Predicting Depression Among Canadians At-Risk or Living with Diabetes Using Machine Learning

Konrad Samsel[1,*], Amrit Tiwana[1,*], Sarra Ali[1], Aryan Sadeghi[2], Aziz Guergachi[2,3], Karim Keshavjee[2], Mohammad Noaeen[1], and Zahra Shakeri[2]

***Abstract*— Depression is disproportionately prevalent among individuals with diabetes compared to the general populace, underscoring the critical need for predictive mechanisms that can facilitate timely interventions and support. This study explores the use of machine learning to forecast depression in those at risk or diagnosed with diabetes, leveraging the extensive primary care data from the Canadian Primary Care Sentinel Surveillance Network. Six machine learning models including Logistic Regression, Random Forest, AdaBoost, XGBoost, Naive Bayes, and Artificial Neural Networks were trained and evaluated on their ability to predict depression. XGBoost emerged as the most effective model with an AUC of 0.70 on the test data. Sex, age, osteoarthritis, A1c levels, and body mass index emerged as the key contributors to the best-performing model's predictive ability. While the study navigated through the constraints of limited demographic information and potential label bias, it lays a foundational premise for subsequent longitudinal studies aimed at refining depression prediction within this specific clinical cohort.**

I. INTRODUCTION

Depression is a significant comorbidity among individuals at risk or living with diabetes. Continuous self-monitoring, the need for lifestyle adjustments, and consistent anxiety about blood sugar management can result in extended periods of emotional distress [1]. Previous research found the prevalence of depression to be around two to three times higher among individuals with diabetes than among the general population [2, 3]. Additionally, prediabetic individuals with high levels of self-reported depressive symptoms have double the risk of diabetes diagnosis compared to those with lower reported levels [4]. The challenges posed by diabetes management may contribute to the development of depressive symptoms, while the emotional toll of depression could potentially elevate the risk of diabetes onset [5, 6]. This bidirectional link emphasizes the need for a holistic approach in addressing both conditions to enhance overall health outcomes.

The coexistence of diabetes and depression can result in a significant physical and emotional burden on the affected individual and impact the management of this progressive condition. For instance, depression can worsen diabetes prognosis, increase non-compliance to treatments, and negatively impact one's quality of life [7, 8]. The increasing prevalence of diabetes and its associated complications present a significant challenge for healthcare systems. Per capita expenses for patients with diabetes can be two to four times higher than those without diabetes with a significant amount of these costs attributed to the management of comorbidities [9]. Given these challenges, understanding the relationship between diabetes and depression becomes essential for developing effective interventions, improving patient outcomes, and alleviating the economic burden on healthcare systems.

Despite knowledge of the increased incidence of depression among diabetic patients, it has been estimated that only 25-50% of patients meeting the criteria for this mood disorder receive diagnosis and treatment [10]. This highlights a critical gap in identifying and managing mood disorders among patients at risk of or living with diabetes. We sought to explore the efficacy of modelling depression outcomes using cross-sectional electronic medical data from primary care practices across Canada [11–14].

This study will evaluate the ability to use different machine learning approaches to identify depression cases among diabetic and diabetic patients, considering physician-diagnosed depression as the reference standard. In addition, we will use feature selection methods and SHAP (SHapley Additive exPlanations) analysis to identify the most influential variables among candidate models [15]. Through this analysis, we aim to provide early insights for future work exploring the early detection of depression in individuals at risk or living with diabetes, with such efforts offering the potential for timely interventions and support.

II. METHODS

A. Data Collection and Preparation

The dataset (N=10,000) used for this analysis was obtained from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). This data source included de-identified health information from electronic medical records at participating primary care settings across Canada from 2003 to 2015 [12]. These cross-sectional records included 47 variables, encompassing basic demographic information, anthropometric measures, diagnosis of comorbidities, corticosteroid and hypertension prescription use, and biomarkers of blood lipid and hemoglobin A1c levels. The data was subsetted to only include those living with diabetes, defined as controlled or uncontrolled diagnosed diabetes mellitus

*These authors contributed equally to this work and share first authorship.
[1]Konrad Samsel, Amrit Tiwana, Sarra Ali, and Mohammad Noaeen are with the Dalla Lana School of Public Health, University of Toronto, Canada.
[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.

type 1 and type 2 [16], or prediabetes, defined by the indication of a recent hemoglobin A1c value greater than or equal to 5.7 [17]. The inclusion of observations with a confirmed diagnosis of diabetes or those with an A1c value meeting the prediabetic threshold resulted in a final sample size of N=7,862.

The data was explored for external representativity, variable distribution, missing data, collinearity, and class imbalance. In exploring possible selection bias that may impact the generalizability of the model to clinical populations, we examined the distribution of the data by gender, age, and diabetes status. Skew and outliers were assessed in continuous variables using box plots, histograms, and quantile-quantile plots. To assess missing data patterns, we examined differences in age, sex, body mass index (BMI), diabetes, and depression status among observations with missing records. We also used a correlation matrix to check for feature collinearity. We examined feature pairs with collinearity above 0.7 and removed the one with the most missing data.

New features were generated to better structure the dataset for machine learning tasks. This included converting hypertension and corticosteroid medication use from a textual format to a series of binary variables and calculating the duration that an individual has lived with other comorbidities. Medications with fewer than ten instances within the data were excluded, resulting in 45 new binary medication variables. Next, we determined the duration of each comorbidity by subtracting the date of diagnosis from the timestamp denoting a patient's most recent clinical interaction. This calculation quantifies years lived with a specific condition, coding undiagnosed cases as zero.

Following feature engineering, the data was prepared for model training and testing using the Scikit-learn package in Python [18]. Data were split into training (80%) and testing (20%) sets, with the label being the binary classification of depression diagnoses. Given the small amount of missing data and lack of evidence of the data being missing at random, multiple imputations by chained equation (MICE) from the impute package were used to impute missing data [19]. Next, all features were normalized, and a combination of TomekLinks undersampling and Synthetic Minority Oversampling TEchnique (SMOTE) was used to address the class imbalance. Imputation and normalization occurred after data splitting to avoid possible information leakage, and only the training set underwent class re-balancing.

B. Model Development

This analysis included the development of six different machine learning algorithms, including Logistic Regression (LR), Naive Bayes (NB), Random Forest (RF), AdaBoost (AB), XGBoost (XGB), and artificial neural networks (ANN). SelectFromModel (SFM), a feature selection technique, was applied to all models except NB and ANN. In contrast to other approaches such as Recursive Feature Elimination with Cross-Validation (REFCV), SelectFromModel (SFM) was chosen for its computational efficiency in feature selection.

Using selected features for each model, a grid search using cross-validation was then used to tune each model's hyperparameters. For LR, a grid search was performed to identify the ideal regularization technique and alpha value, determined through the mean MSE. Mean F1 validation scores were used for the remaining models, excluding NB and ANN. Similarly, grid search was used in the RF model to identify the best-performing hyperparameters for *'n_estimators', 'max_depth', 'min_samples_leaf', 'min_samples_split', and 'max_features'*. For AB and XGB, *'n_estimators'* and *'learning_rate'* were tuned using the same methodology. We employed a neural network without the use of hyperparameter tuning or feature selection given computational constraints. Incorporating L2 regularization to mitigate overfitting, the ANN featured five hidden layers with 40 neurons each, trained for 400 epochs using Adam optimization and a learning rate of 0.0001.

The final version of each model was then evaluated on training data and the unseen test data. Model performance was assessed by AUC, F1 (weighted average), precision and recall scores. The comparison of these scores among training and test sets allowed for the assessment of model overfitting. The F1 score is valuable when class imbalance is evident and provides an overall performance metric for binary classification. To explore the broader context of our work, we performed a literature scan to identify the extent to which concepts related to equity, diversity, and inclusion (EDI) are included in similar work. Through OVID Embase, we identified 58 research papers focused on the application of machine learning in the diagnosis or prognosis of comorbidities among diabetic (or prediabetic) adults. After extracting abstracts and discussions from each paper, we employed Word2Vec embeddings to explore contextual associations in EDI terminology. To promote replicability and support future research endeavors, the source code for all machine learning models discussed in this study has been made publicly accessible on GitHub[1].

III. RESULTS

About one-fifth of the sample has depression (20.6%). Most of the sample has diabetes (65.4%) compared to prediabetes (34.6%). The mean age is 45 years (sd = 12.3 years) and the mean BMI is 30.9 kg/m^2 (sd = 6.8 kg/m^2). There are slightly more females (52.3%) than males (47.7%). Hypertension is a prevalent condition in the majority of the sample, with a prevalence rate of 69.2%, and 76.9% taking at least one hypertension medication. Additional patient characteristics are detailed in **Table 1**.

Approximately 3% of observations contained at least one missing data value, with consistent distributions across key demographic variables, including age, sex, BMI, diabetes status, and depression status. This suggests that the missing data pattern is not likely to be MAR. Most continuous variables displayed a normal distribution, except for triglyceride levels, which displayed a left skew. Categorical variables demonstrated sufficient variability when stratified by depression,

[1]https://github.com/tiwanaam/mlforhealthdata

TABLE I: Patient characteristics are presented as mean [sd] for continuous variables or n (%) for categorical variables.

Variable	Total N = 7,862
Age, years	64.9 [12.3]
BMI, kg/m^2	30.9 [6.8]
Sex, female	4,110 (52.3)
Hemoglobin A1c, %	6.5 [0.9]
Fasting blood sugar level, mmol/L	6.6 [1.8]
Total cholesterol, mmol/L	4.6 [1.2]
Missing	149 (1.9)
Depression (Yes)	1620 (20.6)
Hypertension (Yes)	5,439 (69.2)
Years living with hypertension	2.1 [2.5]
Osteoarthritis (Yes)	2,628 (33.4)
Years living with osteoarthritis	2.1 [2.5]
COPD (Yes)	828 (10.5)
Years living with COPD	1.8 [2.0]
Diabetes (Yes)	5,139 (65.4)
Years living with diabetes	2.8 [2.5]
Take at least one hypertension medication	6,042 (76.9)
Take at least one corticosteroid medication	2,298 (29.2)

the outcome of interest. Based on the correlation matrix, we observed a high correlation between A1c and fasting blood sugar levels and between low-density lipoprotein levels and total cholesterol. We opted to exclude fasting blood sugar levels rather than A1c, considering its relevance as a marker for blood sugar control over the past 3 months.

Among the six models evaluated on the unseen test data, XGB had the highest AUC and F1 scores, at 0.70 and 0.73, respectively (**Fig. 1, Table 2**). The XGBoost model utilized 28 features selected using SelectFromModel, with a grid search informing the selection of hyperparameters, which included a learning rate of 0.05 with 80 estimators. The second-best performing model was LR, which utilized 21 selected features and employed L1 regularization with a C-value of 0.068. Interestingly, the logistic regression model displayed a low degree of overfitting, with similar AUC values seen in training and test sets (0.70 and 0.69, respectively).

Finally, as illustrated in **Fig. 2**, we utilized SHAP to assess the contribution of features for the best-performing model (XGBoost). The top five most influential features included sex, age, osteoarthritis, AIc levels, and BMI. Lower ages, lower A1c levels, and higher BMI positively contributed to the model's performance on unseen test data.

TABLE II: Performance Evaluation of Models in Training and Test Sets

Model	Training		Testing			
	F1 Score	AUC	Prec.	Recall	F1 Score	AUC
LR	0.64	0.70	0.76	0.62	0.65	0.69
NB	0.65	0.65	0.76	0.65	0.68	0.64
RF	0.75	0.83	0.75	0.69	0.71	0.69
AB	0.74	0.83	0.75	0.70	0.72	0.68
XGB	0.73	0.81	**0.75**	**0.71**	**0.73**	**0.70**
ANN	0.74	0.85	0.74	0.51	0.54	0.68

IV. DISCUSSION

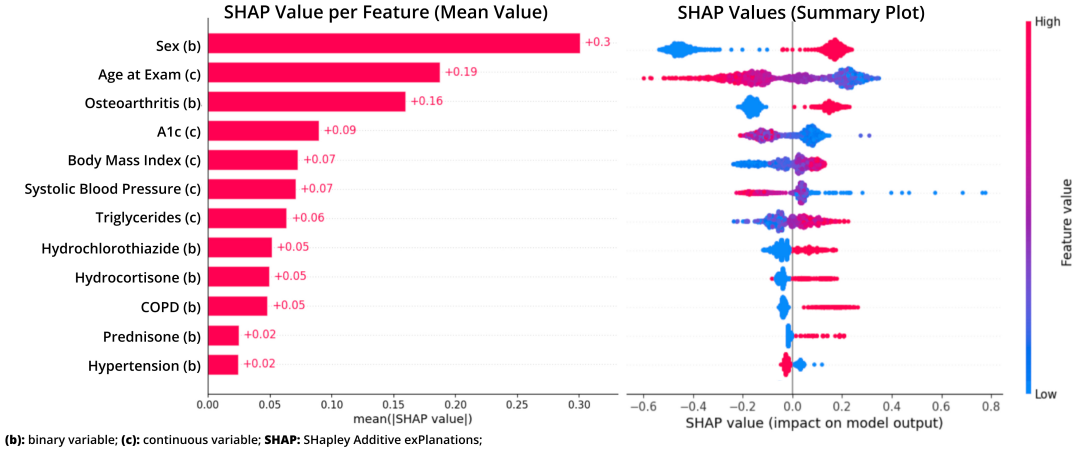
Our models displayed a moderate ability to identify depression among diabetic and pre-diabetic patients using routine primary care data. The XGB model had the highest AUC (0.70) and weighted F1 score (0.73), with a moderate degree of overfitting when comparing AUC values between training and test sets. The LR model had the second-highest AUC (0.69), with a lower degree of overfitting likely attributable to the use of regularization techniques. While feature engineering was employed to better capture the impact of comorbidities among this cohort, these variables did not substantially augment model performance. As identified through the SHAP analysis of XGBoost, non-engineered features had a higher mean absolute value.

Earlier research has demonstrated that individuals with diabetes tend to report lower self-perceived health, diminished psychological well-being, and reduced overall quality of life compared to those without diabetes. Contributing factors to this disparity include being female, experiencing depression, lack of physical activity, and obesity [20]. Another study suggested a tenuous correlation between HbA1c levels and quality of life. However, the presence of depressive symptoms in Type 2 Diabetes Mellitus is notably linked to poorer health status and diminished quality of life [21]. Hence, emphasizing the importance of addressing both diabetes and depression is crucial for enhancing individuals' quality of life and their perception of health status.

Drawing on primary care data from CPCSSN, this research offers key insights into the ability to identify depression within a diverse adult population [12]. While our study provides an exploration of the performance of multiple discriminative models in predicting depression, it is essential to acknowledge limitations. The absence of demographic data limited our ability to assess the potential under-representation of certain groups in the data. Moreover, the results may be impacted by label bias, which in this case, relates to a potential discrepancy between the actual prevalence of depression in the population and the cases that were recognized and diagnosed by clinicians. Using the results of screening surveys employed in clinics, such as PHQ-9 may help reduce potential label bias and the reliance on clinical diagnoses as the 'gold standard'. Lastly, the cross-sectional nature of the dataset limits our ability to explore how changes over time affect the prediction of depression. Future explorations could utilize longitudinal datasets and different model architectures to better investigate prognostic models and inform meaningful interventions.

Given that physician-diagnosed depression served as the study's ground truth, we investigated the impact of antidepressant use on model performance. We considered the influence of depression treatment effects, which may have confounded our ability to detect depression cases. Having identified and matched the full medication histories for 8.8% of our study sample through additional data sources, we discovered that the use of prescribed medications typically employed for depression treatment was similar among those with and without depression (5.8% and 6.1%, respectively). While we did not have enough data to confidently report model performance stratified by the use of these medications, this may be an interesting area of further study.

Our EDI analysis found discussions around bias, inclusion, diversity, race, and ethnicity within machine learning health research, emphasizing the importance of a comprehensive



(b): binary variable; **(c):** continuous variable; **SHAP:** SHapley Additive exPlanations;

Fig. 1: SHAP Analysis on XGBoost model. The top 12 features are displayed.

understanding of contextual usage and acknowledging the multifaceted nature of these discussions. The positive sentiment associated with terms like 'minority' suggests a potential shift in narrative or emphasis, contributing to ongoing efforts in fostering EDI within the scientific literature on machine learning health research.

V. CONCLUSION

Considering the need for early intervention and mental health support for individuals at risk or living with diabetes, this study explored the predictive capability of multiple machine learning models in predicting depression. The findings reveal that XGBoost was the most effective machine learning model, acheiving an AUC of 0.70 on unseen test data. Despite efforts in feature engineering, non-engineered variables including sex, age, osteoarthritis diagnosis, A1c level, and BMI were most influential in augmenting predictive performance. While this study leveraged a diverse dataset and innovative methodologies, limitations include limited demographic data and potential label bias, prompting the need for future longitudinal research for a more nuanced exploration of the relationship between diabetes and depression.

REFERENCES

[1] D. Canada. "How diabetes can affect your mental health." (Accessed November 1, 2023), [Online]. Available: https://www.diabetes.ca/en-CA/archive/managing-my-diabetes---archive/preventing-complications/mental-health.

[2] S. V. Bădescu, C. Tătaru, L. Kobylinska, *et al.*, "The association between diabetes mellitus and depression," *J Med Life*, vol. 9, no. 2, 120–125, 2016.

[3] J.-Y. Lee, D. Won, and K. Lee, "Machine learning-based identification and related features of depression in patients with diabetes mellitus based on the korea national health and nutrition examination survey: A cross-sectional study," *PLoS One*, vol. 18, no. 7, e0288648, 2023.

[4] M. Virtanen, J. E. Ferrie, A. G. Tabak, *et al.*, "Psychological distress and incidence of type 2 diabetes in high-risk and low-risk populations: The whitehall ii cohort study," *Diabetes Care*, vol. 37, no. 8, 2091–2097, 2014.

[5] R. I. G. Holt and W. J. Katon, "Dialogue on diabetes and depression: Dealing with the double burden of co-morbidity," *J Affect Disord*, vol. 142, no. Suppl, S1–3, 2012.

[6] R. I. G. Holt, M. de Groot, and S. H. Golden, "Diabetes and depression," *Curr Diab Rep*, vol. 14, no. 6, p. 491, 2014.

[7] J. S. Gonzalez, M. Peyrot, L. A. McCarl, *et al.*, "Depression and diabetes treatment nonadherence: A meta-analysis," *Diabetes Care*, vol. 31, no. 12, 2398–2403, 2008.

[8] L. E. Egede, P. J. Nietert, and D. Zheng, "Depression and all-cause and coronary heart disease mortality among adults with and without diabetes," *Diabetes Care*, vol. 28, no. 6, 1339–1345, 2005.

[9] M. Khaledi, F. Haghighatdoost, A. Feizi, *et al.*, "The prevalence of comorbid depression in patients with type 2 diabetes: An updated systematic review and meta-analysis on a huge number of observational studies," *Acta Diabetol*, vol. 56, no. 6, 631–650, 2019.

[10] Centers for Disease Control and Prevention. "Diabetes and mental health — cdc." Accessed December 8, 2023. (2023), [Online]. Available: https://www.cdc.gov/diabetes/managing/mental-health.html.

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

[12] S. Garies, R. Birtwhistle, N. Drummond, *et al.*, "Data resource profile: National electronic medical record data from the canadian primary care sentinel surveillance network (cpcssn)," *Int J Epidemiol*, vol. 46, no. 4, 1091–1092f, 2017.

[13] K. Esser *et al.*, "Predicting diabetes in canadian adults using machine learning algorithms," in *The IEEE Engineering in Medicine and Biology Society (EMBC)*, Under review, 2024.

[14] P. Saha *et al.*, "Predicting time to diabetes diagnosis using random survival forests," 2024, Under review.

[15] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *NeurIPS*, vol. 30, 2017, Accessed January 23, 2024. [Online]. Available: https://api.semanticscholar.org/CorpusID:21889700.

[16] 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," *Ann Fam Med*, vol. 12, no. 4, 367–372, 2014.

[17] C. Lorenzo, L. E. Wagenknecht, A. J. G. Hanley, *et al.*, "A1c between 5.7 and 6.4 percent as a marker for identifying pre-diabetes, insulin sensitivity and secretion, and cardiovascular risk factors: The insulin resistance atherosclerosis study (iras)," *Diabetes Care*, vol. 33, no. 9, 2104–2109, 2010.

[18] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python," *MACHINE LEARNING IN PYTHON*, 2011.

[19] S. V. Buuren and K. Groothuis-Oudshoorn, "Multivariate imputation by chained equations," *J. Stat. Soft. [electronic article]*, vol. 45, no. 3, 2011, Accessed December 8, 2023. [Online]. Available: http://www.jstatsoft.org/v45/i03/.

[20] M. M. Esteban y Peña, V. Hernandez Barrera, X. Fernández Cordero, *et al.*, "Self-perception of health status, mental health and quality of life among adults with diabetes residing in a metropolitan area," *Diabetes Metab*, vol. 36, no. 4, 305–311, 2010.

[21] M. Sundaram, J. Kavookjian, J. H. Patrick, *et al.*, "Quality of life, health status and clinical outcomes in type 2 diabetes patients," *Qual Life Res*, vol. 16, no. 2, 165–177, 2007.