2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) | 979-8-3503-7149-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/EMBC53108.2024.10782210

Predicting Time to Diabetes Diagnosis Using Random Survival Forests

Priyonto Saha[1], Yacine Marouf[1], Hunter Pozzebon[1], Aziz Guergachi[2,3], Karim Keshavjee[3], Mohammad Noaeen[1], and Zahra Shakeri[3]

***Abstract*— Type 2 Diabetes Mellitus (T2DM) is a chronic metabolic disorder with increasing population incidence. However, T2DM takes years to develop, allowing onset prediction and prevention to be a clinically effective treatment strategy. In this study we propose and assess a novel approach to diabetes prediction which integrates a specialized extension of the random forest algorithm known as random survival forest (RSF). Rather than predicting a binary outcome, this machine learning model incorporates survival analysis methodology to predict the time until a patient will receive a diabetes diagnosis if their current lifestyle is maintained. We trained a baseline model on 7,704 electronic medical records from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN) with 14 biomarker and comorbidity features across different measurement dates. Although tuning parameters were purposefully chosen for quick training rather than for predictive performance, our model exceeded expectations with a concordance index of 0.84. Thus, RSF models have been shown to produce accurate timelines of diabetes onset trajectory, providing patients with quantifiable and relatable risks that are easy to understand. The results of our study have substantial implications for advancing machine learning in clinical decision support and patient outcome predictions, emphasizing the role of innovative models in improving predictive accuracy.**

I. INTRODUCTION

Type 2 Diabetes Mellitus (T2DM) is a chronic metabolic disorder where patients have difficulty controlling blood glucose levels due to insulin insensitivity [1]. As of 2019, 8.8% of Canadians are living with diabetes and there are approximately 549 new diagnoses daily [2]. Type 2 diabetes is incurable, so prevention and delaying onset is the best defense against this growing epidemic [3, 4]. The predominant methods of prevention include lifestyle and diet adjustments [5], and multiple studies report significant correlations between metabolic biomarkers, exercise rates, and diabetes incidence [6, 7].

Prior studies attempting to predict diabetes using Electronic medical records (EMR) often use traditional machine learning models that do not capture longitudinal measurements [8–14]. Furthermore, studies using models with temporal features do not properly adjust for potential right censoring, that is, when a patient may develop diabetes after the study concludes.

Survival analysis methodology such as joint models and landmarking can address the right censoring prevalent in time-to-event data by predicting time to diabetes diagnosis. However, these techniques are often computationally intensive, make assumptions regarding underlying distributions, or struggle with analytical complexity [15].

To address these concerns, we propose the adoption of random survival forests (RSFs) for prognostic modeling applications. RSF is a non-parametric ensemble method of survival trees which account for time to event data and can utilize multiple longitudinal factors found in EMR data [16]. Notably, RSF was found to outperform both joint models and landmarking in both computational complexity and predictive performance [17]. By providing explicit time to diabetes diagnosis estimates, our model allows clinicians to better advise patients on their diabetes risk and collaborate towards personalized prevention plans for patients, thereby enhancing the efficacy of patient care strategies.

II. METHODS

A. Data Collection and Preparation

The Canadian Primary Care Sentinel Surveillance Network (CPCSSN) provided a dataset of EMRs consisting of a random sample of 10,000 records with 43 features from 8,602 unique adult patients for model training. The dataset was generated by compiling systolic blood pressure (sBP) measurements with the closest in time clinical measurements from within a certain time frame. Clinical measurements of body mass index (BMI), low-density lipoprotein (LDL), high-density lipoprotein (HDL), total cholesterol (TC), and triglyceride (TG) were within one year, HbA1c ($A1c$) was within three months, and fasting blood sugar (FBS) was within one month. The dataset also included age and sex of each patient alongside binary comorbidity indicators of depression, hypertension (HTN), osteoarthritis (OA), and chronic obstructive pulmonary disease (COPD). Dates for all clinical measurements and health condition diagnoses were also included, with clinical measurement dates ranging between 2003 and 2015 and health conditions dates ranging from 1989 to 2015.

In the preliminary data exploration phase, we conducted an analysis of missing data, computed summary statistics, and generated a correlation matrix to assess the relationships between variables. We employed histograms to examine the distribution of predictor variables and utilized boxplots to identify potential outliers. This suite of visualizations

[1]Priyonto Saha, Yacine Marouf, Hunter Pozzebon, and Mohammad Noaeen are with the Dalla Lana School of Public Health, University of Toronto, Canada.

[2]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.

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

was instrumental in uncovering patterns, anomalies, errors, and imbalances in the dataset, facilitating a more informed preprocessing strategy.

Survival time was calculated by taking the difference between the earliest date of a lab test or comorbidity diagnosis and their diabetes diagnosis date. The end of the study date, June 30th, 2015, was used for right-censored records where diabetes onset went unobserved. Left-censored individuals were removed, resulting in 7,756 eligible records with 14 predictors for the final dataset.

To identify the types of missing data, a series of logistic regression models was generated with the Statsmodels package[18]. For each feature, a binary missingness indicator was used as the response with all other features as covariates. Type of missingness was deemed missing completely at random (MCAR) if the logistic regression showed no significant covariate associations and missing at random (MAR) otherwise. Records where the only missing values were MCAR were dropped, and the remaining MAR missing data was imputed by chained equations through `IterativeImputer` from Scikit-Learn [19].

B. Random Survival Forest Model

The random survival forest model is an extension of the random forest model that accounts for the presence of censoring in survival data. It uses the log-rank splitting rule that splits nodes by maximization of the log-rank test statistic [16, 20]. The survival forest model was implemented through scikit-survival library in Python 3.10.12 [19, 21].

Hyperparameters for our model were explicitly chosen for rapid training rather than for optimal performance in order to assess the baseline (i.e. unoptimized and untuned) capabilities of RSF. This baseline random survival forest was trained with a 75:25 train-test split of our EMR data, consisting of 100 survival trees each with a max depth of 15, a minimum leaf sample of 100, and a minimum leaf split of 150. To ensure replicability and facilitate further research, all source code is available on GitHub[1].

III. Results and Discussion

A. Preliminary Results

The initial exploratory analysis revealed no anomalies, with visualizations indicating that all clinical measures fell within expected ranges and exhibited no significant imbalances. The correlation matrix (Figure 1) showed no unexplained correlation. High correlation between total cholesterol, HDL, and LDL is explained by cholesterol in blood primarily consisting of HDL and LDL. Patients with higher $A1c$ and FBS show higher rates of diabetes outcomes, which is expected since they are used as a diagnostic tool for diabetes [22]. Table I summarizes the 7,756 eligible records and their 14 non-chronological features before missing data processing, stratified on records with diabetes observed and diabetes unobserved. We determined that HDL and sBP were the only two features that were MCAR, which allowed us to drop records in which only these features were missing (52 records). LDL (39 records), TG (48 records), total cholesterol (175 records) and the remaining HDL records (16) were MAR and imputed, resulting in a sample size of 7,704. Since the imputed records make up less than 10% of the data we do not expect it to affect our results.

[1] https://github.com/P-Saha/RSF-Diabetes-Prevention

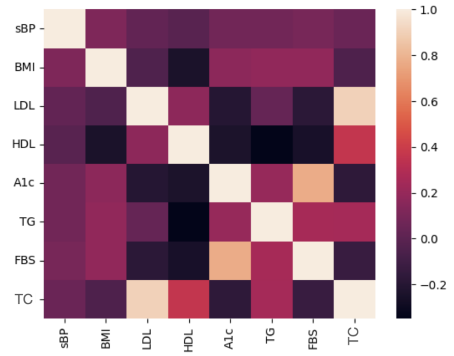

Fig. 1: Correlation matrix of measured biomarkers. Zero represents no correlation and 1 represents perfect correlation.

TABLE I: Summary statistics for 7,756 eligible records, stratified on records with diabetes observed and diabetes unobserved. Results are presented as mean [sd] for continuous variables or n (%) for categorical variables and missing values. P-values displayed are results of t-tests and chi-square tests for continuous and categorical features respectively

Features	Diagnosis(No)	Diagnosis(Yes)	p-value	Missingness
N (total=7756)	4861 (62.7)	2895 (37.3)	-	-
Age	60.86 [13.70]	65.30 [11.97]	< 0.001	0 (0.0%)
Sex (Male)	2058 (42.3)	1426 (49.3)	< 0.001	0 (0.0%)
Biomarkers				
sBP	129.25 [17.09]	132.46 [17.00]	< 0.001	3 (0.039%)
BMI	29.20 [6.48]	32.16 [6.95]	< 0.001	0 (0.0%)
LDL	2.85 [0.91]	2.34 [0.93]	< 0.001	39 (0.50%)
HDL	1.43 [0.43]	1.22 [0.35]	< 0.001	65 (0.84%)
A1c	5.71 [0.36]	6.70 [1.04]	< 0.001	0 (0.0%)
TG	1.42 [0.95]	1.74 [1.12]	< 0.001	48 (0.62%)
FBS	5.33 [0.63]	7.13 [1.95]	< 0.001	0 (0.0%)
Total Cholesterol	4.92 [1.08]	4.34 [1.12]	< 0.001	175 (2.26%)
Comorbidities				
Depression (Yes)	1090 (22.4)	723 (25.0)	0.011	0 (0.0%)
Hypertension (Yes)	2682 (55.2)	2300 (79.4)	< 0.001	0 (0.0%)
Osteoartitis (Yes)	1413 (29.1)	1051 (36.3)	< 0.001	0 (0.0%)
COPD (Yes)	441 (9.1)	325 (11.2)	0.002	0 (0.0%)

Our model had a concordance index (C-Index) of 0.84, which is the evaluation measure for random survival forest models [21]. This means that the model could accurately predict the relative order of diabetes diagnosis time of two randomly selected individuals 84% of the time [23].

Figure 2 shows the distribution of the median time of survival for diabetes diagnosis amongst records in the test

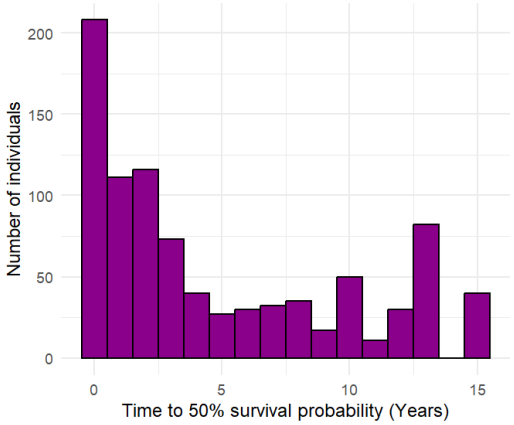

Fig. 2: Median diabetes diagnosis time histogram for test set.

set. In this context, the median time of survival is interpreted as the time before the model predicts that a record has over a 50% probability of diabetes diagnosis. Specifically, the bin into which a record is categorized signifies the estimated years until a patient is predicted to surpass a 50% probability of receiving a diabetes diagnosis. The median times for these records range from within one year to fifteen years after the earliest biomarker measurement or health condition diagnosis date. Only 923 records of 1926 records in the test set are represented in the figure, with the remaining 1023 records predicted to never go above a 50% chance of diabetes diagnosis. Table II provides feature importance means, calculated using `permutation_importance` from Scikit Learn [19, 21]. The most important model features were $A1c$, FBS, and LDL, with feature importance means of 0.12, 0.072, and 0.0036, respectively. The standard deviations for these means were all less than 0.01. The importance of these features support current literature and diabetes monitoring standards [22].

TABLE II: Feature Importance estimated with permutation

Features	Mean	Standard Deviation
A1c	0.12	0.0052
FBS	0.072	0.0050
LDL	0.0036	0.00087
Total Cholesterol	0.0030	0.00076
BMI	0.00079	0.00050
Age at Exam	0.00072	0.00026
HTN	0.00070	0.00044
HDL	0.00067	0.00030
TG	0.00055	0.00054
Sex	0.00011	0.000073
COPD	0.000045	0.000032
Depression	0.000030	0.000074
OA	0.000008	0.000075
sBP	-0.000067	0.000099

B. Key Findings

Overall, we found biomarkers and comorbidities commonly measured in health examinations to be important predictors for diabetes onset. To our knowledge, we are the first to utilize the RSF model to predict time to diabetes diagnosis through biomarkers and other comorbidites from a diverse Canadian population. Previous machine learning research has been conducted to predict diabetes, however they either utilize common random forests [24] or use RSFs for predicting diabetic complications [25]. One previous study employed the RSF model to predict the onset of diabetes, focusing on cardio-respiratory fitness and waist-to-height ratio as key indicators [26]. This research was a prospective study confined to healthy male subjects, thus not encompassing a wider demographic spectrum.

Gender Diabetes Disparity–Based on Table I, men had a 32% higher likelihood of diabetes diagnosis compared to women. (OR: 1.32, 95%; CI: [1.20;1.45]). Gender differences in diabetes diagnosis are noteworthy, with Canadian census data showing higher prevalence among men. Notably, women in lower-income brackets experience even greater disparities in prevalence rates [27].

Diabetes Biomarker Differences–T-tests from Table I showed that *all* measured factors were significantly different between patients who became diabetic and patients who remained non-diabetic. On average, patients with diabetes had sBP that was in the healthy range ($<$130) whereas patients with diabetes had sBP in the high-normal range. Diabetic patients had higher BMI ($BMI > 30$) and differences in LDL, HDL, and total cholesterol compared to non-diabetics, but both groups remained within recommended ranges [28]. Diabetic patients exhibited borderline high TG levels, whereas non-diabetic patients maintained normal TG levels [29]. Average $A1c$ and fasting blood glucose are also highly different between patients diagnosed with diabetes and healthy patients with varying health categories. Diabetes can be diagnosed through a fasting blood sugar measurement of over 7.0 mmol/L or through an A1c level above 6.5%. The average diabetic patient was above both those thresholds, while on average non-diabetic patients were well below the diabetic threshold and would not even fall within the prediabetic class, which requires a FBS between 6.1-6.9 mmol/L or an $A1c$ of 6.0%- 6.4% for its diagnosis [30].

Enhanced RSF Prediction Granularity–Figure 2 elucidates the innovative aspect of our methodology; in contrast to conventional models that would categorize 923 records as *true* for a diabetes diagnosis, the RSF model provides a more nuanced differentiation. Predictions are rendered in days to facilitate precise and individualized care strategies, albeit aggregated into yearly intervals for histogram representation. As illustrated in Figure 2, the predominant segment of patients at risk are forecasted to manifest diabetes within five years from their initial measurement. Records not represented in the histogram, failing to surpass a 50% risk threshold, imply a low likelihood of these individuals developing diabetes in their lifespan. Owing to RSF's capability to generate both a survival function and a cumulative hazard function per record, this inherent 50% benchmark can be adjusted to align with more conservative or liberal prognostic timelines. Therefore, the RSF model adeptly balances interpretability with comprehensive analytical depth, a challenge often encountered with alternative predictive models.

A C-Index of 0.84 indicates high predictive performance and shows great potential in the baseline capabilities of RSF. Furthermore, the simplicity of our pipeline leaves lots of room for optimization through parameter tuning. Although the flexible design of RSF was specifically proposed to accommodate the complexities of real world EMR, addressing data sparsity and imbalance provides another avenue for improvement. In particular for our dataset, the majority of patients had only a single record, multiple records were left-censored, and various records contained missing variables.

Longitudinal Data Challenges– Several advancements in random forest methodologies have been proposed to account for correlations among repeated measurements in longitudinal datasets [31]. Despite the potential of these enhancements to elevate predictive accuracy, our dataset's absence of repeated measures precludes their application. In practical scenarios, Electronic Medical Records (EMR) from clinical settings are anticipated to offer an abundance of repeated measurements, thus paving the way for enhanced

model refinement.

It is critical to distinguish between the dates of diagnosis and the actual onset of conditions, recognizing that patients may live with diabetes for a period before it is formally diagnosed. This scenario typifies interval censoring, a common challenge in survival analysis. The exact duration of this interval remains elusive in our study due to the prevalence of single-record patients, leading our model to estimate the time to diagnosis rather than the onset. Nonetheless, given the complications and potential emergencies arising from unmanaged diabetes, we posit that the interval between onset and diagnosis is likely minimal.

Looking ahead, our future work will concentrate on data generation processes to further refine our model's framework and assess the effectiveness of random survival forest model extensions. This endeavor aims to optimize predictive performance and applicability in real-world settings.

IV. Conclusion

Diabetes is a lifelong condition that may greatly reduce one's quality of life. The best defense against this growing epidemic is proactive prevention planning through lifestyle changes. However, implementing and maintaining these lifestyle changes can be challenging, especially without a solid understanding of the risks. This study advocates the use of random survival forest to predict the time to diabetes diagnosis, offering a quantifiable and relatable risk that is easy to comprehend. With a concordance index of 0.84 at baseline, random survival forest models have proven to be effective and accessible as a clinical tool with significant potential for further improvement. The complexities observed in our data reflect real-world constraints in electronic medical records. The random survival forest model excels under these conditions due to its robust non-parametric design. With such a promising initial performance, we are eager to explore additional applications of random survival forest in other prognostic modeling tasks.

References

[1] A. B. Olokoba, O. A. Obateru, and L. B. Olokoba, "Type 2 diabetes mellitus: A review of current trends," *Oman medical journal*, vol. 27, no. 4, p. 269, 2012.

[2] A. G. LeBlanc, Y. J. Gao, L. McRae, and C. Pelletier, "At-a-glance - twenty years of diabetes surveillance using the canadian chronic disease surveillance system," *Health Promotion and Chronic Disease Prevention in Canada*, vol. 39, pp. 306–309, 11 Nov. 2019, ISSN: 2368-738X.

[3] H. Sagesaka *et al.*, "Type 2 diabetes: When does it start?" *Journal of the Endocrine Society*, pp. 476–484, 5, ISSN: 2472-1972.

[4] D. P. P. R. Group, "Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin," *New England Journal of Medicine*, vol. 346, pp. 393–403, 6 Feb. 2002, ISSN: 0028-4793.

[5] A. D. Association, "3. Prevention or Delay of Type 2 Diabetes: Standards of Medical Care in Diabetes—2021," *Diabetes Care*, vol. 44, no. Supplement$_1$, S34–S39, Dec. 2020, ISSN: 0149-5992.

[6] M. Biavaschi, V. M. Melchiors Morsch, L. F. Jacobi, A. Hoppen, N. Bianchin, and M. R. Chitolina Schetinger, "Predisposition to type 2 diabetes in aspects of the glycemic curve and glycated hemoglobin in healthy, young adults: A cross-sectional study," *Canadian Journal of Diabetes*, vol. 47, no. 7, 587–593, May 2023.

[7] F. Ahmed, M. AL-Habori, E. Al-Zabedi, and R. Saif-Ali, "Impact of triglycerides and waist circumference on insulin resistance and -cell function in non-diabetic first-degree relatives of type 2 diabetes," *BMC Endocrine Disorders*, vol. 21, no. 1, 2021.

[8] S. Perveen, M. Shahbaz, K. Keshavjee, and A. Guergachi, "Prognostic modeling and prevention of diabetes using machine learning technique," *Scientific Reports*, vol. 9, p. 13 805, 1 Sep. 2019, ISSN: 2045-2322.

[9] S. Perveen, M. Shahbaz, M. S. Ansari, K. Keshavjee, and A. Guergachi, "A hybrid approach for modeling type 2 diabetes mellitus progression," *Frontiers in Genetics*, vol. 10, Jan. 2020, ISSN: 1664-8021.

[10] I. Naveed, M. F. Kaleem, K. Keshavjee, and A. Guergachi, "Artificial intelligence with temporal features outperforms machine learning in predicting diabetes," *PLOS Digital Health*, vol. 2, e0000354, 10 Oct. 2023, ISSN: 2767-3170.

[11] A. Dekamin, M. Wahab, K. Keshavjee, and A. Guergachi, "High cardiovascular disease risk-associated with the incidence of type 2 diabetes among prediabetics," *European Journal of Internal Medicine*, vol. 106, pp. 56–62, 2022.

[12] K. Lu *et al.*, "Identifying prediabetes in canadian populations using machine learning," in *Proceedings of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2024.

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

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

[15] L. Ferrer, H. Putter, and C. Proust-Lima, "Individual dynamic predictions using landmarking and joint modelling: Validation of estimators and robustness assessment," *Statistical Methods in Medical Research*, vol. 28, pp. 3649–3666, 12 Dec. 2019, ISSN: 14770334.

[16] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Annals of Applied Statistics*, vol. 2, pp. 841–860, 3 Sep. 2008, ISSN: 19326157.

[17] K. L. Pickett, K. Suresh, K. R. Campbell, S. Davis, and E. Juarez-Colunga, "Random survival forests for dynamic predictions of a time-to-event outcome using a longitudinal biomarker," *BMC Medical Research Methodology*, vol. 21, 1 Dec. 2021, ISSN: 14712288.

[18] S. Seabold and J. Perktold, "Statsmodels: Econometric and statistical modeling with python," in *9th Python in Science Conference*, 2010.

[19] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[20] J.-C. Hsu, Y.-Y. Yang, S.-L. Chuang, L.-Y. Lin, and T. H.-H. Chen, "Prediabetes as a risk factor for new-onset atrial fibrillation: The propensity-score matching cohort analyzed using the cox regression model coupled with the random survival forest," *Cardiovascular Diabetology*, vol. 22, no. 1, pp. 1–11, 2023.

[21] S. Pölsterl, "Scikit-survival: A library for time-to event analysis built on top of scikit-learn," *Journal of Machine Learning Research*, vol. 21, no. 212, pp. 1–6, 2020.

[22] R. Goyal, M. Singhal, and I. Jialal, *Type 2 Diabetes*. 2023.

[23] A. Alabdallah, M. Ohlsson, S. Pashami, and T. Rögnvaldsson, "The concordance index decomposition - a measure for a deeper understanding of survival prediction models," *SSRN Electronic Journal*, 2022.

[24] T. Ooka, H. Johno, K. Nakamoto, Y. Yoda, H. Yokomichi, and Z. Yamagata, "Random forest approach for determining risk prediction and predictive factors of type 2 diabetes: Large-scale health check-up data in japan," *Health*, vol. 4, p. 200, 2021.

[25] M. Li *et al.*, "Multicomponent prediction of 2-year mortality and amputation in patients with diabetic foot using a random survival forest model: Uric acid, alanine transaminase, urine protein and platelet as important predictors," 2023.

[26] R. A. Sloan *et al.*, "A fit-fat index for predicting incident diabetes in apparently healthy men: A prospective cohort study," 2016.

[27] Canada, Public Health Agency, "Inequalities in diabetes and related risk factors: Comparing canadian adults by income level,"

[28] informedhealth.org, *High Cholesterol: Overview*. Institute for Quality and Efficiency in Health Care, Sep. 2017.

[29] Mount Sinai Hospital, *Triglyceride level*, Nov. 2022.

[30] Z. Punthakee, R. Goldenberg, and P. Katz, "Definition, classification and diagnosis of diabetes, prediabetes and metabolic syndrome," *Canadian Journal of Diabetes*, vol. 42, S10–S15, Apr. 2018, ISSN: 23523840.

[31] J. Hu and S. Szymczak, "A review on longitudinal data analysis with random forest," *Briefings in Bioinformatics*, vol. 2023, pp. 1–11, 2.