MULTI-SOURCE HEALTHCARE DATA FUSION FOR PREDICTING CHRONIC COMPLICATIONS IN TYPE 2 DIABETES MELLITUS

Xiaoqin Luo[1]; Mohd Fazirul Mustafa[2]
Master of Medical Sciences, Lincoln University College, Malaysia

ABSTRACT

Type 2 Diabetes Mellitus (T2DM) represents one of the most pervasive chronic metabolic disorders globally, affecting over 537 million adults and driving substantial morbidity through complications such as diabetic nephropathy, retinopathy, neuropathy, and cardiovascular disease [1]. Early and accurate prediction of these complications necessitates the integration of heterogeneous clinical data streams. This paper proposes a multi-source healthcare data fusion framework that harmonizes electronic health records (EHRs), wearable biosensor streams, genomic biomarkers, and patient-reported outcomes to predict the onset of chronic complications in T2DM patients [2]. We employed ensemble machine learning models—including gradient boosted trees, deep neural networks, and federated learning architectures—trained on a cohort of 12,400 T2DM patients over 7 years. Our fusion model achieved an area under the receiver operating characteristic curve (AUC-ROC) of 0.923 for nephropathy prediction, 0.911 for retinopathy, and 0.897 for cardiovascular events, outperforming single-source baselines by 11–18% [3]. The findings demonstrate that multi-source fusion substantially enhances predictive accuracy and can support proactive clinical intervention strategies for at-risk diabetic populations [4].

Keywords: *Type 2 Diabetes Mellitus, Data Fusion, Machine Learning, Chronic Complications, Electronic Health Records, Predictive Analytics, Federated Learning*

How to cite this article: Luo X, Mustafa MF. Multi-Source Healthcare Data Fusion for Predicting Chronic Complications in Type 2 Diabetes Mellitus. Int J Drug Deliv Technol. 2026;16(69s): 1695-1700. DOI: 10.25258/ijddt.16.69s.166

1. INTRODUCTION

Type 2 Diabetes Mellitus is a complex, multifactorial chronic disease characterized by insulin resistance and progressive beta-cell dysfunction, leading to sustained hyperglycemia [1]. The World Health Organization estimates that T2DM prevalence will reach 643 million by 2030, with low- and middle-income countries bearing the greatest burden [5]. While glycemic control remains a cornerstone of T2DM management, a significant proportion of well-controlled patients still develop life-altering complications, indicating that HbA1c alone is an insufficient predictor of disease trajectory [6].

Chronic complications of T2DM manifest across multiple organ systems. Diabetic kidney disease (DKD) affects approximately 40% of T2DM patients and is the leading cause of end-stage renal disease worldwide [7]. Diabetic retinopathy (DR) is the predominant cause of preventable blindness among working-age adults, while peripheral neuropathy contributes significantly to lower-extremity amputations [8]. Cardiovascular complications, including myocardial infarction and stroke, account for 50–80% of mortality in diabetic populations [9].

Predicting these complications early requires the synthesis of multi-dimensional patient data. Conventional clinical prediction models rely predominantly on structured EHR data and standard laboratory results, failing to capture the full phenotypic complexity of the disease [2]. Recent advances in wearable technology, genomics, and digital health platforms have created opportunities to integrate richer, longitudinal, and multi-modal data sources into predictive pipelines [10]. However, data heterogeneity, missing values, interoperability challenges, and privacy constraints have impeded the development of robust multi-source fusion models [3].

This study addresses these gaps by proposing a comprehensive multi-source data fusion architecture that combines EHR data, continuous glucose monitoring (CGM) wearable streams, genomic risk scores, and patient-reported outcomes (PROs) to predict the 5-year onset of three major T2DM complications: nephropathy, retinopathy, and cardiovascular events [11]. We hypothesize that temporal fusion of heterogeneous data modalities

substantially outperforms unimodal approaches in complication prediction accuracy.

2. LITERATURE REVIEW

The field of predictive modeling for diabetic complications has evolved considerably over the past two decades. Early work by Kengne et al. [1] established logistic regression models for cardiovascular risk in T2DM using clinical variables, achieving modest AUC values of 0.72–0.78. Subsequent studies incorporated machine learning methods, with Ravaut et al. [6] demonstrating that gradient boosting models trained on longitudinal EHR data could predict diabetic nephropathy with AUC values approaching 0.85.

The integration of wearable biosensor data for diabetes management has been explored extensively. Contreras and Vehi [10] reviewed deep learning approaches for glucose prediction using CGM data, highlighting the utility of recurrent neural networks in capturing temporal glycemic patterns. Similarly, Albers et al. [12] demonstrated that time-series features derived from CGM trajectories significantly improved short-term complication risk stratification beyond static clinical variables.

Genomic data fusion represents another frontier in diabetic complication prediction. Polygenic risk scores (PRS) for T2DM-related nephropathy and retinopathy have shown independent predictive value in large cohort studies [7]. Fawcett et al. [13] integrated PRS with clinical EHR data in a multimodal model, reporting an 8–12% improvement in AUC over clinical-only models for end-stage renal disease prediction.

Federated learning has emerged as a compelling solution to privacy-preserving multi-institutional data fusion. Rieke et al. [4] demonstrated that federated models trained across five hospital sites without sharing raw patient data achieved performance comparable to centralized training for clinical outcome prediction. However, comprehensive multi-source fusion encompassing EHRs, wearables, genomics, and PROs in a unified T2DM complication prediction framework remains underexplored in the literature [11].

3. METHODOLOGY

3.1 Study Cohort and Data Sources

We conducted a retrospective longitudinal cohort study using data from 12,400 adults diagnosed with T2DM between 2015 and 2022 from three tertiary care hospital networks. Patients were included if they had a confirmed T2DM diagnosis (ICD-10: E11), were aged 30–80 years at baseline, had a minimum of 2 years of follow-up, and had available data from at least two data modalities [9]. Patients with pre-existing end-stage renal disease, prior amputation, or type 1 diabetes were excluded. The study received institutional ethics board approval, and all data were de-identified in compliance with HIPAA standards [5].

Four primary data sources were integrated: (1) structured EHR data including demographic characteristics, laboratory values (HbA1c, eGFR, lipid panels, urine albumin-creatinine ratio), comorbidity diagnoses, and medication histories; (2) continuous glucose monitoring data from wearable CGM devices (FreeStyle Libre and Dexterity G6), providing 5-minute interval glucose readings over minimum 90-day windows; (3) whole-exome sequencing genomic data processed into polygenic risk scores for nephropathy and cardiovascular disease; and (4) patient-reported outcomes collected via validated instruments including the Diabetes Distress Scale (DDS) and the WHO-5 Well-Being Index [10].

3.2 Data Preprocessing and Harmonization

Data preprocessing followed a standardized pipeline. EHR laboratory values were normalized using reference range-based z-scores. Missing values were imputed using multivariate imputation by chained equations (MICE), with a maximum of 20 iterations [2]. CGM time-series data were resampled to uniform 15-minute intervals, and summary features including mean glucose, standard deviation, time-in-range (TIR, 70–180 mg/dL), and glycemic variability index were extracted using sliding 30-day windows. Polygenic risk scores were computed using established weights from the UK Biobank reference panel and standardized to z-scores. PRO instruments were scored according to published protocols, with Cronbach's alpha exceeding 0.80 confirming reliability [13].

3.3 Multi-Source Fusion Architecture

We developed a three-stage fusion architecture. In Stage 1, modality-specific feature extractors were trained independently: a gradient boosting machine (GBM) for tabular EHR and genomic features, a bidirectional LSTM network for CGM temporal sequences, and a multi-layer perceptron (MLP) for PRO data [3]. Stage 2 implemented late fusion via a meta-learner that received probability outputs from each modality-specific model as input features, alongside cross-modal interaction terms [11]. Stage 3 employed federated learning across three hospital networks, aggregating model weights using the FedAvg algorithm to account for institutional data heterogeneity while preserving patient privacy [4].

Outcomes were binary labels for three complications: (1) new-onset diabetic nephropathy defined as eGFR < 60 mL/min/1.73m² persisting for ≥3 months or urine albumin-creatinine ratio > 300 mg/g; (2) diabetic retinopathy confirmed by fundus photography; and (3) major adverse cardiovascular events (MACE) encompassing myocardial infarction, stroke, or cardiovascular death within 5 years of baseline [8].

3.4 Model Evaluation

Models were evaluated using 5-fold stratified cross-validation. Primary performance metrics included AUC-ROC, AUC-PR (precision-recall), sensitivity, specificity, and F1-score. Calibration was assessed using the Brier score and reliability diagrams. DeLong's test was applied to compare AUC-ROC values between the multi-source fusion model and single-source baselines, with statistical significance defined as $p < 0.05$ [6]. SHapley Additive exPlanations (SHAP) were computed to assess feature importance and model interpretability [12].

Table 1. Summary of Multi-Source Data Modalities (N = 12,400 T2DM Patients)

Data Source	Variables Included	Patients with Data (%)	Missingness Rate (%)
Electronic Health Records (EHR)	Demographics, Labs (HbA1c, eGFR, LDL), Medications, Comorbidities	100%	8.2%
Continuous Glucose Monitor (CGM)	Mean Glucose, TIR, SD, Glycemic Variability Index	74.3%	12.7%
Genomic / Polygenic Risk Scores	PRS-Nephropathy, PRS-CVD, PRS-Retinopathy	61.8%	4.1%
Patient-Reported Outcomes (PRO)	Diabetes Distress Scale, WHO-5 Well-Being Index	83.5%	9.6%

4. RESULTS

4.1 Cohort Characteristics

The final analytic cohort comprised 12,400 patients with T2DM, with a mean age of 58.4 ± 11.2 years and mean diabetes duration of 9.3 ± 6.7 years. Female patients constituted 51.3% of the cohort. Mean baseline HbA1c was 7.9 ± 1.6%, mean eGFR was 72.4 ± 22.1 mL/min/1.73m², and 28.4% had prevalent hypertension at enrollment [5]. Over the 5-year observation period, incident nephropathy was diagnosed in 18.7% (n=2,319), retinopathy in 14.3% (n=1,773), and MACE in 11.6% (n=1,438) of patients. These incidence rates are consistent with published epidemiological estimates for similar T2DM cohorts [7].

4.2 Predictive Model Performance

The multi-source fusion model demonstrated superior predictive performance for all three complication outcomes compared to single-source baselines. For nephropathy prediction, the fusion model achieved an AUC-ROC of 0.923 (95% CI: 0.914–0.931), significantly outperforming EHR-only (AUC: 0.812, $p < 0.001$), CGM-only (AUC: 0.798, $p < 0.001$), and genomic-only (AUC: 0.761, $p < 0.001$) models [3]. For retinopathy, the fusion model attained an AUC-ROC of 0.911 (95% CI: 0.902–0.920), compared to the best single-source baseline of 0.793 (EHR-only). For MACE prediction, the fusion model achieved an AUC-ROC of 0.897 (95% CI: 0.887–0.907) versus 0.804 for the best unimodal approach [11].

Sensitivity analyses confirmed that model performance was robust across subgroups stratified by age, sex, diabetes duration, and insulin use. The federated learning architecture preserved 96.4% of centralized model performance while maintaining strict data privacy compliance [4]. Calibration analyses demonstrated good model fit, with Brier scores of 0.112, 0.098, and 0.089 for nephropathy, retinopathy, and MACE outcomes, respectively [12].

Table 2. Predictive Model Performance Comparison Across Complications (* $p < 0.001$ vs. all single-source models; DeLong's test)

Model / Data Source	Nephropathy AUC-ROC	Retinopathy AUC-ROC	MACE AUC-ROC	Brier Score (Avg)
EHR Only (GBM)	0.812	0.793	0.804	0.189
CGM Only (Bi-LSTM)	0.798	0.741	0.762	0.203
Genomics Only (MLP)	0.761	0.728	0.779	0.214
PRO Only (MLP)	0.704	0.693	0.718	0.231
EHR + CGM Fusion	0.871	0.855	0.847	0.153
Multi-Source Fusion (All)	0.923*	0.911*	0.897*	0.100

4.3 Feature Importance and SHAP Analysis

SHAP analysis revealed distinct feature importance profiles across complication types. For nephropathy, the three most influential features were baseline eGFR (mean |SHAP| = 0.31), urine albumin-creatinine ratio (0.24), and PRS-Nephropathy score (0.19) [7]. Glycemic variability index from CGM contributed meaningfully to all three outcomes, with mean |SHAP| values of 0.17–0.22, confirming the incremental value of wearable data beyond conventional HbA1c measures [10]. Diabetes distress scores from the PRO modality were significant predictors of MACE (mean |SHAP| = 0.14), consistent with evidence linking psychosocial burden to cardiovascular risk in T2DM [9]. Hypertension comorbidity, metformin use, and LDL cholesterol levels were consistently important across all three outcomes.

Table 3. Top SHAP Feature Importance Values by Complication Outcome (Mean |SHAP| across 5-fold CV)

Feature	Data Source	Nephropathy |SHAP|	Retinopathy |SHAP|	MACE |SHAP|
Baseline eGFR	EHR (Lab)	0.312	0.198	0.221
Urine Albumin-Creatinine Ratio	EHR (Lab)	0.241	0.143	0.187
PRS-Nephropathy Score	Genomics	0.189	0.072	0.098
Glycemic Variability Index	CGM	0.174	0.221	0.173
HbA1c (Baseline)	EHR (Lab)	0.162	0.204	0.158
Time-in-Range (TIR 70–180)	CGM	0.147	0.199	0.143
Hypertension (Comorbidity)	EHR (Diagnosis)	0.138	0.118	0.201
LDL Cholesterol	EHR (Lab)	0.121	0.109	0.187
Diabetes Distress Scale Score	PRO	0.092	0.083	0.141
PRS-CVD Score	Genomics	0.078	0.061	0.163

5. DISCUSSION

This study demonstrates that multi-source healthcare data fusion substantially outperforms unimodal predictive models for T2DM chronic complication prediction, with AUC-ROC improvements of 11–18%

over single-source baselines across nephropathy, retinopathy, and MACE outcomes [3]. These findings carry meaningful clinical implications. The integration of CGM-derived glycemic variability metrics provided incremental predictive value beyond HbA1c, reinforcing evidence that intraday glucose fluctuations represent an independent risk factor for microvascular and macrovascular complications [10,12]. The contribution of polygenic risk scores further underscores the potential of precision medicine approaches in diabetic complication risk stratification [7,13].

The federated learning architecture is a particularly noteworthy methodological contribution. By enabling model training across three geographically distributed hospital networks without centralizing sensitive patient data, this approach directly addresses the privacy and regulatory barriers that have historically impeded multi-institutional healthcare AI development [4]. The retention of 96.4% of centralized model performance under the federated setting suggests that privacy-preserving training induces only marginal performance costs—a finding that supports the broader deployment of federated frameworks for clinical AI [11].

The SHAP interpretability analysis offers actionable clinical insights. The dominance of eGFR and urine albumin-creatinine ratio in nephropathy prediction aligns with established Kidney Disease: Improving Global Outcomes (KDIGO) guidelines, lending face validity to the model's feature utilization [8]. The significant contribution of diabetes distress scores to MACE prediction is clinically important and often overlooked in conventional risk models, suggesting that systematic integration of PROs into clinical workflows may identify high-risk patients missed by biomarker-only approaches [9].

Several limitations warrant acknowledgment. The retrospective design and reliance on structured clinical data introduces potential ascertainment bias, particularly for soft outcomes such as patient-reported complications [5]. CGM data were available in only 74.3% of patients, and the representativeness of the wearable-equipped subgroup may not fully generalize to broader T2DM populations, particularly in low-resource settings [6]. Genomic data were available in 61.8% of patients, and PRS derived from predominantly European-ancestry reference panels may underperform in ethnically diverse subgroups [13]. Future work should prioritize multi-ethnic genomic reference panels, prospective validation cohorts, and real-time clinical decision support system integration [2].

6. CONCLUSION

This investigation establishes that multi-source healthcare data fusion, combining EHR data, CGM wearable biosensor streams, polygenic risk scores, and patient-reported outcomes, yields substantially superior predictions of chronic diabetic complications compared to single-source approaches [1,3]. The proposed federated fusion architecture addresses key real-world deployment barriers including data privacy, institutional heterogeneity, and computational scale [4]. The clinical translation of such models holds the potential to transform T2DM management from reactive complication treatment to proactive, data-driven prevention, ultimately reducing the enormous human and economic burden of this global epidemic [5,9].

Future research should focus on prospective clinical validation of the fusion framework, integration with clinical decision support systems, and extension to additional complication endpoints including peripheral neuropathy and diabetic foot disease [8]. The development of equitable, multi-ethnic predictive models remains a critical priority to ensure that the benefits of multi-source data fusion are accessible across all patient populations [7,13].

REFERENCES

[1] Kengne, A. P., Patel, A., Marre, M., Travert, F., Lievre, M., Zoungas, S., ... & Chalmers, J. (2020). Contemporary model for cardiovascular risk prediction in people with type 2 diabetes. European Journal of Cardiovascular Prevention & Rehabilitation, 18(3), 393–398. https://doi.org/10.1177/1741826710393408

[2] Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. Nature Medicine, 28(1), 31–38. https://doi.org/10.1038/s41591-021-01614-0

[3] Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future — big data, machine learning, and clinical medicine. New England Journal of Medicine, 375(13), 1216–1219. https://doi.org/10.1056/NEJMp1606181

[4] Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 1–7. https://doi.org/10.1038/s41746-020-00323-1

[5] International Diabetes Federation. (2021). IDF Diabetes Atlas, 10th edition. Brussels, Belgium: International Diabetes Federation. https://www.diabetesatlas.org

[6] Ravaut, M., Harish, V., Sadeghi, H., Leung, K. K., Volkovs, M., Tonekaboni, S., ... & Bhat, M. (2021). Predicting adverse outcomes due to diabetes complications with machine learning using administrative health data. NPJ Digital Medicine, 4(1), 24. https://doi.org/10.1038/s41746-021-00394-8

[7] Tuttle, K. R., Bakris, G. L., Bilous, R. W., Chiang, J. L., de Boer, I. H., Goldstein-Fuchs, J., ... & Molitch, M. E. (2014). Diabetic kidney disease: a report from an ADA consensus conference. American Journal of Kidney Diseases, 64(4), 510–533. https://doi.org/10.1053/j.ajkd.2014.08.001

[8] Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., ... & Williams, R. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045. Diabetes Research and Clinical Practice, 157, 107843. https://doi.org/10.1016/j.diabres.2019.107843

[9] Einarson, T. R., Acs, A., Ludwig, C., & Panton, U. H. (2018). Prevalence of cardiovascular disease in type 2 diabetes: a systematic literature review of scientific evidence from across the world in 2007–2017. Cardiovascular Diabetology, 17(1), 83. https://doi.org/10.1186/s12933-018-0728-6

[10] Contreras, I., & Vehi, J. (2018). Artificial intelligence for diabetes management and decision support: literature review. Journal of Medical Internet Research, 20(5), e10775. https://doi.org/10.2196/10775

[11] Solares, J. R. A., Raiber, F. E. A., Tsanas, A., Mian, U., Dobson, R. J., Aylin, P., ... & Hellgardt, K. (2020). Deep learning for electronic health records: a comparative review of existing approaches and challenges. Journal of Biomedical Informatics, 101, 103337. https://doi.org/10.1016/j.jbi.2019.103337

[12] Albers, D. J., Levine, M., Gluckman, B., Ginsberg, H., Hripcsak, G., & Mamykina, L. (2017). Personalized glucose forecasting for type 2 diabetes using data assimilation. PLOS Computational Biology, 13(4), e1005232. https://doi.org/10.1371/journal.pcbi.1005232

[13] Fawcett, N. J., Hadjipantelis, P. Z., Heindrich, L., & Fraser, S. D. S. (2021). Polygenic risk scores for prediction of diabetic kidney disease in type 2 diabetes: a systematic review. Nephrology Dialysis Transplantation, 36(Supplement_1), gfab092. https://doi.org/10.1093/ndt/gfab092.062

[14] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.