2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3503-7149-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/EMBC53108.2024.10782174

Identifying Prediabetes in Canadian Populations Using Machine Learning

Katherine Lu[1,*], Paijani Sheth[1,*], Zhi Lin Zhou[1,*], Kamyar Kazari[2], Aziz Guergachi[2,3], Karim Keshavjee[2], Mohammad Noaeen[1], and Zahra Shakeri[2]

***Abstract*— Prediabetes is a critical health condition characterized by elevated blood glucose levels that fall below the threshold for Type 2 diabetes (T2D) diagnosis. Accurate identification of prediabetes is essential to forestall the progression to T2D among at-risk individuals. This study aims to pinpoint the most effective machine learning (ML) model for prediabetes prediction and to elucidate the key biological variables critical for distinguishing individuals with prediabetes. Utilizing data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN), our analysis included 6,414 participants identified as either nondiabetic or prediabetic. A rigorous selection process led to the identification of ten variables for the study, informed by literature review, data completeness, and the evaluation of collinearity. Our comparative analysis of seven ML models revealed that the Deep Neural Network (DNN), enhanced with early stop regularization, outshined others by achieving a recall rate of 60%. This model's performance underscores its potential in effectively identifying prediabetic individuals, showcasing the strategic integration of ML in healthcare. While the model reflects a significant advancement in prediabetes prediction, it also opens avenues for further research to refine prediction accuracy, possibly by integrating novel biological markers or exploring alternative modeling techniques. The results of our work represent a pivotal step forward in the early detection of prediabetes, contributing significantly to preventive healthcare measures and the broader fight against the global epidemic of Type 2 diabetes.**

I. Introduction

Type 2 diabetes (T2D) is a detrimental condition that affects millions of people worldwide [1]. Over 5 million Canadians had diabetes in 2023, with a projected increase of 26% in 10 years [2]. The precursor to T2D is prediabetes, a condition where blood glucose levels are elevated but below the diagnosis threshold for T2D [3]. Prediabetes is a common but reversible condition; according to an American Diabetes Association panel, up to 70% of individuals with prediabetes will develop diabetes [4]. Therefore, it is imperative to address prediabetes during the onset to prevent it from developing into T2D in the future.

Prediabetes is specifically defined by a fasting blood sugar (FBS) level between 6.1 and 6.9 mmol/L [5]. Many unhealthy lifestyle factors contribute to the development of T2D, so it is important to implement preventative measures such as exercise programs when individuals reach the prediabetes stage. Prediabetes also shares many of the same indicators as T2D. For example, hypertension [6], high total cholesterol [7], depression [8], glucocorticoids [9], chronic obstructive pulmonary disease (COPD) [10], osteoarthritis [11], and increased age have all been found to be associated with greater risk of developing T2D. Because risk factors of diabetes are well-known, many studies have applied machine learning (ML) models to clinical data to predict diabetes [12–16]. However, there are no Canadian studies developing prescriptive ML models to predict prediabetes. The onset of prediabetes is insidious, and most prediabetic individuals are unaware of their condition [17]. Therefore, accurately detecting prediabetes with ML models will improve health outcomes and reduce the burden on the healthcare system. Our study aims to identify the best ML model to predict prediabetes, and determine important factors in prediabetic individuals.

Building on the foundation laid by previous research, this study takes a novel approach by creating machine learning models specifically designed for the Canadian population to predict prediabetes. By identifying key variables and leveraging cutting-edge predictive analytics, we aim to advance early detection methods, ultimately fostering preventative measures against the transition from prediabetes to Type 2 diabetes. This effort not only aims to enhance awareness and management of prediabetes but also sets a precedent for utilizing technology-driven approaches in public health strategies, marking a significant step towards mitigating the global challenge of diabetes.

II. Methods

A. Data Source and Study Population

Our study uses data from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) [18]. Briefly, the CPCSSN compiles de-identified electronic medical record (EMR) data from 14 participating primary care networks across Canada. CPCSSN data is typically extracted from EMRs twice per year, and only structured data (e.g., not physician notes) is extracted. Details of CPCSSN data are described elsewhere [18, 19].

The data for our study is a random sample (n=10,000) of a CPCSSN subset that is comprised of patient records that precede the onset of diabetes. We restricted our study cohort to nondiabetic and prediabetic patients resulting in a final sample size of 6,414. Diabetic patients were excluded if they either had a fasting blood sugar (FBS) greater than 6.9 or an A1c level over 6.4% [5].

*These authors contributed equally to this work and share first authorship.
[1]Paijani Sheth, Katherine Lu, Zhi Lin Zhou, & Mohammad Noaeen are with the Dalla Lana School of Public Health, University of Toronto, Canada.
[2]Kamyar Kazari, 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[dot]shakeri[@]utoronto[dot]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.

B. Data Processing and Feature Engineering

With the available data, we selected a total of 10 variables based on existing literature supporting an association with diabetes [6–11]. This focus on diabetes, rather than prediabetes, was due to the limited evidence linking prediabetes with the available variables in the dataset. We then removed variables with over 50% missing observations. If variables were collinear, the most appropriate variable was selected; for example, total cholesterol includes low-density lipoprotein (LDL) cholesterol and high-density lipoprotein (HDL) cholesterol, so we kept total cholesterol and removed LDL and HDL. The final variables we included in our analysis were: age, body mass index (BMI), total cholesterol, depression status, hypertension status, presence of osteoporosis, chronic obstructive pulmonary disease (COPD), corticosteroid use, hypertension medication use (HMU), and sex. Corticosteroid and HMU were recategorized from a descriptive variable with all medication names to a binary variable of whether patients used any hypertension medications. The distributions, correlations, and missingness of all selected variables were explored. The class balance of the target, prediabetes, was also explored.

The outcome variable, presence of prediabetes, was created based on both FBS and A1c values. Patients were labeled as prediabetic if they had an FBS score between 6.1–6.9 mmol/L, or an A1c between 6.0–6.4% [5].

C. Prediction Models Development

We selected seven machine learning models to predict for prediabetes: logistic regression, Random Forest, XGBoost (extreme gradient boosting), mixed Naive Bayes, KNN (K-Nearest Neighbours), SVM (support vector machine), and Deep Neural Network (DNN). These models were chosen to explore model types with different strengths such as computational efficiency and robustness to overfitting, and for their proven performance in diagnosing prediabetes or T2D in previous studies [12, 20, 21]. The data engineering process for all models involved splitting the data into training (75%), validation (12.5%) and testing (12.5%) sets. The purpose of the validation set is to evaluate model performance post-training. The highest performing model from this process will subsequently be tested on a separate test set, which provides the final performance metrics. The test set adds an additional layer of generalizability. Following splitting, all data sets were imputed for missing values using the median. The only variable that had missing values were total cholesterol. Imputing missing data in general can be subject to bias if a group with similar characteristics has more missing data. However, only 2.4% of the data was missing, resulting in minimal potential bias by imputation. Median imputation, chosen for its preservation of the total cholesterol distribution and outlier resilience, was applied to the sole variable with missing values. Post-imputation, all datasets underwent normalization.

The DNN model's architecture involved defining aspects of the neuron and hidden layers. Before training, the data was stored into PyTorch tensors to improve efficiency during the forward pass and back propagation processes. The model was run on the training set, and the Binary Cross-Entropy Loss function between the predictions and true value was computed and visualized as loss over epochs. This was important in determining whether the model was learning over epochs. To reduce risk of overfitting, three DNN models with different types of regularization techniques were explored: L2 regularization, drop out, and early stopping. Hyperparameters were tuned manually. The highest performing parameters were: learning rate of 0.001, batch size of 25, 10 hidden units, and drop out probability of 0.5. Using the optimal hyperparameters, model performance was assessed on the validation set using all three regularization techniques and the best performing model was finally run on the test dataset.

Lastly, to identify the important variables in predicting prediabetes, a SHAP (SHapley Additive exPlanations) analysis was conducted on Google Colab and a SHAP dot plot was displayed. To ensure replicability and facilitate further research, the source code of all the presented machine learning models is available on GitHub[1].

TABLE I: CPCSSN subset patient characteristics stratified by prediabetes status. Results are presented as median [min, max] or n (%).

Variables	Nondiabetic	Prediabetic	Total
N (%)	3798 (59.2)	2616 (40.8)	6414 (100.0)
Age	60.0 [18.0,90.0]	66.0 [20.0,90.0]	63.0 [18.0,90.0]
BMI	27.8 [14.7,69.5]	29.9 [16.0,70.0]	28.6 [14.7,70.0]
Total Cholesterol	4.8 [2.1,12.7]	4.5 [0.8,8.8]	4.7 [0.8,12.7]
Depression (Y)	867 (22.8)	569 (21.8)	1436 (22.4)
Hypertension (Y)	1968 (51.8)	1859 (71.1)	3827 (59.7)
Osteoporosis (Y)	1070 (28.2)	895 (34.2)	1965 (30.6)
COPD (Y)	316 (8.3)	305 (11.7)	621 (9.7)
Corticosteroid (Y)	1036 (27.3)	793 (30.3)	1829 (28.5)
HMU (Y)	2039 (53.7)	2010 (76.8)	4049 (63.1)
Sex (Female)	2196 (57.8)	1370 (52.4)	3566 (55.6)

III. Results and Discussion

Patient demographics are illustrated in Table I. Among all variables selected, only total cholesterol had missing data. Our sample had 59% nondiabetic patients and 41% prediabetic patients, indicating no issue of class imbalance. All continuous variables (age, BMI, total cholesterol) demonstrated normal distributions. For prediabetic patients, age and BMI were slightly higher while total cholesterol was slightly lower compared to nondiabetic patients. A correlation matrix of the continuous variables showed that age and BMI had slight correlations with prediabetes (correlation coefficient around 0.1-0.3), and no correlation with total cholesterol.

A. Prediction Models Results

Final model evaluation metrics are presented in Table II. For model evaluation, we produced and compared the following scores: accuracy, precision, recall, F1-score, and AUC. However, recall values were primarily used when assessing the performance because our goal is to minimize the number of false negatives in our study. The DNN model performed the best with a recall value of 60%, followed by logistic regression (57%), Random Forest (52%), XGBoost (51%), KNN (38%) and Naive Bayes (23%).

[1] https://github.com/lilyzhizhou/Identifying-Prediabetes-ML.git

TABLE II: Performance metrics for each type of machine learning model using test sets for predicting prediabetes.

Models	TP	FP	FN	TN	Acc	Precision	Recall	F1	AUC
LR	215	308	164	596	0.63	0.41	0.57	0.48	0.61
RF	350	101	324	829	**0.74**	**0.78**	0.52	0.62	0.71
XGBoost	272	190	265	556	0.65	0.59	0.51	0.54	0.63
NB	69	51	232	450	0.65	0.58	0.23	0.33	0.56
DNN	198	87	134	383	0.72	0.69	**0.60**	**0.64**	**0.76**
KNN	130	210	210	260	0.49	0.38	0.38	0.38	0.66
SVM	254	269	166	594	0.66	0.60	0.49	0.54	0.63

Note: Bold indicates the highest value of a metric in the column.

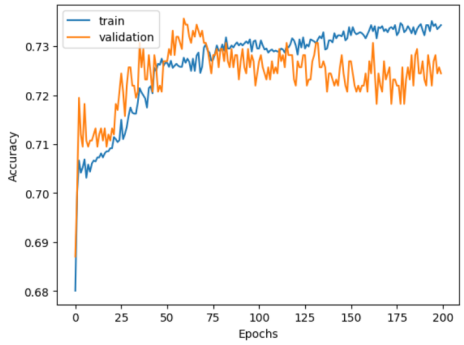

Fig. 1: Performance plots for the DNN model for prediabetes prediction

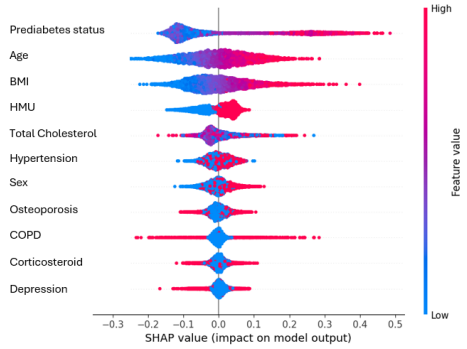

Fig. 2: SHAP scatter plot for prediabetes prediction DNN features.

Using the optimal hyperparameters for the DNN model, the loss over 2000 epochs plot showed a sudden decline within the first 50 epochs before reaching a plateau around 50%. We implemented 3 regularization techniques: L2 regularization, dropout regularization, and early stopping regularization. Early stop regularization yielded the highest performance, with a validation recall of 53% after 200 epochs. Furthermore, the recall improved over epochs, suggesting that the model learned. Therefore, an early stop regularization technique was used when assessing the performance on the test data. The test recall ended up being 60% and there was a slight overall increase in performance over 200 epochs. We opted for 200 epochs instead of 2000, as used in the loss over epochs plot, because this amount was sufficient to capture the highest performance gains. Extending the training to 2000 epochs resulted in diminishing returns on recall and introduced more noise as training progressed. To assess the generalizability of the model and its learned state, train and test accuracies over 200 epochs were visualized (Figure 1). Figure 2 displays the SHAP dot plot for our selected features from highest to lowest importance in predicting prediabetes.

B. Discussion

After conducting the exploratory process, it was found that the majority of the categorical features were relatively balanced between the prediabetic and nondiabetic groups (Table I). The continuous variables, median age and BMI, were higher in prediabetic individuals compared to nondiabetic individuals. Furthermore, a correlation matrix between the continuous variables suggested minimal correlation among covariates with prediabetes.

The DNN model with early stop regularization had the best overall performance, having the highest recall, F1 score, and AUC (Table II). However, the primary performance metric of this research was recall, which aims to minimize the number of false negatives and minimize the number of prediabetic individuals left undiagnosed. Failing to identify prediabetic individuals would result in a higher risk of them developing diabetes, and interventions to manage the disease become more invasive and costly. Looking solely at accuracy, the Random Forest model appeared best performing (accuracy = 0.74), but its low recall of 0.52 indicated that more individuals with prediabetes were falsely diagnosed as healthy. The Neural Network model, on the other hand, exhibited an accuracy of 0.71 and a recall of 0.60 (Table II), which underscored its superior performance among the evaluated models. Despite this, we observed notable limitations in its application. Firstly, there was evidence of some overfitting to the training data, as illustrated in Figure 1. While both test and training accuracy increased and stabilized in tandem, suggesting effective learning from the training dataset, the presence of more pronounced peaks and dips in test accuracy indicates potential overfitting. This pattern suggests that the model may not generalize as effectively to unseen data. Additionally, there is low predictive power, evidenced by the suboptimal performance across all models. This aspect further highlights the challenges in achieving high accuracy in prediabetes prediction, underlining the need for continued refinement of machine learning approaches in this domain.

Our results are in contrast with some existing literature that use similar ML models to predict prediabetes/diabetes with better performance. For example, Lai et al. [12] obtained a recall of 73.4 using logistic regression to predict diabetes, and Choi et al. [22] obtained a recall of 74.3 using SVM to predict prediabetes. A potential reason for the worse performance of our models could be that biological characteristics of nondiabetic and prediabetic individuals may overlap more compared to nondiabetic and diabetic individuals, making it more difficult to predict prediabetes. Another explanation could be that the biological features we selected were not adequate for distinguishing those with prediabetes from nondiabetics.

The results of our SHAP analysis, on the other hand, align with findings from current literature. We identified age, BMI, and hypertension medication use as the top three important features for predicting prediabetes. A 2016 study highlighted that individuals with prediabetes typically have a higher BMI [23]. Moreover, advanced age and hypertension are also well-

documented factors associated with prediabetes [24, 25].

The rapid expansion of ML applications for healthcare has prompted discussion regarding the unique ethical concerns of ML. For instance, if the training data for an ML model does not include different racial groups, the model may overlook biological differences that affect disease diagnosis and presentation [26]. Such oversights can introduce bias in healthcare and exacerbate health inequities. Diversity, equity, and inclusion (DEI) is an important consideration for diabetes research because thresholds and symptoms of diabetes and prediabetes can vary across populations. In Canada's diverse population, studies have highlighted an association between lower income levels and certain racial groups with an increased diabetes risk [27, 28]. A review of relevant literature in Medline and Embase revealed a paucity of discussions on DEI within the context of machine learning applications in diabetes research. Our dataset's exclusion of race/ethnicity information precludes DEI exploration. Enhancing DEI in diabetes ML research necessitates the inclusion of comprehensive demographic data for analysis.

IV. CONCLUSION

An enhanced machine learning model for prediabetes diagnosis is essential for the early detection of individuals at risk for T2D. Such advancements enable healthcare professionals to implement proactive preventative interventions, thereby mitigating the onset and progression of diabetes-related complications. Our study aimed to leverage machine learning models to detect prediabetes within Canadian populations, utilizing data from the CPCSSN dataset. We sought to determine the key clinical variables for identifying individuals with prediabetes. Among the seven machine learning models tested, DNN demonstrated the highest performance. Age, BMI, and the use of hypertension medication emerged as the most influential features in predicting prediabetes, aligning with findings in existing literature. Future research should focus on enhancing model accuracy by incorporating a broader array of variables, such as demographics, dietary habits, physical activity levels, psychological factors, and smoking status, and considering sex-specific differences due to varying biomarker levels. To validate further, traditional clinical risk assessment tools can be employed as well.

REFERENCES

[1] M. Guasch-Ferré *et al.*, "Metabolomics in prediabetes and diabetes: A systematic review and meta-analysis," *Diabetes care*, vol. 39, no. 5, pp. 833–846, 2016.

[2] Diabetes Canada, "Diabetes in canada: Backgrounder," *Ottawa: Diabetes Canada*, 2023.

[3] N. Bansal, "Prediabetes diagnosis and treatment: A review," *World journal of diabetes*, vol. 6, no. 2, p. 296, 2015.

[4] A. G. Tabák, C. Herder, W. Rathmann, E. J. Brunner, and M. Kivimäki, "Prediabetes: A high-risk state for developing diabetes," *Lancet*, vol. 379, no. 9833, p. 2279, 2012.

[5] Z. Punthakee, R. Goldenberg, and P. Katz, "Definition, classification and diagnosis of diabetes, prediabetes and metabolic syndrome," *Canadian journal of diabetes*, vol. 42, S10–S15, 2018.

[6] V. Tsimihodimos, C. Gonzalez-Villalpando, J. B. Meigs, and E. Ferrannini, "Hypertension and diabetes mellitus: Coprediction and time trajectories," *Hypertension*, vol. 71, no. 3, pp. 422–428, 2018.

[7] C. Jing *et al.*, "The ability of baseline triglycerides and total cholesterol concentrations to predict incidence of type 2 diabetes mellitus in chinese men and women: A longitudinal study in qingdao, china," *Biomedical and Environmental Sciences*, vol. 32, no. 12, pp. 905–913, 2019.

[8] B. Mezuk, W. W. Eaton, S. Albrecht, and S. H. Golden, "Depression and type 2 diabetes over the lifespan: A meta-analysis," *Diabetes care*, vol. 31, no. 12, pp. 2383–2390, 2008.

[9] H. E. Tamez-Pérez, D. L. Quintanilla-Flores, R. Rodríguez-Gutiérrez, J. G. González-González, and A. L. Tamez-Peña, "Steroid hyperglycemia: Prevalence, early detection and therapeutic recommendations: A narrative review," *World journal of diabetes*, vol. 6, no. 8, p. 1073, 2015.

[10] S. Gläser, S. Krüger, M. Merkel, P. Bramlage, and F. J. Herth, "Chronic obstructive pulmonary disease and diabetes mellitus: A systematic review of the literature," *Respiration*, vol. 89, no. 3, pp. 253–264, 2015.

[11] K. Louati, C. Vidal, F. Berenbaum, and J. Sellam, "Association between diabetes mellitus and osteoarthritis: Systematic literature review and meta-analysis," *RMD open*, vol. 1, no. 1, e000077, 2015.

[12] 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.

[13] T. A. Andargie, B. Mengistu, L. D. Baffa, K. A. Gonete, and A. K. Belew, "Magnitude and predictors of pre-diabetes among adults in health facilities of gondar city, ethiopia: A cross-sectional study," *Frontiers in Public Health*, vol. 11, p. 1 164 729, 2023.

[14] K. Samsel *et al.*, "Predicting depression among canadians at-risk or living with diabetes using machine learning," in *medRxiv*, 2024.

[15] P. Saha *et al.*, "Predicting time to diabetes diagnosis using random survival forests," in *Proceedings of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024.

[16] K. Esser *et al.*, "Predicting diabetes in canadian adults using machine learning," in *Proceedings of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024.

[17] Centers for Disease Control and Prevention (CDC), *The surprising truth about prediabetes*, Jul. 2022. (visited on 10/31/2023).

[18] *Canadian primary care sentinel surveillance network*, Available online at https://cpcssn.ca/. (visited on 01/04/2024).

[19] S. Garies, R. Birtwhistle, N. Drummond, J. Queenan, and T. Williamson, "Data resource profile: National electronic medical record data from the canadian primary care sentinel surveillance network (cpcssn)," *International Journal of Epidemiology*, vol. 46, no. 4, 1091–1092f, 2017.

[20] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics*, vol. 7, p. 21, 2013.

[21] Z. Zhang, "Introduction to machine learning: K-nearest neighbors," *Annals of translational medicine*, vol. 4, no. 11, 2016.

[22] S. B. Choi *et al.*, "Screening for prediabetes using machine learning models," *Computational and mathematical methods in medicine*, vol. 2014, 2014.

[23] Z. Gholi, M. Heidari-Beni, A. Feizi, B. Iraj, and G. Askari, "The characteristics of pre-diabetic patients associated with body composition and cardiovascular disease risk factors in the iranian population," *Journal of research in medical sciences: the official journal of Isfahan University of Medical Sciences*, vol. 21, 2016.

[24] M. H. Alijanvand, A. Aminorroaya, I. Kazemi, M. Amini, S. A. Yamini, and M. Mansourian, "Prevalence and predictors of prediabetes and its coexistence with high blood pressure in first-degree relatives of patients with type 2 diabetes: A 9-year cohort study," *Journal of Research in Medical Sciences: The Official Journal of Isfahan University of Medical Sciences*, vol. 25, 2020.

[25] Z. Yan, M. Cai, X. Han, Q. Chen, and H. Lu, "The interaction between age and risk factors for diabetes and prediabetes: A community-based cross-sectional study," *Diabetes, Metabolic Syndrome and Obesity*, pp. 85–93, 2023.

[26] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, "Ensuring fairness in machine learning to advance health equity," *Annals of internal medicine*, vol. 169, no. 12, pp. 866–872, 2018.

[27] Z. Lysy, G. L. Booth, B. R. Shah, P. C. Austin, J. Luo, and L. L. Lipscombe, "The impact of income on the incidence of diabetes: A population-based study," *Diabetes research and clinical practice*, vol. 99, no. 3, pp. 372–379, 2013.

[28] N. A. Khan *et al.*, "Ethnicity and sex affect diabetes incidence and outcomes," *Diabetes care*, vol. 34, no. 1, pp. 96–101, 2011.