Mapping Neighborhood-Level Drivers of Type 2 Diabetes: A Predictive-Causal Approach for Precision Public Health

Mohammad Noaeen
University of Toronto
Amirhosein Rostami
University of Toronto
Ibrahim Ghanem
University of Toronto
Olli Saarela
University of Toronto
Karim Keshavjee
University of Toronto
Jeffrey R. Brook
University of Toronto
Zahra Shakeri
zahra.shakeri@utoronto.ca

University of Toronto

Article

Keywords:

Posted Date: June 6th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-6648511/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Scientific Reports on January 5th, 2026. See the published version at https://doi.org/10.1038/s41598-025-34287-7.

Mapping Neighborhood-Level Drivers of Type 2 Diabetes: A Predictive-Causal Approach for Precision Public Health

Mohammad Noaeen[1,2,+], **Amirhosein Rostami**[2,+], **Ibrahim Ghanem**[3,+], **Olli Saarela**[1], **Karim Keshavjee**[2], **Jeffrey R. Brook**[1,4,5], **and Zahra Shakeri**[1,2,6,7,*]

[1]Dalla Lana School of Public Health, University of Toronto, Toronto, Canada
[2]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Canada
[3]Department of Geography, Geomatics and Environment, University of Toronto, Toronto, Canada
[4]Department of Mineral and Civil Engineering, University of Toronto, Toronto, Canada
[5]Research Institute, The Hospital for Sick Children, Toronto, Canada
[6]Faculty of Information, University of Toronto, Toronto, Canada
[7]Schwartz Reisman Institute, University of Toronto, Toronto, Canada
[*]corresponding author (email: zahra.shakeri@utoronto.ca)
[+]These authors contributed equally to this work and share first authorship

ABSTRACT

Type 2 diabetes has become an urban epidemic influenced by neighbourhood environments. However, conventional risk models focusing solely on individual factors fail to account for these community influences and often require detailed patient data that may not be available. To address this gap, we developed an integrated approach combining machine learning and causal inference to map type 2 diabetes risk at the community level. Using demographic, health, and socioeconomic data from 1,149 Census Tracts (CTs) in a large metropolitan region, we trained seven machine learning models to identify neighbourhoods with high diabetes prevalence. Although neighbourhood-level diabetes data were available for this study area, our model's high predictive accuracy on external validation data (area under the curve (AUC) = 0.95), particularly from a distinct geographical region, demonstrates its potential utility in predicting diabetes risk for other regions in Canada or elsewhere where such data are unavailable. The top models achieved high recall ($> 90\%$) and AUC up to 0.96 on test data, indicating accurate identification of high-risk neighbourhoods with few false positives. Survey-derived community health indicators, including obesity rate, physical inactivity, and median age, were strong predictors of diabetes prevalence. We then applied a Causal Forest approach to estimate the impact (Conditional Average Treatment Effect, τ) of modifiable factors. Higher work stress ($\tau = 0.312$) and daily smoking ($\tau = 0.155$) were moderately associated with increased risk, whereas better mental health ($\tau \approx -1.1$) was protective, highlighting mental health as a critical intervention priority, especially in neighbourhoods predicted to have high diabetes prevalence. These findings illustrate how community-level factors can guide targeted interventions and advance health equity, particularly for immigrant and visible-minority populations. Our integrated machine-learning and causal framework lays the groundwork for precision public health, demonstrating how modifiable neighbourhood factors can indicate diabetes risk when patient-level data are scarce. Furthermore, our methodology is adaptable to other chronic diseases influenced by social and environmental determinants, potentially guiding targeted prevention efforts beyond type 2 diabetes.

Introduction

Type 2 diabetes is a growing global health challenge, affecting approximately 422 million people (5.2% of the world population)[1]. In Canada, over three million individuals (8.9% of the population) have been diagnosed with diabetes, and 6.1% of adults aged 20 to 79 are prediabetic[2]. Prevalence is rising at approximately 3.3% per year[2]. Although traditional perspectives emphasize biological factors, growing evidence shows that social determinants of health (SDoH) strongly shape diabetes incidence and outcomes[3–5]. SDoH covers the conditions in which people live and work, including neighborhood-level resources and stressors. Features such as access to green space, housing stability, and walkability have been linked to better metabolic health outcomes[6–9]. Similarly, socioeconomic disadvantages, such as differences in income, employment, and education, correlate with a higher diabetes burden[10,11]. Regional studies in Ontario and Finland further highlight that these social and environmental factors vary across communities, complicating prevention efforts[12–14]. These complexities suggest the need for analytic methods that capture both the broad influence of environment and the localized nature of risk. Neighborhood-level approaches can directly link environmental factors to disease dynamics in ways that align with clinical priorities and social service needs[15]. Community-level SDoH data can act as a proxy for missing individual information and can capture broader contextual influences on health behaviors, strengthening the case for integrating fine-grained geographic data into predictive models.

Recent advances in machine learning (ML) have opened new possibilities for health predictive analytics[16–18]. Researchers have developed ML models for a range of diabetes-related applications, including prognostic models for diabetes complications[19], ensemble and deep learning approaches for predicting diabetes and prediabetes[20–26], neural networks to identify mental health comorbidities in diabetes patients[27,28], population-level models to explore epidemiological trends[29], and survival analyses to estimate time to disease onset[30,31]. More recently, large language models and generative AI have been explored for diabetes prevention and self-management support[32,33]. However, deploying such AI systems at a neighborhood scale remains challenging due to the need for detailed contextual data. Traditional classification and regression methods often provide clearer interpretability of community-level factors (e.g., local socioeconomic conditions), especially when data are aggregated or partially observed. Indeed, most ML studies on diabetes rely on individual-level clinical data[34], whereas a neighborhood-oriented perspective can reveal spatial patterns that purely clinical datasets might overlook, particularly in urban areas with sharp socioeconomic gradients[35].

Conventional ML is powerful for detecting correlations but cannot tell us if changing a factor will change outcomes[36]. For example, an algorithm might find that neighborhoods with more green space tend to have lower diabetes prevalence, but it cannot tell if adding green space would actually *cause* diabetes rates to drop. Causal machine learning (Causal ML) combines principles from statistical inference and ML to address such counterfactual 'what if' questions, providing insight into potential outcomes of targeted interventions[37–39]. This approach can directly guide public health efforts by indicating how factors such as walkability or income could serve as key prevention targets rather than mere correlations. Although public health agencies collect extensive patient-level diabetes data, these surveillance systems usually focus on confirmed cases and often overlook psychosocial or environmental factors at finer geographic scales. Integrating rich neighborhood-level variables into predictive models enables ML methods to identify high-risk communities and guide localized interventions that account for underlying social and environmental conditions.

In this study, we apply both standard ML and Causal ML methods to predict neighborhood-level type 2 diabetes prevalence in the Greater Toronto Area (the Toronto Census Metropolitan Area (CMA), population: 6.2 million) and its central municipality, the City of Toronto[40]. Toronto's cultural and ethnic diversity makes it an ideal testbed for examining how multiple determinants interact across communities. Our integrated predictive–causal approach is designed to support precision public health (i.e., using data to deliver the right interventions to the right populations) by finding 'hotspots' of elevated risk and identifying plausible intervention targets. The insights from this work can help clinicians, policymakers, and urban planners allocate resources and design programs that explicitly account for neighborhood socioeconomic and environmental conditions. This framework directly links predictive modeling results to real-world impact, illustrating how SDoH profiles vary between communities and how data-driven models can guide targeted prevention and improved care.

Results

We evaluated seven machine learning algorithms: Logistic Regression (LR), K-Nearest Neighbors (KNN), Neural Networks (NN), Support Vector Machines (SVM), Decision Trees (DT), Random Forests (RF), and Extreme Gradient Boosting (XGBoost). Table 1 compares the performance of these algorithms for detecting high-diabetes neighborhoods in the Toronto CMA and the City of Toronto, evaluated by accuracy, precision, recall, and F1-score. The SVM and NN models were identified as the top-performing algorithms. The SVM correctly identified all high-prevalence neighborhoods ($y_i = 1$), achieving 100% recall on the test set. The NN followed closely, with 95% recall on the test set and 78% on the external validation set.

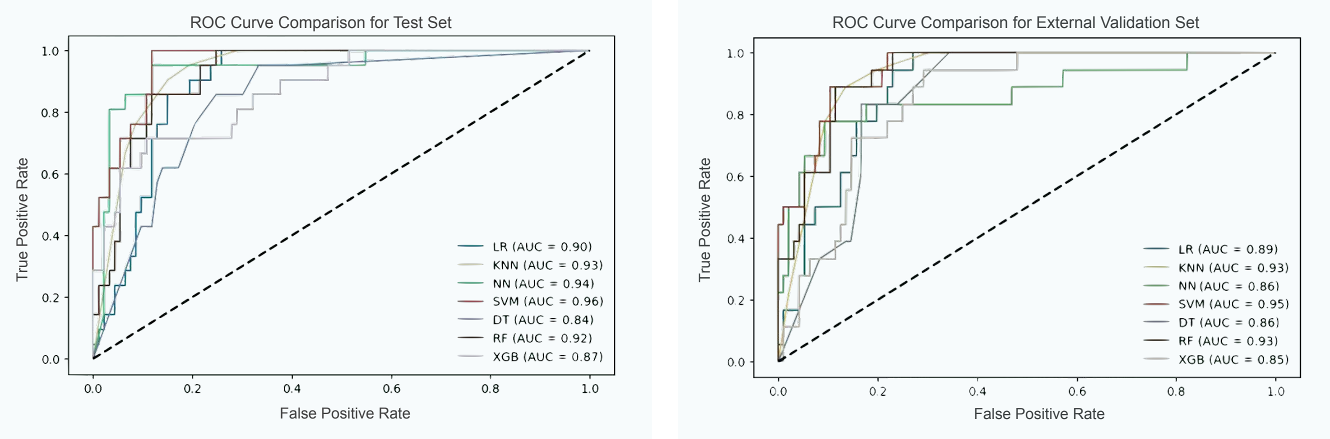
Figure 1 shows that, for the Toronto CMA, the SVM achieved the highest AUC at 0.96 on the test data and 0.95 on external validation, while the NN reached an AUC of 0.94 on the test set and 0.86 on external data. The external validation set for the Toronto CMA and the city of Toronto respectively consist of geographically distinct subsets: Brampton in Peel Region (an area with higher-than-average diabetes prevalence) for the Toronto CMA model and Old Toronto for the City-of-Toronto model. Both subsets include high- and low-prevalence neighbourhoods to ensure a thorough evaluation.

Obesity prevalence, overweight rates, and physical inactivity, as indicated by low active transportation usage, were especially influential in distinguishing neighborhoods with higher diabetes prevalence, aligning with known links between sedentary lifestyles, excess weight, and metabolic risk[4,6]. The SVM and NN models appear to capitalize on non-linear relationships among these features by using richer representations. For instance, the SVM with a radial basis function kernel can capture complex decision boundaries influenced by socioeconomic and environmental factors, while the NN's multi-layer architecture learns interactions among variables such as education level, walkability, and obesity. Other algorithms in our set of models (e.g., DT, RF, and XGBoost) can also model non-linearities, but the SVM and NN achieved the highest accuracy and recall on this dataset. For instance, the XGBoost model, despite its popularity, did not outperform the SVM or NN in our study. One

	Toronto CMA								City of Toronto							
Model	Accuracy		Precision		Recall		F1		Accuracy		Precision		Recall		F1	
	T	E	T	E	T	E	T	E	T	E	T	E	T	E	T	E
LR	0.82	0.82	0.51	0.44	0.90	0.67	0.66	0.53	0.81	0.80	0.40	0.40	1.00	0.73	0.54	0.51
KNN	0.86	0.87	0.58	0.55	0.90	0.89	0.70	0.67	0.83	0.84	0.41	0.43	0.93	0.91	0.56	0.59
NN	0.87	0.89	0.59	0.61	0.95	0.78	0.73	0.68	0.82	0.84	0.45	0.45	1.00	0.90	0.62	0.61
SVM	0.86	0.85	0.57	0.52	1.00	0.94	0.72	0.67	0.83	0.77	0.45	0.41	1.00	0.91	0.62	0.62
DT	0.79	0.83	0.45	0.48	0.62	0.83	0.52	0.61	0.84	0.84	0.38	0.45	0.74	0.85	0.50	0.56
RF	0.86	0.82	0.58	0.47	0.86	0.89	0.69	0.62	0.79	0.74	0.37	0.37	0.85	0.86	0.51	0.51
XGBoost	0.75	0.75	0.41	0.38	0.71	0.83	0.52	0.54	0.74	0.78	0.39	0.39	0.63	0.78	0.51	0.51

Table 1. Summary of performance metrics for each model, including accuracy, precision, recall, and F1 score, on test (T) and external validation (E) datasets for the Toronto CMA and the City of Toronto.

possible reason is the relatively limited size of our dataset. Complex ensemble methods like XGBoost can exhibit high variance and may overfit when data are scarce or noisy, which can reduce generalization performance. In such data-constrained settings, the advantages of boosting algorithms are less pronounced, and a well-regularized classifier like the SVM can end up achieving better results.



(a) Test Dataset

(b) External Validation

Figure 1. Receiver Operating Characteristic (ROC) Curve Comparison for Diabetes Prevalence Prediction. Each curve shows the true positive rate (TPR) and false positive rate (FPR) across different thresholds. Higher TPRs and lower FPRs shift the curve further away from the diagonal line, which represents random guessing (TPR = FPR). The SVM model outperforms other methods, achieving the highest AUC of 0.96 on the test set and 0.95 on the external validation set.

The SVM's performance improved markedly after targeted feature engineering. After we log-transformed the skewed median-age variable (hereafter referred to as log median age) and removed highly collinear features, the SVM's recall on the test set rose from 88% to 100% with no loss of precision. This outcome suggests that normalizing feature distributions and mitigating multicollinearity helped the SVM learn a more effective decision boundary, enabling it to correctly identify all high-prevalence neighborhoods (achieving 100% sensitivity) without increasing false positives. A subsequent Recursive Feature Elimination with Cross-Validation (RFECV) procedure identified five leading predictors that were consistently retained by the SVM, NN, RF, KNN, and LR models: obesity rate, overweight rate, active transportation rate, rate of physically active population, and log-transformed median age. These factors are closely tied to local lifestyle and demographic conditions, highlighting the value of combining individual health-behavior indicators with environmental context when modeling diabetes risk.

Geographical Applicability and Model Scalability

The Toronto CMA encompasses a large and diverse region, whereas the City of Toronto represents the dense urban core with somewhat more homogeneous demographics and narrower socioeconomic variability. We observed that applying the SVM model (trained on the CMA) to data from the City of Toronto led to a drop in precision (from 57% to 45% on the test set, and from 52% to 41% on external validation), while recall remained 100% on the test set and dipped only slightly (from

94% to 91%) on external validation (Table 1). One likely explanation is that the City of Toronto has less variability in certain predictors, which caused the classifier to flag a larger share of neighborhoods as high prevalence (i.e., increasing false positives). In particular, in the City of Toronto, data on 20% of the neighborhoods are a truly high prevalence, so a precision of 45% still represents more than double the base rate. From a public health perspective, we therefore prioritized recall (i.e., sensitivity) over precision because reliably identifying every high-prevalence neighborhood, even at the expense of some false alarms, is preferable to missing communities that need intervention.

Causal Analysis of Diabetes Risk Factors

We conducted a causal analysis to examine how modifying certain neighborhood factors might influence diabetes prevalence. Three sets of features were defined to explore different dynamics: (1) a *base* set consisting of the SVM's top predictive features (i.e., log median age, obesity rate, overweight rate, active transportation rate, and active population rate), (2) an *extended* set that added key psychosocial factors to the base (i.e., mental health score, visible minority rate, work stress score, and daily smokers rate), and (3) a *full* feature set incorporating a broader range of attributes (Table 2). The extended set was designed to introduce modifiable social and behavioral variables (e.g., stress and smoking) while keeping model complexity in check.

Feature Set	**Treatment**	$\tau(x)$
SVM set (Log median age, Obese rate, Overweight rate, Active transportation rate, Rate of active population)	Active transportation rate	0.003
	Rate of active population	-0.011
Extended set (SVM features + Mental health score, Visible minority rate, Work stress score, Daily smokers rate)	Active transportation rate	0.004
	Rate of active population	-0.001
	Mental health score	**-1.113**
	Work stress score	-0.021
	Daily smokers rate	**0.155**
Full set (Log median age, Obese rate, Overweight rate, Rented rate, Residential instability index, Material deprivation, Ethnic concentration index, Food insecurity score, Visible minority rate, Work stress score, Mental health score, Regular alcohol drinkers rate, Daily smokers rate, Rate of active population, Active transportation rate, High education rate, Average income, Unemployment rate, Recent immigration rate)	Food insecurity score	-0.039
	Work stress score	**0.312**
	Mental health score	**-1.045**
	Regular alcohol drinkers rate	-0.005
	Rate of daily smokers	0.003
	Rate of active population	-0.017
	Active transportation rate	-0.014
	High education rate	0.001
	Average income	0.000
	Unemployment rate	-0.001

Table 2. Summary of Causal Forest results for each feature set (SVM base, extended, and full), reporting the estimated $\tau(x)$ for selected treatments in neighborhood-level diabetes prevalence.

We used Causal Forest to estimate the Conditional Average Treatment Effect (CATE), denoted by $\tau(x)$, for selected features and to explore the relationships between predictors and diabetes outcomes. Some candidate variables were not treated as direct 'interventions' in this analysis. For example, we did not assign treatments for log median age, obesity, overweight, visible minority rate, or recent immigrant rate, because these factors are difficult to manipulate in isolation or are inherently tied to other variables. Reducing obesity at a population level, for instance, requires multifaceted programmes involving policy changes, dietary interventions, and increased physical activity, and recent immigrant status is closely linked to income and education. We therefore focused on feasible intervention-related features.

For the base feature set, the Causal Forest estimated almost no direct effect from increasing active transportation ($\tau(x) = 0.003$) or the active population rate ($\tau(x) = -0.011$). This outcome does not mean that promoting physical activity has no value, only that within our dataset and model these variables did not show a strong independent causal influence on diabetes prevalence. With the extended feature set, which added mental health, visible minority status, work stress, and smoking, the analysis revealed a meaningful protective effect of better mental health, $\tau(x) = -1.113$, and a positive moderate effect of daily smoking, $\tau(x) = 0.155$, consistent with the known metabolic risks of smoking.

In the model incorporating the full feature set, work stress showed a moderate positive estimated effect on diabetes prevalence, $\tau(x) = 0.312$, whereas mental health remained strongly protective, $\tau(x) = -1.045$. Food insecurity displayed a slight negative coefficient, $\tau(x) = -0.039$, which may reflect a complex mix of household factors or simply noise given the small magnitude. Alcohol use, unemployment rate, and recent immigration produced essentially negligible effects in our Causal Forest results, suggesting that other psychosocial and environmental features play larger roles in determining neighbourhood-level diabetes vulnerability.

Discussion

Neighborhood-level data provided a clear advantage in identifying communities with high diabetes prevalence, beyond what individual-level or highly aggregated data alone can offer. When combined with physical activity indicators, factors such as obesity, overweight rates, and age structure were identified as important predictors of local risk, showing how the built environment and lifestyle factors shape disease patterns[41,42]. These findings are consistent with previous research linking excess body weight and sedentary behavior to cardiometabolic conditions in urban settings[4]. However, many health agencies lack standardized neighborhood-level data on outcomes and risk factors, making it difficult to monitor socioeconomic disparities in diverse populations[43]. Systematically incorporating local characteristics could refine resource allocation by identifying areas where interventions are most urgently needed[44].

In this study, the positive class (i.e., $y_i = 1$: high-prevalence neighborhoods) accounted for roughly 20% of the sample, producing a class imbalance that can bias ML models toward the majority class. To address this, we applied Random Under-Sampling (RUS) to the training data: we retained all minority-class examples and down-sampled the majority class to achieve a desired ratio. Let N_{maj} and N_{min} denote the numbers of majority and minority observations, respectively. We drew a subset $S \subseteq N_{\text{maj}}$ with $|S| = r N_{\text{min}}$ for a chosen ratio r, then trained the model on $S \cup N_{\text{min}}$. This strategy preserved authentic high-prevalence neighborhoods rather than synthetically generating new ones. We chose RUS over oversampling methods, such as the synthetic minority oversampling technique (SMOTE)[45], which creates synthetic data points that could distort complex socioeconomic feature distributions. For example, SMOTE interpolates between minority instances as

$$x_{\text{new}} = x_i + \lambda \left(x_j - x_i\right), \quad \lambda \sim U(0,1),$$

where λ is a random interpolation weight drawn from the continuous uniform distribution $U(0,1)$ that determines how far along the segment from x_i to x_j the synthetic point lies. Although SMOTE-Nominal Continuous (SMOTE-NC) adapts this approach for mixed data, synthetic points may misrepresent certain categorical or survey-derived characteristics essential to our local context. RUS preserves authentic minority samples, aligning with our goal of maximizing recall for higher-risk neighborhoods[46].

We also explored probability calibration with methods such as Platt scaling or isotonic regression. Platt scaling fits a logistic function,

$$P(y = 1 \mid s) = \frac{1}{1 + \exp(As + B)},$$

where s is the uncalibrated score, and (A, B) are fitted parameters. Isotonic regression learns a non-decreasing piecewise-constant function to map raw scores to probabilities. While these calibration techniques can improve the interpretability of probability outputs, they add model complexity and require additional data to fit reliably. In our case, with a relatively limited dataset, we prioritized identifying high-prevalence neighborhoods (i.e., recall and F1) over perfectly calibrated probabilities. Our primary objective was classification performance, flagging as many true high-risk neighborhoods as possible, so we elected not to apply post-hoc calibration, as it was not essential for the intervention-focused goals of this study.

Careful curation of neighbourhood variables improved model performance and identified policy-relevant levers such as material deprivation and residential instability, which can guide infrastructure upgrades and resource allocation for diabetes prevention. This same feature-driven framework can be adapted to other chronic diseases influenced by environment and behaviour, including hypertension.

Assessing Model Reliability and Sensitivity

To evaluate model reliability, we measured the sensitivity of the SVM to each predictor through permutation tests[47]. For a given performance metric $M(f)$, the values of feature x_j were randomly permuted across the dataset, breaking its link to the outcome, and the resulting drop in performance was recorded. Feature importance was then defined as

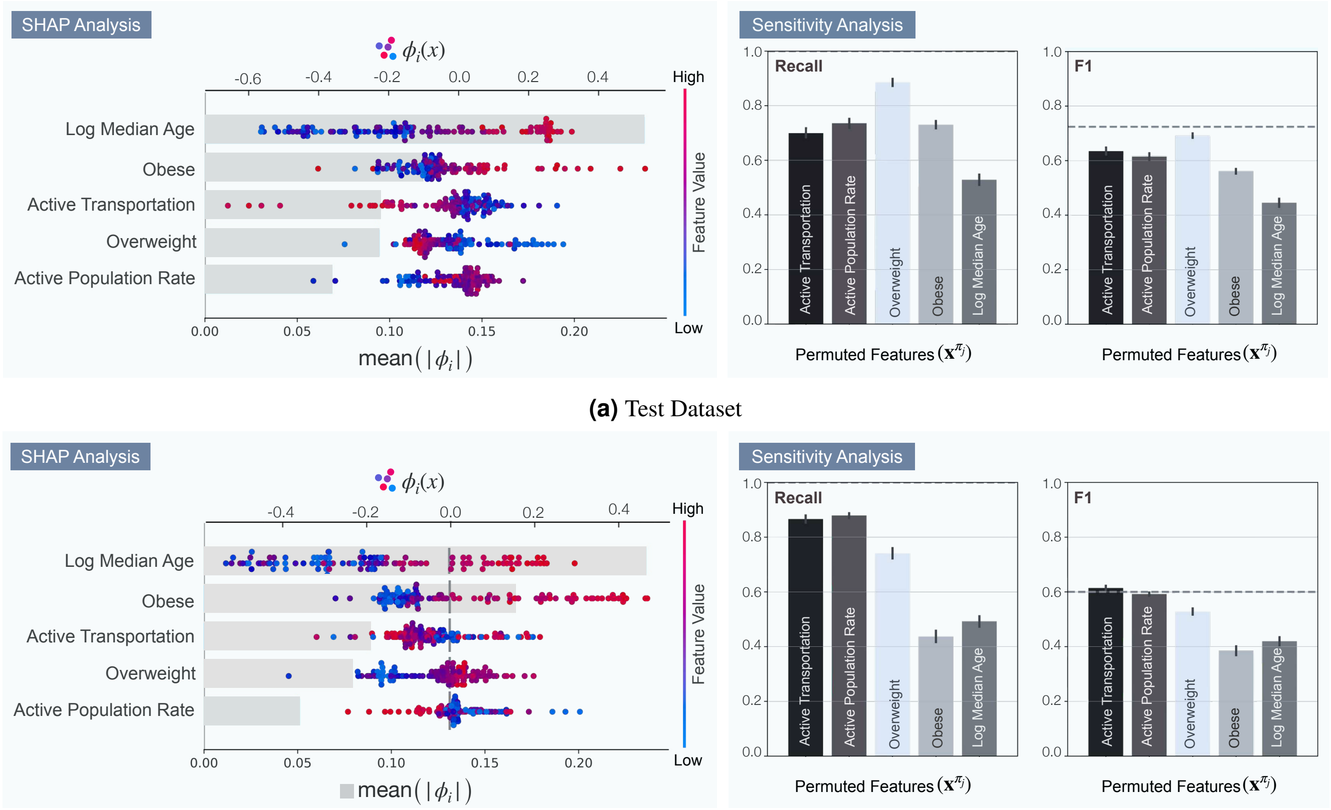
$$S(x_j) = M\big(f(\mathbf{x})\big) - M\big(f(\mathbf{x}^{\pi_j})\big),$$

where $M\big(f(\mathbf{x})\big)$ is the original score and $f(\mathbf{x}^{\pi_j})$ is the model after permuting feature j. A larger $S(x_j)$ therefore indicates a stronger contribution of that feature[48]. Figures 2a and 2b show that permuting `Log Median Age` produced the greatest reduction in recall and F1, making it the most influential variable for identifying high-prevalence neighbourhoods. `Obesity`, `Overweight`, and `Physical Activity` followed in importance, a pattern consistent with evidence that active living reduces metabolic risk in urban settings[49].

We complemented this global analysis with SHapley Additive exPlanations (SHAP), which allocate the difference between an individual prediction and a baseline prediction to each feature i[50]:

$$\phi_i(x) = \mathbb{E}_{S \subseteq \mathscr{F} \setminus \{i\}} \Big[f(S \cup \{i\}, x) - f(S, x) \Big].$$

The beeswarm plots in Figure 2a (i.e., internal test set) and Figure 2b (i.e., external validation set) display these local contributions. Neighbourhoods with higher Log Median Age or Obesity shift SHAP values to the right, pushing predictions toward the 'high prevalence' class, whereas greater use of active transportation shifts points to the left and signals lower risk. The accompanying bar charts confirm that the features with the largest SHAP magnitudes also produce the steepest declines in recall and F1 when permuted. The agreement between these two methods suggests that interventions targeting age structure, obesity, and physical-activity levels could meaningfully benefit communities at elevated risk, in line with prior work[51,52].



(a) Test Dataset

(b) External Validation

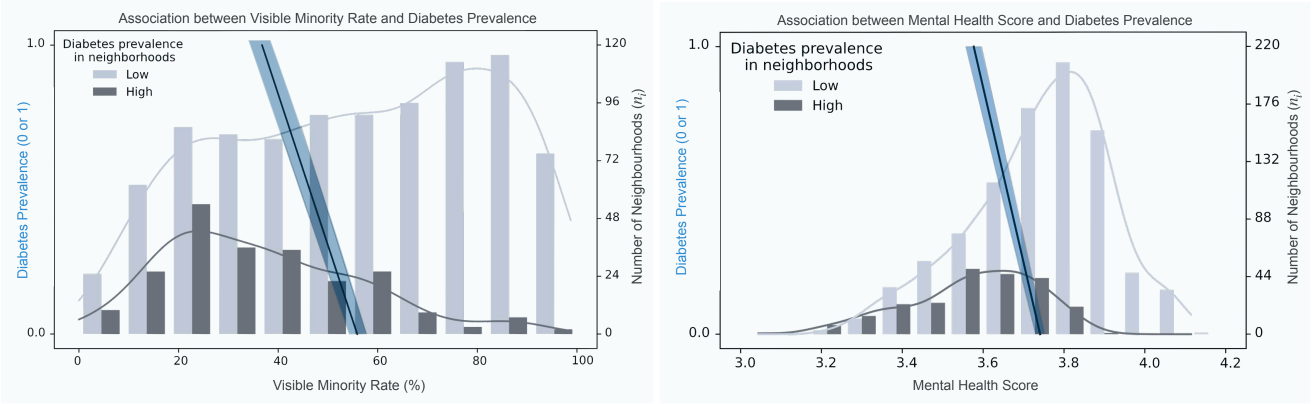
Figure 2. Integrated SHAP and sensitivity analysis of SVM interpretability and robustness. (a) Test dataset. The upper beeswarm displays local Shapley values $\phi_i(x)$ for every feature, coloured from pink to blue to denote low and high feature values. Points to the right of zero increase the predicted probability of high prevalence, whereas points to the left decrease it. The lower Barnhart bar chart reports global importance through mean absolute Shapley values $\mathbb{E}[|\phi_i|]$. The adjacent bars show the drop in Recall and F1 when feature j is permuted, $\mathbf{x}^{\pi_j}$, thereby measuring sensitivity. (b) External-validation dataset. The same structure confirms that Log Median Age, Obesity, Overweight, Active Population Rate, and Active Transportation remain the dominant predictors, and that the permutation analysis yields comparable robustness across data sources.

Urban Lifestyle Interventions. Our results indicate that urban design elements that facilitate physical activity can influence neighbourhood-level diabetes risk. Prior work shows that the built environment shapes daily movement patterns and metabolic health[53]. Although the conditional average treatment effect for active transportation, $\tau(x)$, was close to zero in our analysis, this finding reflects limitations of the current dataset rather than the irrelevance of the intervention. In practice, redesigning neighbourhoods, for example by adding bike lanes or continuous sidewalks, can promote more active routines and may lower diabetes prevalence in high-risk areas. Well-planned infrastructure is widely recognized as a catalyst for healthier behaviours and can, over time, reduce the metabolic burden on urban communities.

Evidence from other cities supports this interpretation. Neighborhoods designed with pedestrians and cyclists in mind tend to have lower diabetes prevalence[9,54]. In Toronto, for instance, debates over removing bike lanes have raised concerns about negative impacts on residents' daily activity and well-being[55]. Likewise, experiences in cities like Copenhagen and

Melbourne show that urban designs promoting active transportation are associated with improved metabolic health indicators[56]. Importantly, urban lifestyle interventions carry co-benefits beyond diabetes prevention: features that support walking and cycling also contribute to cleaner air and stronger social cohesion in communities[57]. Policymakers can use these insights to align urban planning with public health goals, clarifying how specific design choices in the built environment may reinforce healthier lifestyles.

Ethnocultural Influences on Diabetes. We used an identical set of five features for the Toronto CMA and the City of Toronto ML models to ensure comparability on the two geographic scales. To achieve this, we removed two variables, mental health score and work stress score, from the City of Toronto model due to multicollinearity. To explore the impact of these psychosocial factors on ML prediction, we conducted an extended analysis on the Toronto CMA data that explicitly included mental health score, work stress score, log median age, obesity, overweight, active transportation, and active population rate (seven features). Adding the mental health and work stress variables led to slight performance changes: approximately a 2% gain in accuracy, a 6% gain in precision, and a 1% gain in F1, accompanied by a 3% drop in recall. This outcome aligns with earlier observations that expanding a model's feature set can boost accuracy and precision at the cost of a small decrease in sensitivity[58,59]. The trade-off arises because additional predictors introduce variability, making the model more confident (and precise) in classifying many positive cases, but causing it to miss a few borderline cases it previously captured. Nonetheless, incorporating mental health and work stress measures enhanced the model's ability to capture local sociocultural factors. In a diverse region like the Toronto CMA, these variables noticeably affect diabetes risk, thus their inclusion helps the model better reflect real-world community differences.



(a) Visible minority rate

(b) Mental health score

Figure 3. Visible Minority Rate and Mental Health Score in relation to diabetes prevalence. (a) shows a grouped bar chart of binned visible-minority rates in neighbourhoods classified by diabetes prevalence (low: $y_i = 0$, high: $y_i = 1$). Bars (right y-axis) indicate the number of neighbourhoods n_i in each bin, and smoothed density curves overlay the bars to highlight where each class is concentrated. The fitted line visualizes the mean visible minority rate for each diabetes class, with diabetes prevalence class on the y-axis and the visible minority rate on the x-axis. For each diabetes class, the mean visible minority rate is calculated across the set of census tracts (CTs) labeled with diabetes class i as $\hat{x}_i = \frac{1}{n_i}\sum_{\ell\in\mathscr{C}_i} x_\ell$, where $\mathscr{C}_i$ denotes the set of CTs with diabetes class i, $n_i = |\mathscr{C}_i|$ is the number of such CTs, and x_ℓ is the visible minority rate of CT ℓ. (b) applies the same approach to mental health scores, where x_ℓ denotes the mental health score of tract ℓ.

One notable insight from our analysis is the association between neighborhood ethnic composition and diabetes. The *visible minority rate* was a key predictor in our models. Furthermore, the line graph in Figure 3a indicates an inverse relationship: neighborhoods with higher proportions of visible minority residents tended to have lower average diabetes prevalence. This pattern is consistent with some earlier findings on how neighborhood racial/ethnic composition and segregation relate to diabetes disparities[60,61]. It may suggest that communities with many recent immigrants or minority residents maintain certain protective cultural practices (e.g., traditional diets or strong social networks) that help mitigate diabetes risk[62]. The well-documented 'healthy immigrant effect' may also contribute to this pattern. Recent immigrants often arrive in relatively good health because of selection mechanisms, and this advantage can translate into lower chronic disease rates in the short term[63]. In Toronto, where more than 50 % of residents are immigrants[64], these findings show how demographic context can shape the local diabetes burden.

A similar protective gradient emerged for perceived mental health. We observed a clear inverse association between neighbourhood mental health status and diabetes burden. Neighbourhoods with *higher* mean mental health scores, indicating better perceived well-being, exhibited *lower* mean diabetes prevalence, as illustrated in Figure 3b. This pattern is consistent with extensive epidemiological evidence showing that psychological distress and depression elevate the risk of incident type 2 diabetes and worsen glycemic control and insulin sensitivity, contributing to diabetes development[65–69]. At the physiological level, chronic psychosocial stress activates the hypothalamic–pituitary–adrenal axis, elevates cortisol, and promotes systemic insulin resistance, linking poor mental health to metabolic dysfunction[70]. Conversely, the demands of diabetes self-management can negatively affect mental well-being, creating a bidirectional feedback loop between the two conditions[29].

It is important to acknowledge that the Canadian Community Health Survey (CCHS) item is a single, self-rated indicator. Although validated against structured diagnostic interviews[71], self-report instruments can vary in meaning across cultural groups and may underrepresent individuals experiencing severe psychological distress, especially when survey non-response is high[72,73]. Such biases typically weaken rather than exaggerate associations, so the true mental-health gradient in diabetes prevalence may be even steeper than observed. Our supplementary analyses (Supplementary Figures 1 and 2) further illustrate how mental health and work stress influenced the Toronto CMA models under the extended and full feature sets. Taken together, these results suggest that strategies addressing both psychological well-being and metabolic health could yield better outcomes than focusing on either in isolation.

Causal Forest Insights on Diabetes Risk

Our Causal Forest analysis clarified which risk factors have the strongest causal influence on neighbourhood-level diabetes prevalence. `Work stress` stood out as the most detrimental factor: higher stress was linked to a moderate increase in diabetes prevalence, $\tau(x) = 0.312$ in the full model, consistent with evidence that psychosocial stress impairs metabolic function[74]. `Daily smoking` showed a positive causal effect as well, with $\tau(x) = 0.155$, aligning with literature that connects smoking to inflammation and insulin resistance. In contrast, `mental health` acted as a protective factor. We estimated $\tau(x) \approx -1.1$ for the mental health score across both the extended and full feature sets, implying that better average well-being in a neighbourhood could noticeably lower its diabetes prevalence.

These causal findings are supported by our model interpretability analysis. Although active transportation did not exhibit a notable causal effect ($\tau(x) \approx 0$) in the Causal Forest results, it remained an important predictor in the ML models' SHAP rankings. This discrepancy suggests indirect or interaction pathways. For example, neighbourhoods that promote walking may also provide other health-supporting amenities. Age structure, captured by `log median age`, consistently ranked among the top predictors and was associated with higher diabetes risk, a pattern expected for an age-related disease.

Obesity repeatedly appeared as a key driver in the SHAP summaries. Our combined results further suggest an interaction between work stress and weight status. The adverse effect of stress was stronger in neighbourhoods with high rates of overweight residents than in those with high obesity rates. A possible explanation is that individuals who are overweight but not yet obese may be more susceptible to stress-induced metabolic changes, with stress acting as a tipping point[75]. Such context-specific interactions would be missed in a single-factor analysis.

Together, causal inference and SHAP analysis provide a more comprehensive view of the pathways that link psychosocial and lifestyle factors to diabetes risk[76]. Supplementary Figures 3 and 4 present additional results from the base, extended, and full feature sets, and the combined evidence clarifies the roles of psychosocial and demographic attributes, such as mental health, stress management, and age structure, in shaping community-level diabetes burden.

Limitations

Although we included a broad range of socioeconomic and behavioural variables, several environmental factors known to influence health were not directly measured. For instance, the dataset lacked explicit information on green-space availability and air-pollution levels[77]. We relied on proxies such as active-transportation metrics and material-deprivation indices to approximate these exposures, but direct measurements would likely improve model performance. The Causal Forest residual term, $\varepsilon = y - \hat{y}$, may therefore capture unobserved environmental influences or latent confounders. Working at the census-tract scale can also obscure finer social and mobility patterns that shape health outcomes, a limitation commonly referred to as the *modifiable areal unit problem* (MAUP)[78].

Many predictors originated from self-reported survey data, introducing potential biases such as recall error and variation in how questions are interpreted across cultural groups[79]. We cross-validated with additional data sources whenever possible. However, more objective clinical measures such as biometric indicators would enhance model fidelity. The cross-sectional design also limits causal interpretation. Longitudinal or quasi-experimental approaches, such as natural experiments, would provide clearer evidence of how changes in social and environmental conditions influence diabetes prevalence over time[80–83]. Longitudinal data or quasi-experimental designs would offer a clearer view of how changes in social or environmental conditions influence diabetes rates over time.

Moreover, our external validation was limited to the Toronto CMA and the City of Toronto, which may constrain generalizability[84]. Applying this framework to other urban regions and to rural settings would strengthen confidence in its broader applicability. Data limitations also prevented inclusion of certain relevant factors, such as neighbourhood health care access or culturally specific dietary behaviours, that can influence diabetes risk and model completeness. In addition, some

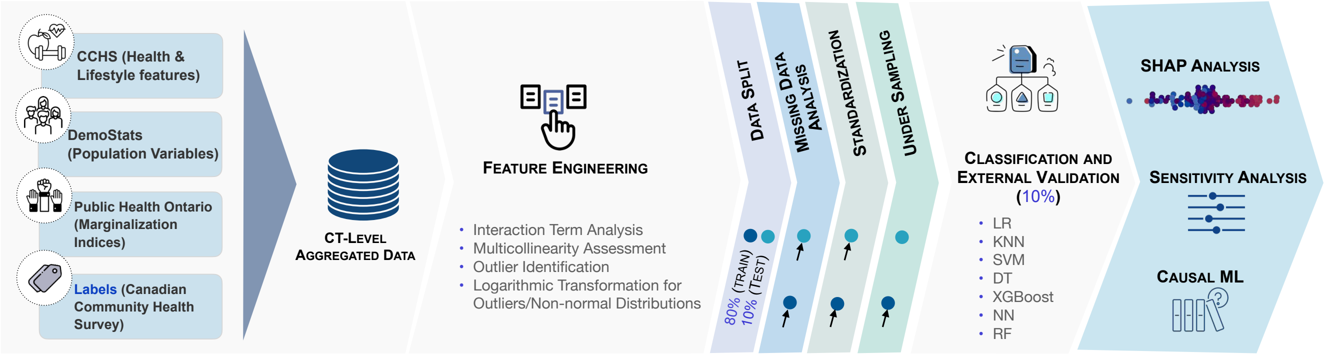

CCHS (Health & Lifestyle features)
DemoStats (Population Variables)
Public Health Ontario (Marginalization Indices)
Labels (Canadian Community Health Survey)
CT-Level Aggregated Data
Feature Engineering
Interaction Term Analysis
Multicollinearity Assessment
Outlier Identification
Logarithmic Transformation for Outliers/Non-normal Distributions
Data Split
80% (Train)
10% (Test)
Missing Data Analysis
Standardization
Under Sampling
Classification and External Validation (10%)
LR
KNN
SVM
DT
XGBoost
NN
RF
SHAP Analysis
Sensitivity Analysis
Causal ML

Figure 4. Overview of the employed ML and Causal ML pipeline. The workflow combines data from multiple sources, addresses interactions and class imbalance, then trains seven ML algorithms (LR, KNN, SVM, DT, XGBoost, NN, RF). SHAP, sensitivity analysis, and Causal ML clarify model outputs and suggest potential neighborhood interventions.

population subgroups may be underrepresented in the data, potentially introducing prediction bias[85]. Future work should incorporate richer information on health status, income and housing, in line with recommendations for comprehensive and equitable ML practice in health care[86].

Despite these limitations, our study demonstrates the unique value of integrating ML prediction with causal inference in public health analytics. This combined approach offers a novel blueprint for precision public health[87]: it clarifies how modifiable neighborhood factors drive disease risk and helps ensure that the right interventions reach the right communities. The same methodology could be adapted to other chronic diseases influenced by social and environmental determinants, potentially guiding targeted prevention efforts well beyond type 2 diabetes.

Methods

Figure 4 provides an overview of our machine learning and causal inference pipeline for identifying neighborhoods with high type 2 diabetes prevalence in the Toronto CMA. The process encompasses data collection, feature engineering, predictive modeling with seven ML algorithms, and interpretative analyses supported by sensitivity checks, SHAP analysis, and causal ML to explore counterfactual outcomes. We formalize our dataset as $\mathscr{D} = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, where each neighborhood i is represented by a feature vector $x_i \in \mathbb{R}^m$ (with m features) and a binary label y_i indicating whether the neighborhood falls in the top quintile of diabetes prevalence (10.5% to 23.8%). We train a classification model M to predict high-prevalence status, such that $\hat{y}_i = M(x_i)$ should correctly flag neighborhoods at elevated risk.

Dataset Overview and Feature Engineering

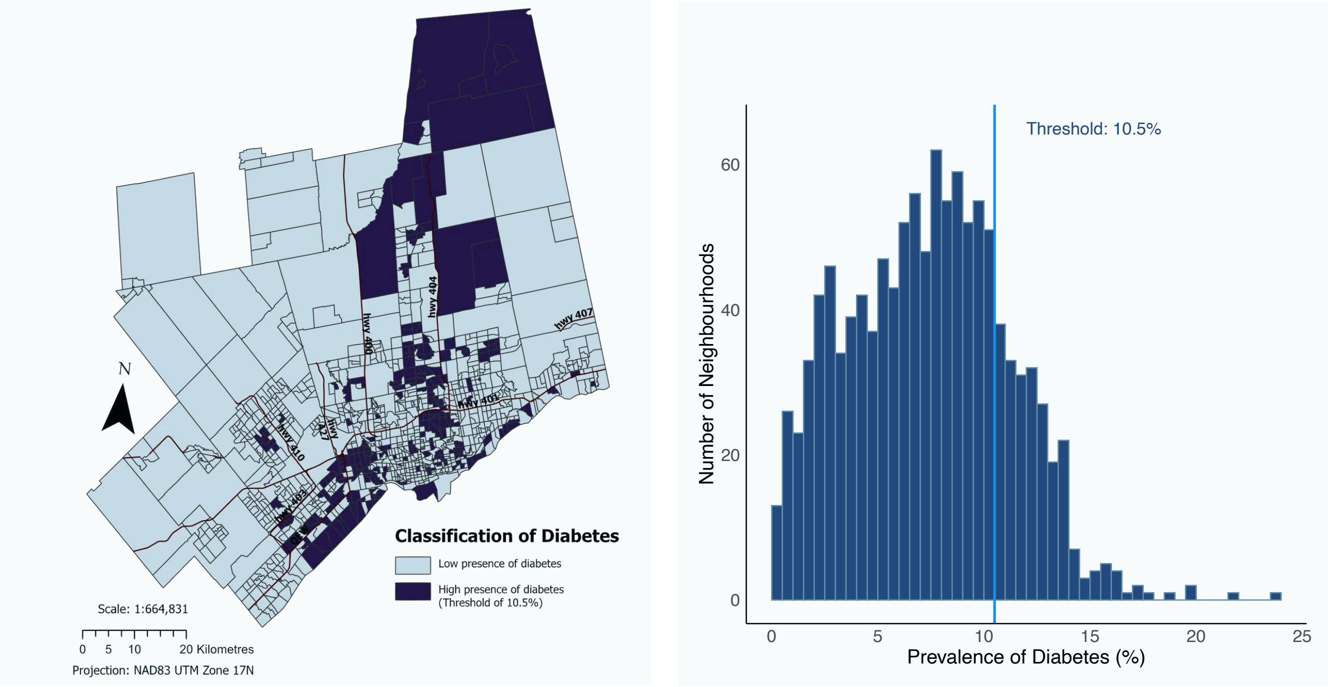
Dataset Overview. Census tracts (CTs) defined by Statistics Canada serve as proxies for neighborhoods with relatively homogeneous socioeconomic and living conditions[88]. We selected predictor variables based on prior studies linking health, socioeconomic, lifestyle, and marginalization indicators to diabetes outcomes[4,6,8,9,11]. Neighbourhood-level diabetes prevalence data came from the 2022 Canadian Community Health Survey (CCHS)[89], which samples Canadians aged 12 and older across 121 health regions and covers about 98% of the population. The remaining 2% live in remote locations such as reserves or Crown Lands and are excluded because of logistical constraints. From the CCHS we identified 1,149 CTs in the Toronto CMA, of which 230 (20.0%) met the definition of high prevalence. We classified a tract as 'high prevalence' if its diabetes rate exceeded the 80th percentile threshold of 10.5%. Figure 5a shows the spatial pattern of these tracts, and Figure 5b presents the full distribution of prevalence values up to the maximum of 23.8%. This binary categorization, contrasting tracts in the top quintile with all others, allowed us to concentrate on areas at distinctly elevated risk.

Data Exploration and Feature Engineering. Neighbourhood predictors were drawn from three main sources: (i) health and lifestyle measures from the 2022 CCHS, (ii) demographic and socioeconomic variables from Environics Analytics' DemoStats dataset[90], and (iii) marginalization indices from the 2021 Ontario Marginalization Index (ON-Marg) compiled by the MAP Centre for Urban Health Solutions[91]. DemoStats provides more than 750 indicators, derived from census counts, economic data, and immigration information, at the CT level[92]. ON-Marg supplies dimension-specific indices of neighbourhood marginalization, such as economic disadvantage and residential instability, together with quintile groupings[93]. We accessed these sources through the SimplyAnalytics platform[94] and Public Health Ontario[95], which yielded an initial pool of 26 candidate features summarized in Table 3.

Table 3. Overview of the 26 study variables sourced from the Canadian Community Health Survey (CCHS), DemoStats, and Public Health Ontario, representing population health, demographics, and socio-economic status.

Feature	Description
Canadian Community Health Survey	
Mental health score	Perceived mental health: Higher = better
Mental health binary *	0 = poor, 1 = good (avg score cutoff)
Regular alcohol drinkers rate	% regular alcohol drinkers
Daily smokers rate	% daily smokers
Food insecurity score	Food insecurity: Higher = more
Food insecurity binary*	0 = low, 1 = high (avg score cutoff)
Work stress score	Perceived work stress: Higher = more
Work stress binary*	0 = low, 1 = high (avg score cutoff)
Rate of active population	% physically active (18+)
Active transportation rate	% using active transport (18+)
Obese rate	% obese population
Overweight rate	% overweight population
DemoStats	
High education rate	% with diploma/degree (25-64)
Average income	Avg household income
Unemployment rate	% unemployed (15+)
Median age	Median population age
Recent immigrant rate	% immigrants (2017-2022)
Visible minority rate	% visible minorities
Visible minority binary*	0 = <25%, 1 = >25%
Rented rate	% rented dwellings
Public Health Ontario	
Ethnic concentration index	% recent immigrants & visible minorities
Ethnic concentration quintiles	Ethnic concentration quintiles
Residential instability index	Residential density & family structure
Residential instability quintiles	Residential instability quintiles
Material deprivation index	Access to basic needs
Material deprivation quintiles	Material deprivation quintiles

* Feature calculated by the authors using the raw feature



(a) Spatial Distribution

(b) Diabetes Rate

Figure 5. Geographic Distribution and Prevalence Rates of Type 2 Diabetes in the Toronto CMA. (a) Map of neighborhoods classified as high-prevalence (10.5%–23.8%) or low-prevalence (lower than 10.5%). (b) Histogram showing the distribution of Type 2 diabetes prevalence rates across 1,149 neighborhoods in the region.

We applied standard preprocessing steps to prepare the features. Categorical variables were one-hot encoded. We created interaction terms to capture higher-order socioeconomic patterns. For instance, we combined the material deprivation and residential instability quintiles into a single 'instability deprivation' indicator that represents neighbourhoods facing both economic and housing challenges[91]. We applied log transformations to skewed variables (e.g., median age) to reduce skewness and checked for multicollinearity, excluding any feature with a Pearson correlation $|r| \geq 0.7$ with another feature.

The dataset was split into a training set (80% of neighborhoods), a test set (10%), and an external validation set (10%). Feature scaling (i.e., normalization) was fitted on the training data and then applied to the test and validation data. Given the 20% prevalence of the positive class, we balanced the training set using RUS, reducing 919 training examples to 340 while leaving the class distribution unchanged in the test and validation sets. For external validation, we chose two geographically distinct areas: Brampton in Peel Region for the Toronto-CMA model and Old Toronto (along Lake Ontario) for the City-of-Toronto model. Peel Region, which includes Brampton, reports a higher diabetes prevalence than the Ontario average[96], making it a stringent test case. Each external subset contained both high- and low-prevalence neighbourhoods to ensure a thorough evaluation. We used five-fold cross-validation during training for hyperparameter tuning and feature selection, which helped improve model stability and prevent overfitting.

Predictive Modeling

A hypothesis class $\mathscr{H}$ with seven ML architectures was chosen to reflect varied learning capacities and interpretability: LR, KNN, NN, SVM, DT, RF, and XGBoost. These algorithms differ in complexity, feature handling, and parameterization, allowing an in-depth examination of neighborhood-level heterogeneity. Each model learns a decision rule $h \in \mathscr{H}$ to minimize a loss function on the training set, with hyperparameters tuned by five-fold cross-validation.

Model performance was evaluated in terms of accuracy, precision, recall, F1-score, and area under the ROC curve (AUC) to provide a comprehensive assessment. Since our primary goal was to correctly identify high-prevalence neighborhoods, we paid particular attention to recall (i.e., sensitivity). We optimized hyperparameters for each algorithm using grid search with cross-validation. Moreover, for models where feature selection could simplify the model without sacrificing performance (i.e., LR, NN, DT, RF, and XGBoost), we employed RFECV to prune away redundant features. We visualized model outcomes using confusion matrices and ROC curves on both training and test sets to check for any overfitting and to compare performance across models. After identifying the strongest performers, we further evaluated those models on the independent external validation sets to assess generalizability. To interpret the predictions, we applied SHAP to each trained model, obtaining both local (neighborhood-level) and global (model-level) explanations by quantifying each feature's contribution to the prediction[50]. Table 4 summarizes the final features retained by each model and the optimal hyperparameters determined through grid search. Interestingly, certain features recurred across multiple top-performing models, including `obesity` and `overweight rates`, `active transportation usage`, overall `physical activity level`, `daily smoking rate`, `visible minority rate`, and `recent immigrant rate`. These features' consistent appearance shows their strong link to diabetes prevalence and highlights them as potential targets for population-level interventions.

Model	# of features	Optimal set of features	Model parameters
LR	5	Obese rate, Overweight rate, Active transportation rate, Rate of active population, Log median age	C[1]: 0.3, *penalty*: l1, *solver*: liblinear
KNN	5[2]	Obese rate, Overweight rate, Active transportation rate, Rate of active population, Log median age	Optimal k (using Elbow method): 6, *algorithm*: auto, *metric*: cosine and *weights*: uniform
NN[3]	7	Obese rate, Overweight rate, Active transportation rate, Rate of active population, Log median age, Visible minority rate, Daily smokers rate	*activation*: tanh, *hidden_layer_sizes*: (128, 64), *solver*: lbfgs
SVM	5[2]	Obese rate, Overweight rate, Active transportation rate, Rate of active population, Log median age	C[1]: 10, *gamma*: 0.1, and *kernel*: rbf[4]
DT	4	Obese rate, Overweight rate, and Log median age, Visible minority rate	*max_depth*: 20, *min_samples_leaf*: 5, *max_features*: square root, and *splitter*: best
RF	7	Obese rate, Overweight rate, Active transportation rate, Rate of active population, Log median age, Visible minority rate, Daily smokers rate	*max_depth*: 10, *min_samples_leaf*: 3, *max_features*: log2, and *n_estimators*: 150
XGBoost	6	Obese rate, Overweight rate, Rate of active population, Visible minority rate, Daily smokers rate, Recent immigrant rate	*max_depth*: 5, *min_samples_leaf*: 10, *n_estimators*: 150, and *learning_rate*: 0.1

[1] *Regularization Parameter*, [2] *Same optimal feature set as LR model*, [3] *Multi-Layer Perceptron*, [4] *Radial Basis Function*

Table 4. Summary of final feature sets (determined by RFECV) and model parameters (determined by grid search with 5-fold cross-validation) across the seven machine learning models.

All models were implemented in `Python` 3.11 using `Scikit-learn` 1.6.1. The code and trained model artifacts are available via our GitHub repository [`https://github.com/HIVE-UofT/diabetes-analysis`], and an interactive dashboard has been deployed for exploration of the results [`https://nediaml.hivelab-uoft.ca/`]. We exported trained models using `PyMilo` [`https://github.com/openscilab/pymilo`], which saves each model's parameters in a secure, non-executable JSON format to ensure cross-platform compatibility and to mitigate security concerns associated with binary serialization, an important consideration in healthcare applications.

Causal Inference through Causal ML: Causal Forest

Causal-ML methods move beyond correlation by estimating how a change in a specific factor would alter the outcome[97,98]. While randomized controlled trials are the gold standard for establishing causality, conducting an RCT at the community level (e.g., randomly altering neighborhood conditions) is often infeasible. Instead, we analyze observational data with causal-ML techniques that adjust for confounding and approximate experimental conditions. After controlling for the covariates (features) listed in Table 3, we assume that the treatment assignment is effectively independent of the diabetes rate each neighbourhood would otherwise have experienced. Under this 'ignorability' assumption, any remaining gap in prevalence between treated and control areas represents the treatment's causal effect and indicates how diabetes rates might shift if the neighbourhood feature were modified.

We work with the same set of neighbourhoods, $\mathscr{D} = \{(x_i, y_i)\}$, and introduce a binary treatment indicator $t_i \in \{0, 1\}$ for each area; a value of $t_i = 1$ denotes the presence of a favourable condition or intervention, whereas $t_i = 0$ denotes its absence. The potential outcomes $Y(0)$ and $Y(1)$ represent the diabetes prevalence under control and treatment, respectively. The Conditional Average Treatment Effect (CATE) for neighbourhoods characterized by $X = x$ is $\tau(x) = \mathbb{E}[Y(1) - Y(0) \mid X = x]$, the expected change in prevalence obtained by switching from control to treatment for areas with profile x.

We implemented Causal Forest using Microsoft Research's `EconML` library, which provides an adaptation of Generalized Random Forests for heterogeneous treatment effect estimation[99]. Causal Forest partition the feature space such that neighborhoods with similar characteristics but different treatment statuses can be compared in a quasi-experimental fashion. The algorithm thus finds subgroups of neighborhoods where an intervention, such as increasing access to green space or healthy-food outlets, has a consistently different effect. By estimating both $\hat{Y}(0)$ and $\hat{Y}(1)$ for every neighbourhood, we infer how local prevalence would change if the intervention were adopted, thereby revealing where public-health measures are likely to yield the greatest benefit.

We used a total of 10,000 trees in each Causal Forest to stabilize the treatment effect estimates. Model parameters (e.g., tree depth and minimum samples per leaf) were tuned via 10-fold cross-validation on the training data. For the outcome and treatment models inside the Causal Forest, we used a Lasso regression as the base learner. The Lasso imposes an L_1 regularization penalty that encourages sparsity. This regularization reduces over-fitting by selecting a concise set of predictors and yields more reliable estimates of $Y(0)$, $Y(1)$, and the treatment propensity, an important consideration for a high-dimensional, observational dataset.

Acknowledgements

This research is supported by the Vector Scholarship in Artificial Intelligence (Vector Institute), the Data Sciences Institute at the University of Toronto (DSI-PDFY3R1P31), and the Artificial Intelligence for Public Health (AI4PH) Trainee Award from the Canadian Institutes of Health Research (CIHR). We also acknowledge funding from the Natural Sciences and Engineering Research Council of Canada (NSERC) through the Canada Research Chairs Program and Discovery Grant (RGPIN-2025-07037).

Author contributions statement

M.N., A.R., and I.G. jointly designed and implemented the machine learning pipeline, conducted the data analysis, and drafted the manuscript. M.N. and A.R. led the interpretation and analysis of results by developing the causal ML framework and creating visualizations. A.R. contributed to the web platform development. Z.S. provided supervision and contributed to data visualization. O.S., K.K., J.R.B,, and Z.S. provided critical feedback throughout the revision process. All authors reviewed and approved the final manuscript.

Conflicts of interest

All authors declare no competing interests.

Data availability

The data used to develop the machine learning and causal inference models in this study are publicly available in the project's GitHub repository at `https://github.com/HIVE-UofT/diabetes-analysis`. All relevant datasets have been provided to facilitate replication of the study findings.

References

1. World Health Organization. Diabetes (2024). World Health Organization health topic page.
2. Public Health Agency of Canada. Framework for diabetes in canada (2022).
3. Frankish, C. J. *et al.* Addressing the non-medical determinants of health: a survey of canada's health regions. *Can. J. Public Heal.* **98**, 41–47 (2007).
4. Hill-Briggs, F. *et al.* Social determinants of health and diabetes: a scientific review. *Diabetes care* **44**, 258 (2021).
5. Awuor, L. & Melles, S. The influence of environmental and health indicators on premature mortality: An empirical analysis of the city of toronto's 140 neighborhoods. *Heal. & Place* **58**, 102155 (2019).
6. Doubleday, A. *et al.* Neighborhood greenspace and risk of type 2 diabetes in a prospective cohort: the multi-ethnic study of atherosclerosis. *Environ. Heal.* **21**, 1–10 (2022).
7. Robinette, J. W., Boardman, J. D. & Crimmins, E. M. Differential vulnerability to neighbourhood disorder: a gene× environment interaction study. *J Epidemiol Community Heal.* **73**, 388–392 (2019).
8. Sheets, L. *et al.* The effect of neighborhood disadvantage on diabetes prevalence. In *AMIA annual symposium proceedings*, vol. 2017, 1547 (American Medical Informatics Association, 2017).
9. Creatore, M. I. *et al.* Association of neighborhood walkability with change in overweight, obesity, and diabetes. *Jama* **315**, 2211–2220 (2016).
10. Collins, G. S., Mallett, S., Omar, O. & Yu, L.-M. Developing risk prediction models for type 2 diabetes: a systematic review of methodology and reporting. *BMC medicine* **9**, 1–14 (2011).
11. Richards, S. E., Wijeweera, C. & Wijeweera, A. Lifestyle and socioeconomic determinants of diabetes: Evidence from country-level data. *Plos one* **17**, e0270476 (2022).
12. Rautio, N. *et al.* Accumulated exposure to unemployment is related to impaired glucose metabolism in middle-aged men: A follow-up of the northern finland birth cohort 1966. *Prim. care diabetes* **11**, 365–372 (2017).
13. Booth, G. L. *et al.* Unwalkable neighborhoods, poverty, and the risk of diabetes among recent immigrants to canada compared with long-term residents. *Diabetes Care* **36**, 302–308 (2013).
14. Rundle, A. G. *et al.* Neighbourhood walkability is associated with risk of gestational diabetes: A cross-sectional study in new york city. *Paediatr. Perinat. Epidemiol.* **37**, 212–217 (2023).
15. Ancker, J. S., Kim, M.-H., Zhang, Y., Zhang, Y. & Pathak, J. The potential value of social determinants of health in predicting health outcomes. *J. Am. Med. Informatics Assoc.* **25**, 1109–1110 (2018).
16. Iparraguirre-Villanueva, O., Espinola-Linares, K., Flores Castañeda, R. O. & Cabanillas-Carbonell, M. Application of machine learning models for early detection and accurate classification of type 2 diabetes. *Diagnostics* **13**, 2383 (2023).
17. Alfalki, A. M. Using machine learning and artificial intelligence to predict diabetes mellitus among women population. *Curr. diabetes reviews* (2025).
18. Tasin, I., Nabil, T. U., Islam, S. & Khan, R. Diabetes prediction using machine learning and explainable ai techniques. *Healthc. Technol. Lett.* **10**, 1–10 (2023).
19. Perveen, S., Shahbaz, M., Keshavjee, K. & Guergachi, A. Prognostic modeling and prevention of diabetes using machine learning technique. *Sci. reports* **9**, 13805 (2019).
20. Lu, K. *et al.* Identifying prediabetes in canadian populations using machine learning. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1–4 (IEEE, 2024).
21. Nguyen, M., Jankovic, I., Kalesinskas, L., Baiocchi, M. & Chen, J. H. Machine learning for initial insulin estimation in hospitalized patients. *J. Am. Med. Informatics Assoc.* **28**, 2212–2219 (2021).
22. De Silva, K., Jönsson, D. & Demmer, R. T. A combined strategy of feature selection and machine learning to identify predictors of prediabetes. *J. Am. Med. Informatics Assoc.* **27**, 396–406 (2020).
23. Ljubic, B. *et al.* Predicting complications of diabetes mellitus using advanced machine learning algorithms. *J. Am. Med. Informatics Assoc.* **27**, 1343–1351 (2020).
24. Swapna, G., Vinayakumar, R. & Soman, K. Diabetes detection using deep learning algorithms. *ICT express* **4**, 243–246 (2018).
25. Choi, S. B. *et al.* Screening for prediabetes using machine learning models. *Comput. mathematical methods medicine* **2014**, 618976 (2014).

26. Zhu, T., Li, K., Herrero, P. & Georgiou, P. Deep learning for diabetes: a systematic review. *IEEE J. Biomed. Heal. Informatics* **25**, 2744–2757 (2020).

27. Feng, W. *et al.* Deep learning based prediction of depression and anxiety in patients with type 2 diabetes mellitus using regional electronic health records. *Int. J. Med. Informatics* 105801 (2025).

28. Samsel, K. *et al.* Predicting depression in canadians with or at risk of diabetes: A cross-sectional machine learning analysis. *medRxiv* 2024–02 (2024).

29. Roy, T. & Lloyd, C. E. Epidemiology of depression and diabetes: a systematic review. *J. affective disorders* **142**, S8–S21 (2012).

30. Saha, P. *et al.* Predicting time to diabetes diagnosis using random survival forests. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1–4 (IEEE, 2024).

31. Liu, B., Li, Y., Sun, Z., Ghosh, S. & Ng, K. Early prediction of diabetes complications from electronic health records: A multi-task survival analysis approach. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32 (2018).

32. Abbasian, M. *et al.* Knowledge-infused llm-powered conversational health agent: A case study for diabetes patients. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 1–4 (IEEE, 2024).

33. Dao, D., Teo, J. Y. C., Wang, W. & Nguyen, H. D. Llm-powered multimodal ai conversations for diabetes prevention. In *Proceedings of the 1st ACM Workshop on AI-Powered Q&A Systems for Multimedia*, 1–6 (2024).

34. Mohsen, F., Al-Absi, H. R., Yousri, N. A., El Hajj, N. & Shah, Z. A scoping review of artificial intelligence-based methods for diabetes risk prediction. *npj Digit. Medicine* **6**, 197 (2023).

35. Feng, C. & Jiao, J. Predicting and mapping neighborhood-scale health outcomes: A machine learning approach. *Comput. Environ. Urban Syst.* **85**, 101562 (2021).

36. Pearl, J. & Mackenzie, D. *The book of why: the new science of cause and effect* (Basic Books, 2018).

37. Athey, S. & Imbens, G. W. The state of applied econometrics: Causality and policy evaluation. *J. Econ. Perspectives* **31**, 3–32 (2017).

38. Diez Roux, A. V. & Mair, C. Neighborhood environments and diabetes risk and control. *Curr. diabetes reports* **16**, 1–9 (2016).

39. Christine, P. J. *et al.* Longitudinal associations between neighborhood physical and social environments and incident type 2 diabetes mellitus: the multi-ethnic study of atherosclerosis (mesa). *JAMA internal medicine* **175**, 1311–1320 (2015).

40. Statistics Canada. Census profile, 2021 census of population (2022).

41. Kahn, S. E., Hull, R. L. & Utzschneider, K. M. Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature* **444**, 840–846 (2006).

42. Eckel, R. H. *et al.* Obesity and type 2 diabetes: what can be unified and what needs to be individualized? *The J. Clin. Endocrinol. & Metab.* **96**, 1654–1663 (2011).

43. Canada, S. Neighbourhood characteristics and life satisfaction of individuals in lower-, middle-, and higher-income families in Canadian Metropolitan Areas (2021).

44. Seligman, H. K., Bindman, A. B., Vittinghoff, E., Kanaya, A. M. & Kushel, M. B. Food insecurity is associated with diabetes mellitus: results from the national health examination and nutrition examination survey (nhanes) 1999–2002. *J. general internal medicine* **22**, 1018–1023 (2007).

45. Chawla, N., Bowyer, K., Hall, L. & Kegelmeyer, W. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002).

46. Sun, Y., Wong, A. & Kamel, M. Classification of imbalanced data: A review. *Int. J. Pattern Recognit. Artif. Intell.* **23**, 687–719 (2009).

47. Breiman, L. Random forests. *Mach. learning* **45**, 5–32 (2001).

48. Fisher, A., Rudin, C. & Dominici, F. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* **20**, 1–81 (2019).

49. Den Braver, N. R. *et al.* Built environmental characteristics and diabetes: a systematic review and meta-analysis. *BMC medicine* **16**, 12 (2018).

50. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In Guyon, I. *et al.* (eds.) *Advances in Neural Information Processing Systems*, vol. 30 (Curran Associates, Inc., 2017).

51. Molnar, C. *Interpretable machine learning* (Lulu. com, 2020).

52. Janizek, J. D., Sturmfels, P. & Lee, S. I. Explaining explanations: Axiomatic feature interactions for deep networks. *J. Mach. Learn. Res.* **22**, 1–54 (2021).

53. Sallis, J. F., Floyd, M. F., Rodriguez, D. A. & Saelens, B. E. Role of built environments in physical activity, obesity, and cardiovascular disease. *Circulation* **125**, 729–737 (2012).

54. Mueller, N. *et al.* Health impact assessment of active transportation: A systematic review. *Prev. Medicine* **76**, 103–114 (2015).

55. Gluckstein. Will removing bike lanes impact cyclist safety? https://www.gluckstein.com/news-item/will-removing-bike-lanes-impact-cyclist-safety (2021). Accessed: 2025-02-20.

56. Koohsari, M. J. *et al.* Public open space, physical activity, urban design and public health: Concepts, methods and research agenda. *Heal. & Place* **33**, 75–82 (2015).

57. Nieuwenhuijsen, M. J. Urban and transport planning, environmental exposures and health-new concepts, methods and tools to improve health in cities. *Environ. Heal.* **15**, 161–171 (2016).

58. Rajkomar, A. *et al.* Scalable and accurate deep learning with electronic health records. *npj Digit. Medicine* **1**, 1–10 (2018).

59. Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Medicine* **25**, 44–56 (2019).

60. Williams, D. R. & Collins, C. Racial residential segregation: a fundamental cause of racial disparities in health. *Am. J. Public Heal.* **91**, 675–682 (2001).

61. Kershaw, K. N., Osypuk, T. L., Do, D. P., De Chavez, P. J. & Diez Roux, A. V. Neighborhood-level racial/ethnic residential segregation and incident cardiovascular disease: the multi-ethnic study of atherosclerosis. *Circulation* **131**, 141–148 (2015).

62. Zhang, D., van Meijgaard, J., Shi, L., Cole, B. & Fielding, J. Does neighbourhood composition modify the association between acculturation and unhealthy dietary behaviours? *J Epidemiol Community Heal.* (2015).

63. Ng, E. The healthy immigrant effect and mortality rates. *Heal. Reports* **22**, C1 (2011).

64. Canada, S. Population and demography statistics (2021). Accessed: 2025-01-02.

65. Anderson, R. J., Freedland, K. E., Clouse, R. E. & Lustman, P. J. The prevalence of comorbid depression in adults with diabetes: a meta-analysis. *Diabetes care* **24**, 1069–1078 (2001).

66. Nouwen, A. *et al.* Type 2 diabetes mellitus as a risk factor for the onset of depression: a systematic review and meta-analysis. *Diabetologia* **53**, 2480–2486 (2010).

67. Holt, R. I., De Groot, M. & Golden, S. H. Diabetes and depression. *Curr. diabetes reports* **14**, 1–9 (2014).

68. Mezuk, B., Eaton, W. W., Albrecht, S. & Golden, S. H. Depression and type 2 diabetes over the lifespan: a meta-analysis. *Diabetes care* **31**, 2383–2390 (2008).

69. Joseph, J. J. & Golden, S. H. Cortisol dysregulation: the bidirectional link between stress, depression, and type 2 diabetes mellitus. *Annals New York Acad. Sci.* **1391**, 20–34 (2017).

70. Yan, Y.-X. *et al.* Investigation of the relationship between chronic stress and insulin resistance in a chinese population. *J. epidemiology* **26**, 355–360 (2016).

71. Mawani, F. N. & Gilmour, H. Validation of self-rated mental health. *Heal. reports* **21**, 61–75 (2010).

72. Wright, E., Pagliaro, C., Page, I. S. & Diminic, S. A review of excluded groups and non-response in population-based mental health surveys from high-income countries. *Soc. Psychiatry Psychiatr. Epidemiol.* **58**, 1265–1292 (2023).

73. Cheung, K. L., Ten Klooster, P. M., Smit, C., de Vries, H. & Pieterse, M. E. The impact of non-response bias due to sampling in public health studies: A comparison of voluntary versus mandatory recruitment in a dutch national survey on adolescent health. *BMC public health* **17**, 276 (2017).

74. Kivimäki, M. *et al.* Long working hours, socioeconomic status, and the risk of incident type 2 diabetes: a meta-analysis of published and unpublished data from 222 120 individuals. *The lancet Diabetes & endocrinology* **3**, 27–34 (2015).

75. Nyberg, S. T. *et al.* Job strain and cardiovascular disease risk factors: meta-analysis of individual-participant data from 47,000 men and women. *PloS one* **8**, e67323 (2013).

76. Kolb, H. & Martin, S. Environmental/lifestyle factors in the pathogenesis and prevention of type 2 diabetes. *BMC medicine* **15**, 131 (2017).

77. Dendup, T., Feng, X., Clingan, S. & Astell-Burt, T. Environmental risk factors for developing type 2 diabetes mellitus: a systematic review. *Int. journal environmental research public health* **15**, 78 (2018).

78. Openshaw, S. *The modifiable areal unit problem* (Concepts and Techniques in Modern Geography, 1984).

79. Xie, Z., Nikolayeva, O., Luo, J. & Li, D. Building risk prediction models for type 2 diabetes using machine learning techniques. *Prev. Chronic Dis.* **16** (2019).

80. Choi, S. G. *et al.* Comparisons of the prediction models for undiagnosed diabetes between machine learning versus traditional statistical methods. *Sci. Reports* **13**, 13101 (2023).

81. Pearl, J. *Causality: Models, reasoning, and inference* (Cambridge University Press, 2009), 2nd edn.

82. Craig, P., Katikireddi, S. V., Leyland, A. & Popham, F. Natural experiments: An overview of methods, approaches, and contributions to public health intervention research. *Annu. Rev. Public Heal.* **38**, 39–56 (2017).

83. Austin, P. C. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivar. Behav. Res.* **46**, 399–424 (2011).

84. Larabi-Marie-Sainte, S., Aburahmah, L., Almohaini, R. & Saba, T. Current techniques for diabetes prediction: review and case study. *Appl. Sci.* **9**, 4604 (2019).

85. Wang, R., Chaudhari, P. & Davatzikos, C. Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies. *Proc. Natl. Acad. Sci.* **120**, e2211613120 (2023).

86. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G. & Chin, M. H. Ensuring fairness in machine learning to advance health equity. *Annals internal medicine* **169**, 866–872 (2018).

87. Khoury, M. J., Iademarco, M. F. & Riley, W. T. Precision public health for the era of precision medicine. *Am. journal preventive medicine* **50**, 398 (2015).

88. Statistics Canada. Dictionary, census of population, 2021-census tract (ct). https://tinyurl.com/45azyjm3 (2021). Government of Canada.

89. Statistics Canada. Canadian community health survey-annual component (cchs) (2024).

90. Environics Analytics. Demostats: Canada's most up-to-date demographic database (2024).

91. MAP Centre for Urban Health Solutions. Ontario marginalization index (on-marg). https://maphealth.ca/on-marg/ (2024). St. Michael's Hospital, Data Tool Description.

92. Environics Analytics. Demostats release notes. Technical Document (2023). Environics Analytics.

93. Matheson, F. I. & van Ingen, T. Ontario marginalization index: User guide. Technical Report (2011). St. Michael's Hospital, Toronto, ON.

94. SimplyAnalytics. Analytics for everyone. https://simplyanalytics.com/ (2024).

95. Public Health Ontario. Ontario marginalization index (on-marg) (2021).

96. Lipscombe, L. L. *et al.* Current state of type 2 diabetes in the peel region. Tech. Rep., Novo Nordisk Network for Healthy Populations (2024). White paper.

97. Athey, S. & Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proc. Natl. Acad. Sci.* **113**, 7353–7360 (2016).

98. Wager, S. & Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* **113**, 1228–1242 (2018).

99. Athey, S. & Wager, S. Estimating treatment effects with causal forests: An application. *Obs. studies* **5**, 37–51 (2019).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfile.pdf