Early Prediction of Diabetic Complications Using Multi-Modal Deep Learning

Comprehensive List of Diabetic Complications

Microvascular Complications
These affect small blood vessels and are unique to diabetes:

1. **Diabetic Retinopathy**: This is the damage to the **Blood vessels** in the retina and can lead to vision loss or blindness.
2. **Diabetic Nephropathy:** This is **Kidney** damage caused by diabetes and may lead to chronic kidney disease or end-stage renal failure
3. **Diabetic Neuropathy**: This is **Nerve damage**, typically in the extremities (hands, feet). Types include: Peripheral neuropathy (most common), Autonomic neuropathy (affecting organs), and Proximal and focal neuropathy.

Macrovascular Complications
These involve larger blood vessels and increase the risk of cardiovascular events:

4. **Cardiovascular Disease (CVD)**: Includes coronary artery disease, heart attacks, and stroke
5. **Peripheral Artery Disease (PAD)**: Narrowing of arteries in limbs and can lead to poor circulation, ulcers, and even amputations
6. **Cerebrovascular Disease**: Includes stroke and transient ischemic attacks (mini-strokes) due to atherosclerosis in cerebral vessels

Title: Predicting the Onset of Diabetes-Related Complications After a Diabetes Diagnosis with Machine Learning Algorithms

Diabetes Research and Clinical Practice (2023)
DOI: 10.1016/j.diabres.2023.110910

Title:
a) What was done:

This study aimed to predict the onset of nine specific diabetes-related complications: hypertension, renal failure, myocardial infarction, cardiovascular disease, retinopathy, congestive heart failure, cerebrovascular disease, peripheral vascular disease, and stroke, within 1, 2, and 3 years after a diabetes diagnosis. Researchers developed and evaluated machine learning (ML) models using a large administrative dataset from Catalonia, comprising 610,019 individuals diagnosed with diabetes. They applied four different ML algorithms (logistic regression, decision tree, random forest, and XGBoost) and also developed a stacked ensemble model to improve predictive accuracy and robustness across different complication types and time horizons.

b) Why it was done (Motivation)

Diabetes complications significantly impact patients' quality of life and increase mortality and healthcare costs. While prior studies focused on individual complications or used longer prediction windows, this study addresses a gap by developing short- and medium-term predictive models. The motivation was to enable earlier risk stratification and personalised prevention by leveraging ML on large-scale population health data. Accurate prediction models could help guide interventions, reduce progression to advanced disease stages, and allocate healthcare resources more effectively. This work contributes to filling a critical gap in real-world, time-sensitive risk modelling for newly diagnosed diabetic populations.

c) How it was done (Methodology)

The dataset combined primary care, hospital, and emergency care data from 2013–2017. Each patient record was preprocessed to generate over 2,800 diagnostic and 463 procedural features, with additional demographic variables. The authors engineered new variables like healthcare usage frequency, encoded categorical features via one-hot encoding, and excluded individuals already diagnosed with target complications. ML models (LR, DT, RF, XGB) were trained on 80% of the data and validated on 20%. Hyperparameters were optimised using grid search and 5-fold cross-validation. SHAP values were calculated to interpret the contribution of each feature to complication predictions.

d) What was achieved (Results)

Across all models and timeframes, predictive performance varied by complication. Retinopathy had the lowest AUC (~60%), while congestive heart failure and hypertension achieved AUCs up to ~69%. Random Forest performed best for hypertension, myocardial infarction, and retinopathy, while Logistic Regression excelled for the remaining complications. The stacked ensemble model generally improved predictive performance. Key

features influencing predictions included age, gender, socioeconomic status, and healthcare usage. Though moderate, the performance was consistent and interpretable, validating ML's utility in short-term complication risk prediction across a large, real-world population.

e) Key findings, conclusions, and limitations

This study demonstrates the feasibility of predicting diabetes-related complications in short timeframes using administrative data and ML techniques. It introduces a novel multi-year, multi-complication risk modelling framework with large-scale data. However, limitations include absence of clinical biomarkers (e.g., blood glucose), lack of data beyond three years, and potential dataset shift over time. Performance was moderate and varied by complication, but key predictors like age and co-payment level were consistent. The findings highlight ML's potential to inform proactive healthcare delivery, although clinical deployment would require integration with richer EHR data and validation in other healthcare systems.

Title: Using Machine Learning Techniques to Develop Risk Prediction Models for the Risk of Incident Diabetic Retinopathy Among Patients With Type 2 Diabetes Mellitus: A Cohort Study

https://doi.org/10.3389/fendo.2022.876559

a) What was done

This study developed and validated five machine learning models, including XGBoost, Random Forest, Logistic Regression, SVM, and KNN, to predict the future risk of diabetic retinopathy (DR) in patients with type 2 diabetes mellitus. The authors trained models using electronic health records of 7,943 patients hospitalized from 2010 to 2018. Performance was evaluated using ROC curves, AUC, accuracy, sensitivity, and specificity. The best-performing model, XGBoost, was capable of predicting DR risk across various follow-up periods up to 10 years. Subgroup analyses, nomogram development, and feature importance analysis were also performed to aid clinical applicability and interpretability.

b) Why it was done

Diabetic retinopathy (DR) is the leading cause of preventable blindness in working-age adults. Existing risk models often rely on cross-sectional data or short-term windows, limiting their clinical utility. Many patients develop DR without early symptoms, delaying diagnosis and intervention. This study addresses these gaps by building long-term, time-sensitive risk prediction models using routinely collected EHR data. By identifying high-risk individuals before clinical symptoms appear, healthcare providers can offer timely interventions, optimize screening intervals, and reduce vision-threatening complications. This research supports the development of proactive, personalized care strategies for DR management in type 2 diabetic populations.

c) How it was done (Methodology)

The cohort consisted of 7,943 adult patients with type 2 diabetes and no baseline DR, collected from Dalian Central Hospital's EHR system. Patients were split into training (n=5,559) and test (n=2,384) sets. Eighteen baseline features (clinical, demographic, and lab data) were used to train five ML models. GridSearchCV and fivefold cross-validation were used for hyperparameter tuning. Prediction performance was assessed across time points (1–10 years) and DR severity levels. XGBoost's feature importance and nomograms were analysed for interpretability. Supplementary analyses used multi-timepoint follow-up data to evaluate robustness. Statistical analysis was conducted using R 4.0.2.

d) What was achieved (Results)

The XGBoost model outperformed others with an AUC of 0.913, accuracy of 79.9%, sensitivity of 90.2%, and specificity of 77.1%. For certain time intervals (1–2 years), AUC reached 0.966. In subgroup analysis, true positive prediction exceeded 87%, peaking at 100% for 7–8 years. Top predictive features included HbA1c, duration of diabetes, fasting blood glucose, and age. Additionally, novel biomarkers like serum uric acid, LDL-C, total cholesterol, eGFR, and triglycerides were identified as strong predictors. Nomograms integrating ML outputs achieved a C-index of 0.921. On average, the XGBoost model predicted DR 2.9 years earlier than clinical diagnosis.

e) Key findings, conclusions, and limitations

The study confirms that ML, especially XGBoost, can accurately predict DR risk years ahead of clinical diagnosis. Feature analysis supports known risk factors and introduces novel ones. The model's long-range performance, interpretability (via nomograms and feature importance), and compatibility with EHR data highlight its potential as a clinical decision support tool. However, the single-centre retrospective design limits generalizability, and the exclusion of certain biomarkers (e.g., hip/neck circumference) due to missing data may reduce model richness. Future directions include multicentre validation and prospective trials. Overall, this study advances predictive modelling for DR, with practical implications for diabetic care pathways.

Title: Analysis for Warning Factors of Type 2 Diabetes Mellitus Complications with Markov Blanket Based on a Bayesian Network Model

https://doi.org/10.1016/j.cmpb.2019.105302

a) What was done

This study built a Bayesian Network (BN) model to identify warning factors for predicting six T2DM complications, DN, DR, DF, DMV, DPN, and DK, using physiological, urine test, and biochemical data. It applied the Markov Blanket (MB) to select minimal yet predictive variables accessible through routine medical exams. The model was trained on a hospital-based dataset of 1,485 patients, evaluated using 10-fold cross-validation. Two BN variants, one with expert priors and one without, were compared to traditional classifiers (Naïve Bayes, Random Forest, C5.0). Performance was assessed with AUC, sensitivity, and specificity, demonstrating superior outcomes using BN-based approaches.

b) Why it was done

T2DM complications severely impair quality of life and impose economic burdens. While prior models used complex clinical or expensive biomarkers, they are impractical for patient self-monitoring. This study sought to develop a practical, explainable, and effective model using accessible features from standard medical exams. The aim was to identify early warning signs that can help patients self-manage and prevent progression. Moreover, the BN structure reveals probabilistic and causal relationships between variables, supporting personalised interventions. The focus on easily obtainable warning factors aims to extend prediction tools to resource-limited settings and empower proactive patient care.

c) How it was done (Methodology)

Data from 1,485 hospital inpatients diagnosed with T2DM complications were pre-processed and imputed using PMM (predictive mean matching). Continuous variables were discretised, and Bayesian networks were built using a combination of expert knowledge and data-driven learning via Bootstrap and Tabu Search. Markov Blanket was used for feature selection. Two BN models, with and without priors, were trained and validated via 10-fold cross-validation. Models predicted each complication using only the selected warning factors. Their performance was compared against baseline models (Naïve Bayes, RF, C5.0) using AUC, sensitivity, specificity, and confidence intervals for robustness.

d) What was achieved (Results)

The BN model achieved strong predictive performance for DN (AUC: 0.831), DF (AUC: 0.905), DMV (AUC: 0.753), and DK (AUC: 0.877) using only warning factors. BN variants outperformed or matched traditional classifiers across most complications. The model identified specific variables (e.g., Cr, HDL, TRIG, Na, U-KET) that were strongly associated with each complication. The use of Markov Blanket allowed dimensionality reduction without compromising performance. The system also visualised probabilistic dependencies among complications, showing that the presence of one may increase risk of another (e.g., DMV affecting DN/DK). This provided clinically interpretable pathways for targeted interventions.

e) Key findings, conclusions, and limitations

The BN model using patient-accessible physiological indicators can effectively predict major T2DM complications. Its interpretability via graphical structure and Markov Blanket makes it suitable for clinical deployment and patient self-management. The study proves that fewer, simpler features can offer actionable insights with high predictive power. However, limitations include single-centre data, moderate sample size, and reliance on discrete-state modelling. Missing data imputation could affect generalizability. Future work should involve external validation, inclusion of temporal features, and refinement of imputation strategies. Nevertheless, the approach shows promise for accessible, low-cost complication prediction in real-world T2DM management.

Title: Machine Learning Models for Prediction of Diabetic Microvascular Complications

https://doi.org/10.1177/19322968231223726

a) What was done

This study presents a comprehensive review of 74 longitudinal studies that developed or validated machine learning (ML) models for predicting diabetic microvascular complications, specifically diabetic retinopathy (DR), diabetic kidney disease (DKD), and diabetic neuropathy (DN). It analysed 256 internally validated and 124 externally validated models based on design, predictor variables, ML methods, outcome definitions, and performance metrics. The review identifies trends, high-performing models, and areas requiring improvement in predictive modelling of these complications.

b) Why it was done

Diabetic microvascular complications significantly contribute to morbidity, but prediction models often focus narrowly on specific conditions or use limited ML techniques. There was a need to comprehensively evaluate ML/AI models that predict the onset, not just detection, of DR, DKD, and DN, using structured clinical data. The goal was to summarise existing approaches, highlight best practices, and identify areas where research is lagging, particularly for DR and DN. This helps guide future development of accurate, generalizable models that can aid early intervention and optimise healthcare resource allocation.

c) How it was done (Methodology)

A PubMed search from 1990 to July 2023 was conducted, focusing on ML models predicting DR, DKD, or DN in type 2 diabetes patients. Eligible studies included longitudinal cohorts and registries with reported predictive performance (e.g., c-statistic). The review excluded image-based detection and short-term hospitalisation outcomes. Predictor variables, model type, validation approach, and outcomes were extracted and analysed. Internal vs. external validation, prediction horizons, and a number of predictors were compared. Model types included XGBoost, Random Forest, Logistic Regression, Neural Networks, and others. Subgroup and statistical analyses were performed using STATA, Python, and R.

d) What was achieved (Results)

DKD models were the most common and best performing (c-statistic 0.81 internal, 0.74 external). DR and DN had fewer validated models and lower performance (DR: 0.74/0.71, DN: 0.71/0.67). XGBoost, Random Forest, and Logistic Regression consistently outperformed other techniques. Models with shorter prediction horizons and fewer predictors tended to perform better. Only 25% of models had true external validation. Common predictors included A1C, blood pressure, duration of diabetes, and eGFR. Diabetic kidney disease models benefited from richer clinical biomarkers and broader international datasets. Diabetic

neuropathy models were the weakest due to inconsistent definitions and fewer reliable predictors.

e) Key findings, conclusions, and limitations

ML models show promise in predicting DR, DKD, and DN, with DKD having the strongest results. Model performance depends heavily on the prediction horizon, type and number of predictors, and ML algorithm. XGBoost and Random Forest were top performers; Survival Analysis and Naïve Bayes underperformed. External validation was scarce, often mislabeled, and had limited generalizability. The review calls for standardisation in model development, better external validation, and incorporation of stable, EHR-friendly predictors. Future research should emphasise longitudinal datasets, diverse populations, and integration into clinical workflows. DN prediction remains underdeveloped due to a lack of robust biomarkers and standardised outcome definitions.

Title: Machine learning-based risk predictive models for diabetic kidney disease in type 2 diabetes mellitus patients: a systematic review and meta-analysis

https://doi.org/10.3389/fendo.2025.1495306

a) What was done

This paper systematically reviewed and meta-analysed 26 studies that used machine learning (ML) algorithms to predict the risk of diabetic kidney disease (DKD) in patients with type 2 diabetes mellitus (T2DM). Across the included studies, 94 different ML models were developed, using diverse datasets including structured EHR data, images, genetic, and metabolic biomarkers. The study assessed internal and external model validation and pooled the Area Under the Curve (AUC) for performance metrics. The review aimed to quantify the accuracy, reliability, and generalizability of these models while identifying methodological limitations and opportunities for improvement in DKD risk prediction research.

b) Why it was done

With DKD affecting 20–40% of T2DM patients and being a major contributor to end-stage renal disease globally, early prediction is essential for timely clinical intervention. Traditional regression models are limited in handling complex, nonlinear relationships present in diverse patient data. Machine learning offers advanced capabilities for risk stratification using large-scale, multi-source datasets. However, performance variability and lack of standardisation hinder clinical adoption. This review was conducted to assess the predictive effectiveness of ML approaches, address inconsistencies across studies, and provide a clearer direction for future research and clinical translation in DKD risk modelling using AI-based methods.

c) How it was done (Methodology)

Following PRISMA and CHARMS guidelines, a comprehensive search of PubMed, Embase, Cochrane Library, and Web of Science was conducted to identify English-language studies using ML to predict DKD risk in T2DM patients. Eligible studies were screened and evaluated using the PROBAST risk of bias tool. Key data including ML model type, input variables, sample size, and validation method were extracted. Meta-analysis was performed using DerSimonian-Laird random effects modelling to pool AUC scores, with subgroup analyses by ML type, study design, and validation method. Sensitivity analysis, heterogeneity testing (I^2), and forest plots were used to evaluate robustness.

d) What was achieved (Results)

The internal validation meta-analysis yielded a pooled AUC of 0.839, while the external validation pooled AUC was 0.830, indicating strong model performance. Deep learning models had the highest pooled AUC (0.863), followed by random forest (0.848). Despite high AUC values, only 8 out of 26 studies performed external validation, and most had a high risk of bias. The included studies varied widely in dataset size, features used, and validation rigor. Heterogeneity was significant across most comparisons. Nevertheless, ML models, especially RF and DL, demonstrated strong potential to support early detection and targeted management of DKD in clinical practice.

e) Key findings, conclusions, and limitations

ML models, especially deep learning and random forests, show excellent potential for predicting DKD risk in T2DM patients, offering superior performance to traditional methods. However, widespread clinical deployment is limited by inconsistent feature selection, limited external validation, varied data types, and high risk of bias. Most models relied on internal validation only, raising concerns about generalizability. The authors recommend multicenter studies, standardised reporting, external validation, and use of interpretable models. Integration of diverse data types, including imaging and genetics, alongside SHAP or other interpretability tools, is suggested for building robust, generalisable, and clinically useful DKD risk prediction systems.

Title: ***Early Prediction of Diabetic Complications Using Multi-Modal Deep Learning***

Summary: This research aims to develop a deep learning framework capable of predicting the early onset of diabetic complications, specifically retinopathy, nephropathy, neuropathy, and cardiovascular disease, using multi-modal data. The increasing burden of diabetes and its associated complications calls for advanced risk-stratification tools that support proactive, personalised interventions. Traditional statistical methods struggle to model the complex, non-linear relationships across diverse clinical indicators. In contrast, deep learning models can harness structured clinical data (e.g., labs, vitals, diagnoses), patient demographics, and derived features to build more accurate and generalizable risk predictors. This project will utilise multiple publicly available datasets, including the Mendeley Diabetic Complications Dataset, the UCI 130-US Hospital Encounters dataset, and Messidor DR-derived features. The pipeline will incorporate multi-label prediction capabilities, support explainability (via SHAP), and emphasise clinical interpretability. The final goal is to build a robust, interpretable, and scalable predictive system that could be integrated into electronic health record workflows for real-world deployment.

Key Project Objectives

1. To build predictive models for common diabetic complications (e.g., retinopathy, nephropathy) using structured clinical data.
2. To integrate multi-modal inputs (e.g., demographics, lab tests, diagnosis codes) into a unified prediction pipeline.
3. To implement explainability tools (e.g., SHAP, LIME) to ensure model transparency and clinical interpretability.
4. Validate model generalizability using cross-dataset evaluation and robustness testing

Datasets Description

Main dataset:

- **Title**: Mendeley Micro & Macro Complications Dataset
- **Link**: https://data.mendeley.com/datasets/dsjcb6pyd8/1

- **Content**: This dataset consists of 3,068 patients with Type II diabetes and includes structured clinical and lifestyle features such as age, BMI, blood pressure, HbA1c, and medication usage. It provides direct binary labels for several complications, including nephropathy (NEP), neuropathy (NEU), retinopathy (RET), cardiovascular disease (CV), and peripheral vascular complications (PER VAS). The dataset is clean, well-structured, and ready for binary or multi-label classification tasks. Its simplicity and labeled outcomes make it ideal for initial experimentation, explainability techniques, and fast prototyping of predictive models.

Supplementary dataset:

- **Title**: *UCI Diabetic Dataset*
- **Link**: https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008

- **Content**: This dataset contains over 100,000 patient encounters collected from 130 US hospitals between 1999 and 2008. It includes rich tabular data such as demographics, admission details, diagnoses, lab procedures, and diabetes-related medications. Importantly, ICD-9 diagnosis codes allow for the derivation of diabetic complications such as retinopathy, nephropathy, and cardiovascular issues. The data structure supports timeline-based modelling, enabling early prediction of complications based on prior visits. It is highly suitable for supervised learning tasks and for simulating longitudinal clinical progression in diabetic patients.

- **Title**: *Messidor DR Features (Debrecen)*
- **Link**: https://archive.ics.uci.edu/dataset/329/diabetic+retinopathy+debrecen
- **Content**: This dataset includes 1,151 rows of tabular features extracted from the Messidor retinal image set. Features represent lesion counts (e.g., microaneurysms, exudates), anatomical descriptors (e.g., macula-disc distance), and binary indicators of image quality. The target variable indicates whether diabetic retinopathy (DR) is present. Although it does not include raw images, it serves as a structured, image-derived representation ideal for binary classification. It is particularly useful for building lightweight diagnostic models or integrating image-derived features into a multimodal prediction system.

Datasets and Descriptions

Dataset Name	Source	Description	Link
Mendeley Micro & Macro Complications Dataset	Mendeley Data (2021)	3,068 patients with structured features (age, BP, HbA1c, etc.) and labeled complication flags: retinopathy (RET), nephropathy (NEP), neuropathy (NEU), cardiovascular (CV), peripheral vascular (PER VAS).	GO
UCI Diabetic 130-US Hospitals Dataset	UCI Machine Learning Repository	101,766 patient encounters from 130 US hospitals (1999–2008). Contains demographics, medications, and ICD-9 codes to derive diabetic complications.	GO
Messidor DR Features (Debrecen)	UCI Machine Learning Repository	Tabular features extracted from fundus images (Messidor) for diabetic retinopathy detection (e.g., microaneurysms, exudates).	GO