developing diabetes. When validating the model on unseen data, the AUC (91%) and accuracy (83%) remained high. This suggests the model generalizes fairly well, indicating that it has learned patterns that are applicable beyond the training data.

Our XGBoost model could potentially serve as a valuable tool for physicians in diagnosing or identifying the risk of diabetes in Canadian patients. The current standard procedures for identifying prediabetes typically rely on an FBS range of 6.1-6.9 [23]. However, HbA1c might offer a more comprehensive assessment. Unlike FBS, HbA1c reflects an average of blood sugar levels over the past 2-3 months, providing a more holistic view of glycemic control [24]. By incorporating HbA1c along with other traditionally significant features in diabetes onset, our model aims to enhance diagnostic accuracy. This refinement could potentially lead to more precise identification and management of individuals at risk of developing diabetes. Additionally, the results of our study are likely to be generalizable given the large, nationally representative sample.

Several limitations of this work should be considered. Firstly, this study was conducted retrospectively, therefore causation cannot be inferred based on the associations found by our predictive models. Secondly, data for this analysis was only captured until 2014, therefore it may not be reflective of current diabetes prevalence or risk factors. Thirdly, our

TABLE III: Comparison of performance metrics between LR, DT, RF, XGB, SVM, CNB and ANN algorithms in predicting diabetes prevalence in the test dataset.

Models	Non-Diabetic			Diabetic			Overall Model Performance		
	Precision	Recall	F1	Precision	Recall	F1	Accuracy	Macro Avg.	AUC
Logistic Regression	**84%**	86%	85%	84%	**81%**	82%	**84%**	**84%**	91%
Decision Tree	81%	**89%**	84%	**86%**	76%	81%	83%	83%	91%
Random Forest	82%	87%	85%	85%	79%	82%	**84%**	83%	**92%**
XGBoost	83%	87%	85%	84%	80%	82%	83%	83%	**92%**
Support Vector Machine	83%	87%	85%	85%	80%	**83%**	**84%**	**84%**	91%
Combined Naive Bayes	81%	83%	82%	81%	77%	79%	81%	81%	89%
Neural Network	83%	88%	**86%**	**86%**	80%	**83%**	**84%**	**84%**	84%

Note: Bold refers to highest metric(s) in column

dataset did not contain a race or ethnicity variable, which limited our ability to explore the impact of race or ethnicity on diabetes prevalence. Future work should use a timely database and investigate differential risks for various sex, ethnicity, and socioeconomic groups, as marginalized populations have been demonstrated to experience disproportionately higher diabetes risk [25].

V. Conclusion

This study used Canada-wide surveillance data to evaluate and compare seven machine learning models to predict diabetes prevalence and identify specific clinical characteristics most important in this prediction. Our analysis identified the XGBoost model as the top-performing model and HbA1c, LDL, hypertension medication, HDL, and BMI as the top five most influential features. This may be disseminated into clinical settings to assist with diagnosis and risk identification, which can inform patient care and resource allocation. For example, interventions for preventing diabetes may consider targeting risk factors through diet, exercise or medication. Earlier diagnosis of diabetes can reduce disease burden and overall health system costs.

References

[1] L. Fregoso-Aparicio, J. Noguez, L. Montesinos, and J. A. García-García, "Machine learning and deep learning predictive models for type 2 diabetes: A systematic review," *Diabetology & metabolic syndrome*, vol. 13, no. 1, pp. 1–22, 2021.

[2] F. Aguirre, A. Brown, N. Cho, *et al.*, *IDF Diabetes Atlas: Sixth edition*, English, Sixth. International Diabetes Federation, 2013, ISBN: 2-930229-85-3.

[3] D. Canada, *Diabetes Canada 2018 clinical practice guidelines for the prevention and management of diabetes in Canada*. Diabetes Canada, 2018.

[4] *Framework for diabetes in canada*, `https://www.canada.ca/en/public-health/services/publications/diseases-conditions/framework-diabetes-canada.html`, Accessed: 2024-01-25.

[5] A. G. LeBlanc, Y. J. Gao, L. McRae, and C. Pelletier, "At-a-glance-twenty years of diabetes surveillance using the canadian chronic disease surveillance system," *Health promotion and chronic disease prevention in Canada: research, policy and practice*, vol. 39, no. 11, p. 306, 2019.

[6] B. Anja and R. Laura, "The cost of diabetes in canada over 10 years: Applying attributable health care costs to a diabetes incidence prediction model," *Health promotion and chronic disease prevention in Canada: research, policy and practice*, vol. 37, no. 2, p. 49, 2017.

[7] P. Saeedi, I. Petersohn, P. Salpea, *et al.*, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas," *Diabetes research and clinical practice*, vol. 157, p. 107 843, 2019.

[8] E. Afsaneh, A. Sharifdini, H. Ghazzaghi, and M. Z. Ghobadi, "Recent applications of machine learning and deep learning models in the prediction, diagnosis, and management of diabetes: A comprehensive review," *Diabetology & Metabolic Syndrome*, vol. 14, no. 1, pp. 1–39, 2022.

[9] K. De Silva, W. K. Lee, A. Forbes, R. T. Demmer, C. Barton, and J. Enticott, "Use and performance of machine learning models for type 2 diabetes prediction in community settings: A systematic review and meta-analysis," *International journal of medical informatics*, vol. 143, p. 104 268, 2020.

[10] A. Tuppad and S. D. Patil, "Machine learning for diabetes clinical decision support: A review," *Advances in Computational Intelligence*, vol. 2, no. 2, p. 22, 2022.

[11] S. Ellahham, "Artificial intelligence: The future for diabetes care," *The American journal of medicine*, vol. 133, no. 8, pp. 895–900, 2020.

[12] V. Jaiswal, A. Negi, and T. Pal, "A review on current advances in machine learning based diabetes prediction," *Primary Care Diabetes*, vol. 15, no. 3, pp. 435–443, 2021.

[13] I. Contreras and J. Vehi, "Artificial intelligence for diabetes management and decision support: Literature review," *Journal of medical Internet research*, vol. 20, no. 5, e10775, 2018.

[14] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques," *BMC endocrine disorders*, vol. 19, pp. 1–9, 2019.

[15] K. Lu, P. Sheth, Z. L. Zhou, *et al.*, "Identifying prediabetes in canadian populations using machine learning," in *The IEEE Engineering in Medicine and Biology Society (EMBC)*, Under review, 2024.

[16] K. Samsel, A. Tiwana, S. Ali, *et al.*, "Predicting depression among canadians at-risk or living with diabetes using machine learning," in *2024 IEEE Engineering in Medicine and Biology Society (EMBC)*, Under review, 2024.

[17] P. Saha, Y. Marouf, H. Pozzebon, *et al.*, "Predicting time to diabetes diagnosis using random survival forests," 2024, Under review.

[18] *Canadian primary care sentinel surveillance network*, `https://cpcssn.ca/`, Accessed: 2024-01-25.

[19] T. Williamson, M. E. Green, R. Birtwhistle, *et al.*, "Validating the 8 cpcssn case definitions for chronic disease surveillance in a primary care database of electronic health records," *The Annals of Family Medicine*, vol. 12, no. 4, pp. 367–372, 2014.

[20] J. H. Lee and J. C. Huber Jr, "Evaluation of multiple imputation with large proportions of missing data: How much is too much?" *Iranian Journal of Public Health*, vol. 50, no. 7, p. 1372, 2021.

[21] M. Li, X. Fu, and D. Li, "Diabetes prediction based on xgboost algorithm," in *IOP conference series: materials science and engineering*, IOP Publishing, vol. 768, 2020, p. 072 093.

[22] M. Athanasiou, K. Sfrintzeri, K. Zarkogianni, A. C. Thanopoulou, and K. S. Nikita, "An explainable xgboost–based approach towards assessing the risk of cardiovascular disease in patients with type 2 diabetes mellitus," in *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, IEEE, 2020, pp. 859–864.

[23] M. R. Rooney, M. Fang, K. Ogurtsova, *et al.*, "Global prevalence of prediabetes," *Diabetes Care*, vol. 46, no. 7, pp. 1388–1394, 2023.

[24] P. Chamnan, R. K. Simmons, N. G. Forouhi, *et al.*, "Incidence of type 2 diabetes using proposed hba1c diagnostic criteria in the european prospective investigation of cancer–norfolk cohort: Implications for preventive strategies," *Diabetes care*, vol. 34, no. 4, pp. 950–956, 2011.

[25] F. Hill-Briggs, N. E. Adler, S. A. Berkowitz, *et al.*, "Social determinants of health and diabetes: A scientific review," *Diabetes care*, vol. 44, no. 1, p. 258, 2021.