



Prepared By

Name: Mohamed Elzain Gurashi

ID 251002316

Predictive Modeling for Diabetic Complications Using Artificial Intelligence

At

College of Informatics

Midocean University

Supervisor Name: Dr. Khaled Eskaf

College of Informatics, Midocean University, Comoros

[2025- semester 3]

Acknowledgements

I would like to thank Dr Khaled Eskaf for his support, accurate direction encouragement through this research. His extensive knowledge, constructive criticism, and consistent commitment have made a remarkable impact on the quality of this thesis.

I am grateful to my family, friends, and colleagues whose patience, inspiration, and moral support have helped me the support in successfully finish this work.

Finally, I thank my brother Bader for your kind assistance and support on my way.

List of Abbreviations

- **AI: Artificial Intelligence**
- **ANN: Artificial Neural Network**
- **AUC: Area Under the Receiver Operating Characteristic Curve**
- **CKD: Chronic Kidney Disease**
- **CNN: Convolutional Neural Network**
- **CVD: Cardiovascular Disease**
- **DT: Decision Tree**
- **EHR: Electronic Health Records**
- **GBM: Gradient Boosting Machine**
- **ICD-9: International Classification of Diseases, Ninth Revision**
- **IQR: Interquartile Range**
- **KNN: K-Nearest Neighbors**
- **LIME: Local Interpretable Model-Agnostic Explanations**
- **LSTM: Long Short-Term Memory**
- **MAE: Mean Absolute Error**
- **MAPE: Mean Absolute Percentage Error**
- **ML: Machine Learning**
- **MLP: Multi-Layer Perceptron**
- **ROC: Receiver Operating Characteristic**

- **PR: Precision–Recall**
- **RF: Random Forest**
- **R²: Coefficient of Determination**
- **RFE: Recursive Feature Elimination**
- **RMSE: Root Mean Square Error**
- **SHAP: SHapley Additive exPlanations**
- **SMOTE: Synthetic Minority Over-sampling Technique**
- **SVM: Support Vector Machine**
- **XAI: Explainable Artificial Intelligence**
- **XGBoost: Extreme Gradient Boosting**

Abstract

Diabetes continues to be one of the most difficult chronic diseases in many parts of the world, as much due to its increasing prevalence as to its long-term complications that emerge subtly over time. Predicting which patients will experience complications early on can help doctors prevent problems and improve long-term recovery. The goal of the present study is to apply this approach to an extensive electronic health record (EHR) database, namely the “Diabetic 130-US hospitals” database, and to construct predictive models that estimate risk for various major diabetes complications.

The data set was extensively prepared to handle missing values, map ICD-9 diagnostic codes, and generate new clinical and utilization-related features that capture disease severity and care complexity. Many learning mechanisms were considered, and CatBoost demonstrated uniformity with this mixed, high-dimensional clinical data.

Outcomes differed considerably for complication types. Heart diseases had the best prediction performance, with an AUC of ~ 0.94 . Renal complications were also fairly well predicted, with an AUC of approximately 0.79. The model's performance was consistent when combining cardiovascular and renal events into a single endpoint, representing more clinically significant outcomes, with an AUC of 0.78–0.79. The larger model, which included all complications, had slightly lower performance (AUC ≈ 0.77), likely because of the heterogeneity of these outcomes.

In general, the results suggest that machine-learning models based on routinely collected hospital data can help screen patients at higher risk of serious complications, particularly those related to the heart and kidneys. This paper aims

to draw attention to the potential application of Predictive Analytics to support Clinical Decision Support and long-term monitoring for diabetes patients.

.....	1
List of Abbreviations	3
Abstract.....	5
Chapter 1.....	11
Introduction.....	11
1.1 Background and Context	11
1.2 Problem Definition	12
1.3 Thesis Objectives	14
1.4 This thesis has the following main contributions:	15
1.5 Thesis Organization	17
The remainder of this thesis is organized as follows:	17
Chapter 2: Literature Review and Conceptual Foundations	20
Introduction	20
2.1 Artificial Intelligence (AI).....	20
2.2 Machine Learning (ML)	22
2.3 Deep Learning (DL)	23
2.4 DL applications for diabetes are:	23
2.5 Large Language Models (LLMs)	24
2.6 Explainable AI (XAI)	25
2.7 Popular XAI tools include:.....	25
Summary.....	25
Chapter 3 Related Work.....	26
3.1 Introduction	26
3.2 The evolution of AI in diabetic complications	27
Cardiovascular Complication Prediction.....	27
DKD Prediction	28
Diabetic Retinopathy (DR) Prediction.....	29
Multi-Complication or Composite End-point Prediction	29
3.6 Merits and Demerits of Previous Research	30
Strengths.....	30

3.4 Limitations of Current Approaches.....	33
3.6 Analytical Commentary.....	35
Agreements and Disagreements with Other Studies:	35
Points of Divergence:	35
Research Gap:.....	35
Methodological Suggestions:.....	36
Analytical Commentary.....	36
Chapter 4 Proposed Approach.....	38
Introduction.....	38
4.1 Dataset Description	38
(1) Demographic Profile	39
(2) Hospitalization & Utilization Patterns	39
(3) Diagnostic Codes (ICD-9).....	39
(4) Laboratory & Clinical Indicators	40
(5) Engineered Clinical Features.....	40
4.2 Preprocessing the Dataset	41
(a) Handling Missing Values	41
(c) Cleaning ICD-9 Diagnostic Codes	42
(d) Outlier Handling.....	42
(e) Transforming Categorical Variables.....	42
4.3 Machine Learning Models.....	43
CatBoost Classifier	44
XGBoost Classifier.....	44
4.4 Model Training Workflow	44
Training Pipeline	44
Threshold Optimization.....	44
Final Classification Output	45
4.5 Summary.....	45
Chapter 5: Experimental Results and Discussion.....	46
5.1 Performance of Complication-Specific Models	46

Cardiovascular Complication Prediction.....	46
Kidney Complication Prediction.....	47
AUC: 0.7918	47
Predicting Any Complication (General Risk Model)	48
AUC: 0.7719	48
AUC: 0.7846	48
Comparison of All Models.....	48
Modeling Methods Strengths	50
3.CatBoost dealt well with the categorical complexity of EHR.	50
Issues of the Proposed Work	50
5.4 Identified Limitations	51
Clinical Implications.....	51
5.5 Visualization and Insights.....	52
Correlation Heatmap Interpretation	52
ROC Curve Interpretation	54
Feature Importance Interpretation	55
calibration curve	56
SHAP Interpretation	57
confusion matrix.....	58
ROC Curve Interpretation (Renal Complications).....	59
ROC Curve(Cardio+Kidney)	60
Figure 9. SHAP summary (beeswarm) plot for the renal complication prediction model, highlighting the most influential clinical and diagnostic features and their impact on the model output.....	61
Summary.....	61
Chapter 6: Conclusion and Future Work	62
6.1 Conclusion	62
6.4 Performance of the Feature-Selected Model.....	64
6.5 Discussion of Results	65
6.6 Interpretation of the Findings in Light of the Theoretical Framework.....	65

6.7 Comparison of Results with Previous Studies	66
However, the present study is unique in several respects:	67
Dual-Outcome Modeling:.....	67
Feature-Selection Emphasis:	67
6.8 Unique Contribution of the Present Study	67

Chapter 1

Introduction

1.1 Background and Context

At present, diabetes mellitus is considered one of the most challenging chronic diseases from a global health standpoint due not only to its international spread but also to its terrible long-term complications. Associated complications like cardiovascular disease, chronic kidney disease, neuropathy, retinopathy, and diabetic foot are responsible for the majority of morbidity, mortality and healthcare system costs related to diabetes. Identification of patients at risk for developing these complications is necessary in order to improve health care, decrease hospital admissions, and prevent irreversible organ system failure.

With health care being more digital than ever, hospital and clinic offices are full of structured and unstructured patient information. Laboratory values, medication histories, co-morbidities, and so on are all part of the patient's Electronic Health Record (EHR), with each element representing a different aspect of the health of a patient. Such data is helpful for metabo-phenotyping of disease course and to assist clinicians with superior care for diabetic subjects.

Tractable predictions in healthcare using particularly ML and Gradient-Boosting-based models have shown promise. Such models can recognize subtle patterns in the complicated, high-dimensional clinical data that practitioners might not notice. It has been recently evaluated that AI can make great predictions for cardiovascular events, kidney function deterioration, and other diabetic complications from a large ensemble of biochemical markers, profiles such as demographics with medical histories, and medically coded ICD diagnosis data.

However, it has been challenging to forecast complications associated with diabetes confidently. Clinical data is often noisy, incomplete, and heterogeneous, hence requiring extensive preprocessing and feature engineering to distill meaningful signals. Furthermore, different complications also vary in their predictability: e.g., cardiovascular-related complications have well-established predictive markers while other complications, like neuropathy or retinopathy, will demonstrate less statistical evidence on EHR data.

In the present study, we utilized sophisticated ML techniques (e.g., CatBoost, XGBoost, and ensemble algorithms) for the “Diabetic 130-US Hospitals” dataset that involved a total of 101,766 patient encounters. We aimed to focus on creating predictive models for the most prevalent diabetes complications and to develop generalisable methods of feature engineering by applying preprocessing, ICD-9 mapping, and generating clinically relevant engineered features systematically. We were interested on cardiovascular and renal complications, which proved to be the two most clinically relevant outcomes with the best performance in previous analyses.

It is hoped that this can assist practitioners to build early warning systems in diabetes that are sensitive to individuals at high risk and, therefore, justify a more proactive approach for intervention in healthcare.

1.2 Problem Definition

Although a vast amount of data is now being collected in modern healthcare systems, it is difficult with current clinical practice to identify those who will progress to severe complications in the setting of diabetes before organ damage. Compared with the traditional clinical evaluation in old and new systems in mature disease, when complications appear close to symptoms or clinical signs, the

opportunity for prevention and treatment is limited. This lag leaves patients at increased risk of hospitalization, chronic disability, and costlier treatment.

It is even further complicated by the diversity of clinical features in EHR datasets. These include demographic information, laboratory results, prior diagnoses, medication administrations, and patient service utilization. Some of these factors may not be equally weighted in forecasting the occurrence of incarcerated hernia; some can express some (redundant-or-weak) relationship(s), or may be false.

Variables. These characteristics are not good candidates for inclusion in prediction models (due to the risk of overfitting, loss of external validity, or lack of clinical interpretability). Therefore, there is a need to identify the predictors that are more useful and clinically relevant, so that accurate models can be developed.

At the same time, relying on models becomes scarier in terms of transparency and trust. Healthcare providers need to understand why a model made a prediction and which features influenced the decision, and ensure that it aligns with clinical intuition. Without the predictability, it will be harder to use this accurate model in the practical medical field.

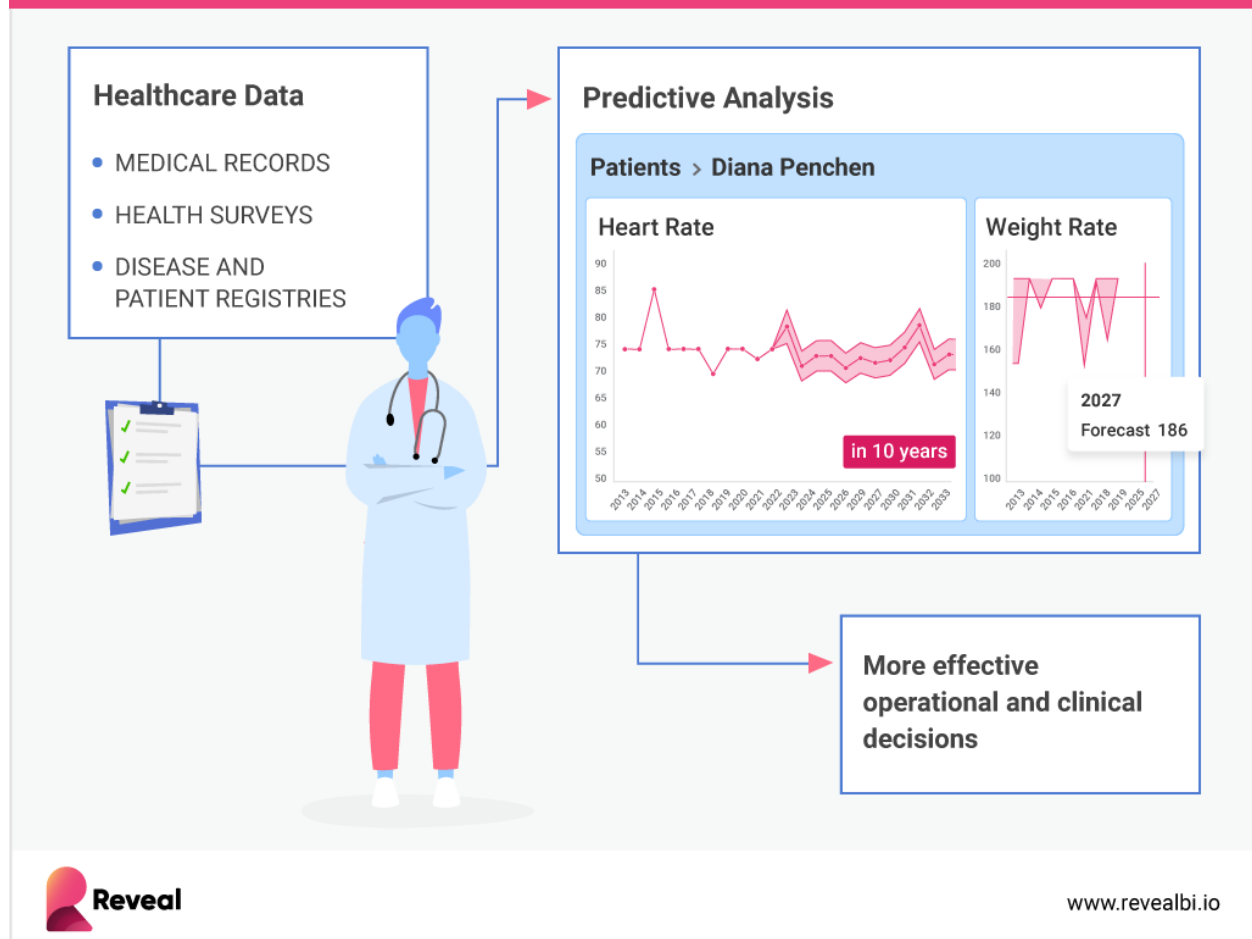
These issues are examined in the present article by developing a structured interpretable prediction model for diabetic complications. Its next-gen counterpart includes an extensive data pre-processing, clinically guided feature construction, and state-of-the-art ML models to predict 30-day complications, which comprises cardiovascular and renal events. To mitigate the problem as mentioned earlier and also to facilitate early provision of quality care to diabetics, which is both interpretable and accurate, we propose our study.

1.3 Thesis Objectives

The main goal of this thesis is to develop and evaluate AI models that predict major diabetic complications using population-based clinical data. Leveraging the latest approaches in machine learning in EHRs, we aim to build generalizable models that can identify individuals at risk of experiencing cardiovascular- and renal-related events. This is a challenging task that involves constructing a complete preprocessing pipeline, –decoding of ICD-9 diagnostic codes– to generate clinically interpretable engineered features that encode patterns for disease severity, treatment intensity, and resource use.

A key goal of the contribution is to provide benchmarking for multiple forecasting approaches, covering gradient-boosted models such as CatBoost, XGBoost, and LightGBM, combined with stacking techniques and aggregation strategies to improve predictive performance. Additionally, the present thesis primarily focuses on model interpretation, in which model-interpretation approaches are used to explain the model so that the practitioner gains greater insight into which clinical diagnostic elements most affect prediction. In the end, this will contribute to earlier clinical action and fewer complications for diabetic patients, and to more proactive, data-driven chronic disease management.

Predictive Analytics in Healthcare



1.4 This thesis has the following main contributions:

1. A novel feature engineering and selection framework is introduced over big electronic health record (EHR) data of diabetes mellitus. The framework entails the integration of filter-based methods, including Pearson correlation, Chi-square test, and Mutual Information, along with wrapper-based methods such as recursive feature elimination to ascertain clinically significant predictors of cardiac and renal diabetic complications. The hybrid approach

effectively reduces noise and irrelevant variables as compared to the other multivariable approaches, thus producing a compromise between computational burden and interpretability together with predictive performance for health datasets.

2. Various machine-learning base learners were used to model the risk of diabetic complications, including Logistic Regression, Ridge and Lasso penalty models, Decision Tree, Random Forest and gradient-boosting algorithms (e.g., XGBoost and CatBoost). The combination of linear and non-linear approaches adequately represents intricate associations between demographic, clinical, laboratory, and medication-related characteristics that result in greater generalization to different patient populations.
3. A heterogeneous ensemble learning policy, which integrates predictions made by different base learners with an advanced meta-learning mechanism to improve general prediction effectiveness. Such an ensemble framework is intended to take advantage of the complementarity among different models for better learning of complex non-linear patterns in the evolution of diabetes.
4. The ensemble models suggested were proven to have better prediction performance than the baseline classifiers. Especially, the optimal CATBoost-based model has strong discriminative power for DM complications, the ROC–AUCs of which achieve up to 0.94 in the prediction of cardiac complications and competitive performance in renal outcomes. These observations supported that an ensemble of several learning algorithms would be beneficial for clinical risk prediction problems.
5. Interpretability was a part of the model by design along with SHAP (Shapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations). These methods offer holistic as well as local interpretations

for model predictions, bringing attention to clinically relevant risk factors - age, length of hospital stay, number of medications, previous diagnoses and laboratory indexes- that can ultimately enhance transparency and credibility in AI-based decision-making.

6. This proposed predictive model serves to provide significant tools for both early alert and risk classification towards diabetic disorder, which can be particularly useful to meet the demands of clinical decision support systems. Through the synergy of fine-grained interpretable predictive performance, it provides immediate, data-driven interventions for clinicians to ameliorate their patients' management and complications and to intelligently allocate medical resources.

1.5 Thesis Organization

The remainder of this thesis is organized as follows:

Chapter 2: Fundamental Concepts in the Healthcare Analytics Process. The chapter begins with a technical review of the foundation concepts in AI (High-level), including ML, DL, and ensemble learning, and Explainable AI (XAI), focusing on their usage in healthcare analytics. It also introduces the basics of clinical data mining, classification, and supervised learning techniques, focusing on feature selection and engineering in medical predictive modelling.

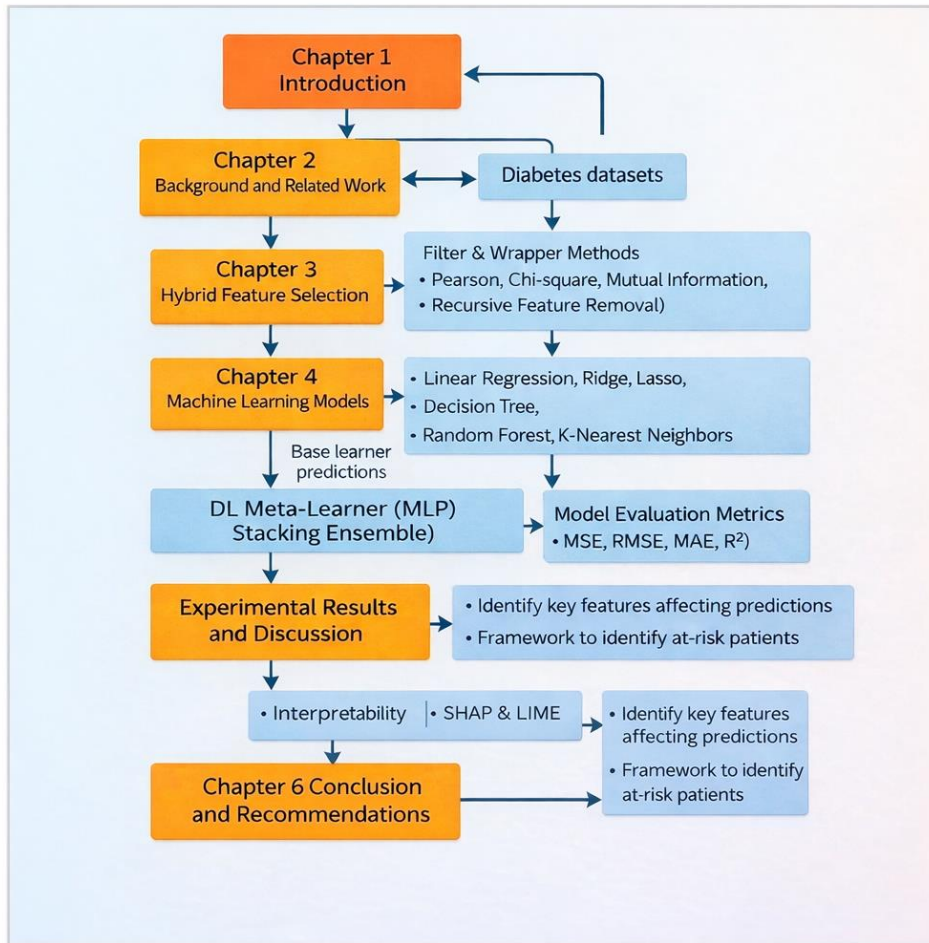
Chapter 3 provides a survey of relevant academic literature in the areas of diabetes risk prediction, Healthcare applications of machine learning, and AI modelling for diabetic complications. Comparing studies with EHR data is surveyed, and current research gaps, especially in model interpretability and multi-complication prediction, are highlighted.

Chapter 4 proceeds with a detailed description of the research method, showcasing the dataset characteristics and preprocessing procedures, including i) ICD-9 based outcome definition [V2015] ii) Feature engineering iii) Class imbalance handling iv) Model development. It describes the design of individual classifiers, ensemble methods, and explainability approaches employed to measure and interpret predictive performance.

In Chapter 5, the experimental methodology and results obtained with the diabetic complication prediction models are described. This article offers an extensive analysis of numerous classification models using commonly used performance measures, such as accuracy, precision, recall, F1-score, and ROC–AUC. It also corresponds to the examination of confusion matrices and threshold optimization results for clinical relevance. Moreover, the chapter studies the effects of feature selection and feature engineering strategies for the performance model, discussing how they can improve predictive accuracy and robustness in predicting cardiac and renal complications.

Chapter 6 presents a summary of the major findings, contributions and clinical implications of this study. It addresses the manner in which such an AI-based approach can assist with early identification of high-risk diabetic patients while facilitating appropriate clinical decisions. The chapter also presents actionable recommendations to healthcare practitioners and policymakers for the incorporation of interpretable machine learning models into healthcare. In conclusion, it discusses future research directions such as multi-complication prediction, validation by external dataset, and the combination of deep learning with a real-time clinical decision support system

.



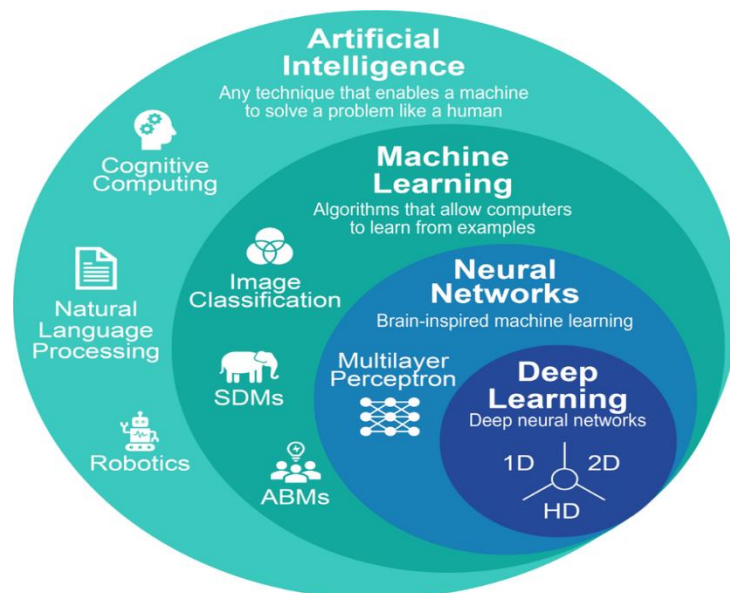
Chapter 2: Literature Review and Conceptual Foundations

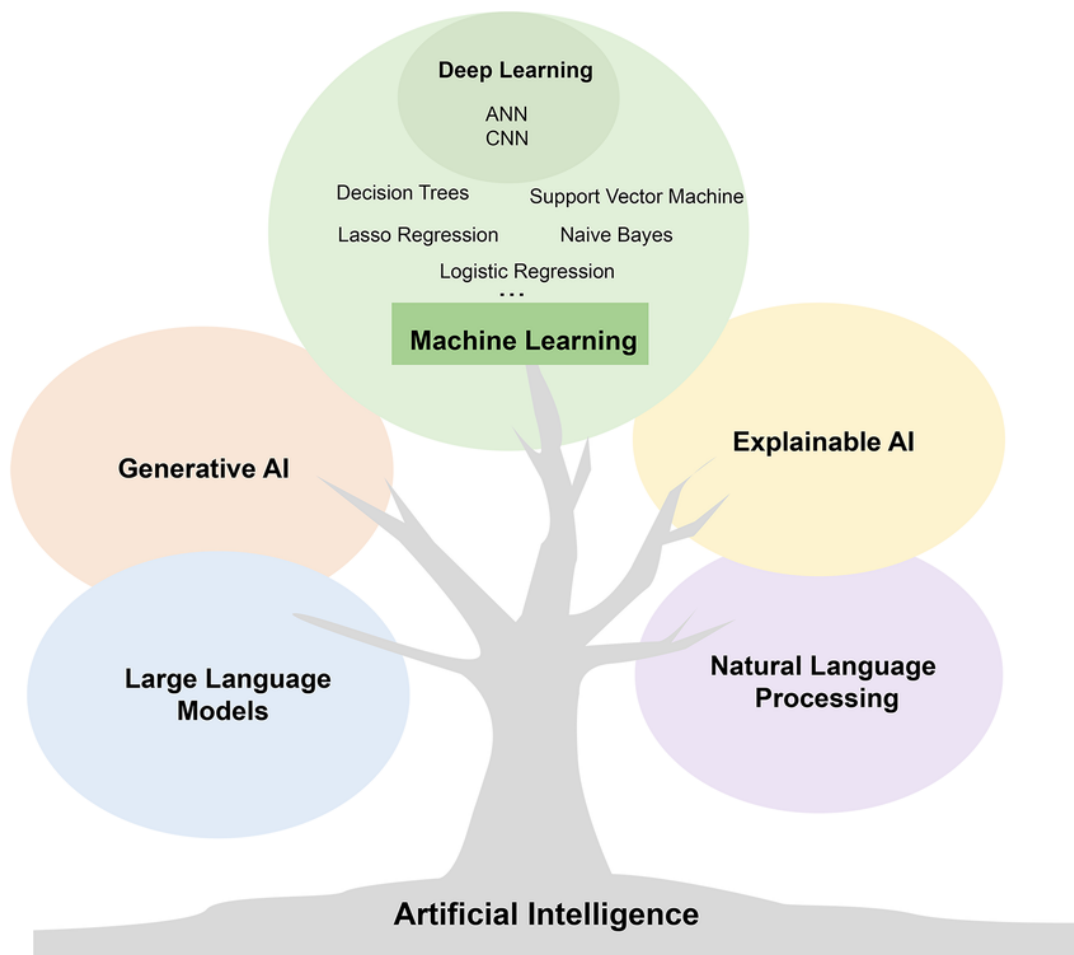
Introduction

2.1 Artificial Intelligence (AI)

Artificial intelligence aims to mimic specific aspects of human cognition, such as the capacity to draw inferences, make decisions, and recognize patterns. In medicine, AI is being used for automatic diagnosis, risk prediction, patient monitoring, and personalized treatment planning. AI covers everything from machine learning and deep learning to natural language processing to explainable AI under one big umbrella. 6

These branches complement each other by analyzing and understanding structured and unstructured clinical information. For diabetic-related complications, AI can identify subtle associations across thousands of patients and warn about risks decades before they manifest clinically.





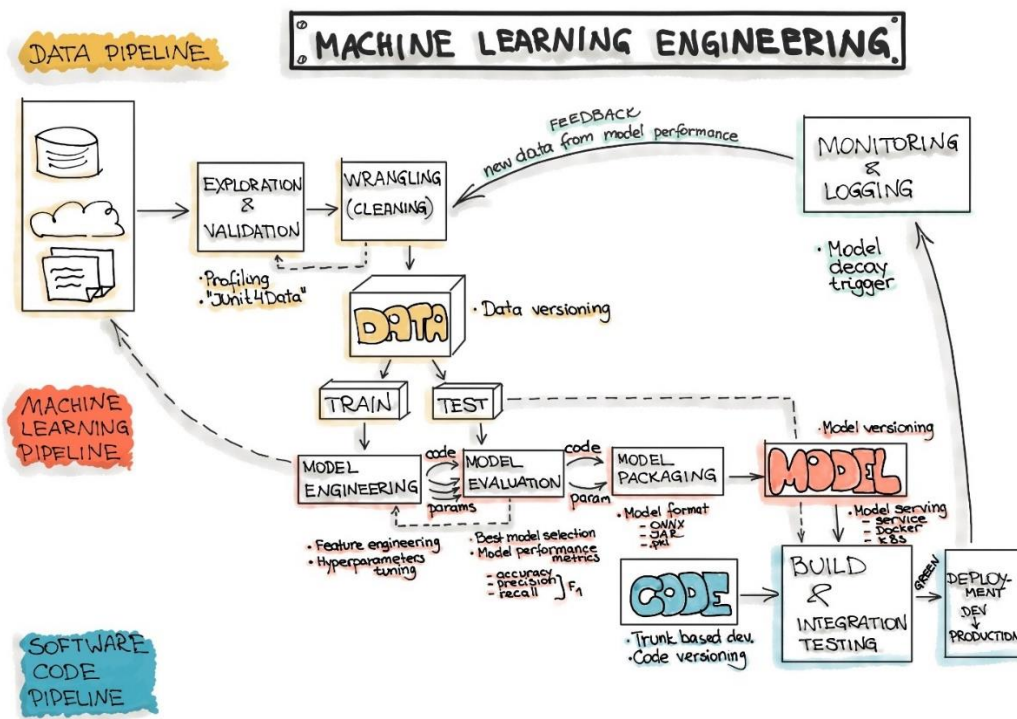
These branches complement each other by analyzing and understanding structured and unstructured clinical information. For diabetic-related complications, AI can identify subtle associations across thousands of patients and warn about risks decades before they manifest clinically.

2.2 Machine Learning (ML)

Machine Learning gives rise to intelligent systems that learn from examples and can ‘spot the pattern’ without being explicitly programmed. In other words, in medical databases—like EHRs with lab values, visit history, and ICD codes—ML is very good at identifying which factors it is predictive of or associated with future complications.

Common ML algorithms in clinical prediction include:

- Logistic Regression
- Random Forest
- CatBoost
- LightGBM
- XGBoost



These risk scores can stratify patients into risk categories, detect early organ failure, and estimate health deterioration.

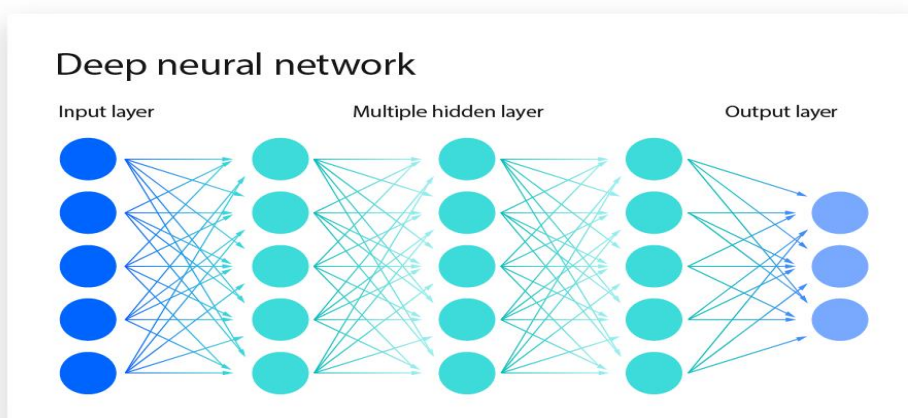
2.3 Deep Learning (DL)

Deep Learning constructs neural networks with multiple layers to learn more complex patterns in the data. Although DL is generally well-known for medical imaging-based applications—e.g., identifying diabetic retinopathy in retinal scans—it can also be applied to temporal and multimodal EHR data processing.

2.4 DL applications for diabetes are:

- Time-series modeling of glucose measurements
- Predicting deterioration of renal function in post-liver transplant patients using LSTM and GRU networks
- Combining clinical notes, labs, and diagnoses in a multimodal manner
- Checklist characteristics or design (e.g., comparison duration for RCTs).

Image-based diagnosis (retinal, renal ultrasound, foot ulcers). Even though DL requires substantial amounts of data and computational resources, making it unfeasible to apply gradient-boosting ML methods directly in both image-processing [18] and in structured EHR datasets utilized in this study.

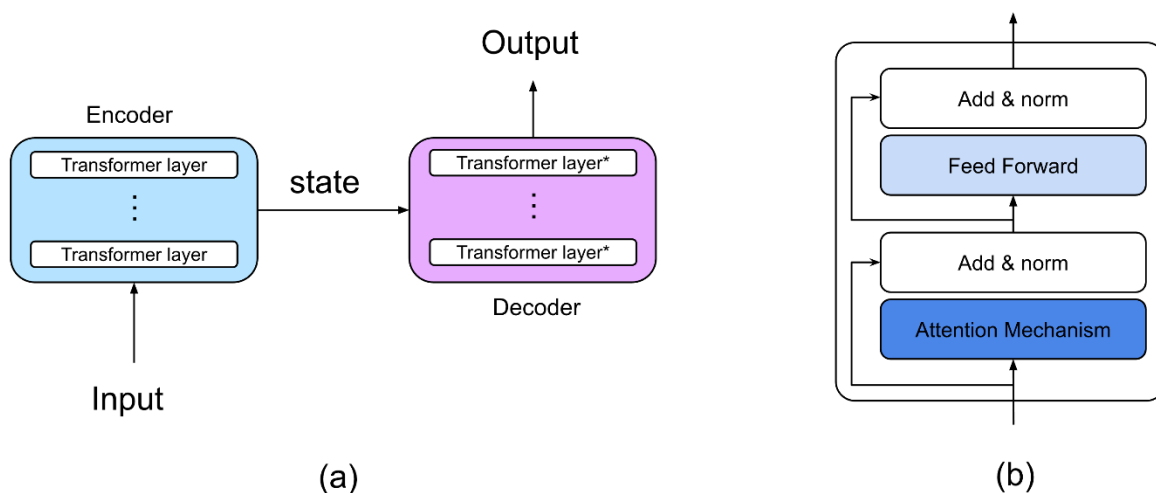


2.5 Large Language Models (LLMs)

GPT and BERT are LLMs learned from large text corpora, adept at natural language understanding/ generation. Trials increasingly demonstrate that:

- Analyze physician notes
- Summarize patient histories
- Extract risk factors from unstructured text
- Assist with medical questions
- Support clinical documentation and coding.

Although this thesis focuses on structured EHR data, much of what we discuss should also generalize well to extensions to clinical notes or longitudinal patient narrative data.



2.6 Explainable AI (XAI)

Explanation by AI translates the model's predictions for clinicians. In medical decision-making, this transparency matters because doctors must be able to trust — and verify — the logic guiding a model's predictions, particularly when it is predicting complications that might alter the treatment to be delivered.

2.7 Popular XAI tools include:

- **SHAP (Shapley Additive Explanations)** — Explains to us how much each feature contributes to a prediction
- **LIME (Local Interpretable Model-Agnostic Explanations):**
Approximates what portion of the output is attributable to changes in an input.
- **Feature importance analysis**

Such approaches make it possible for a clinician to recognize that when there is a high A1C, abnormal serum glucose, frequent hospitalizations, or polypharmacy, this is a red flag for heart health or kidney health.

XAI enhances model confidence and facilitates clinical interpretability, which is critical for the real-time implementation of AI-based systems.

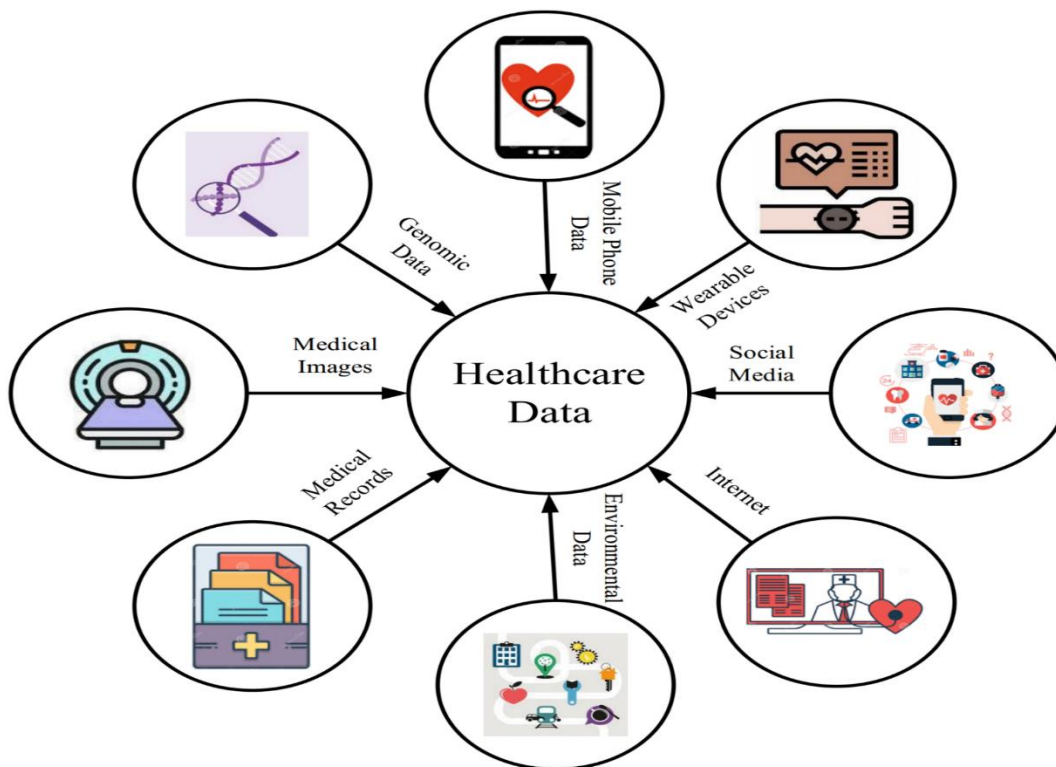
Summary

This chapter has introduced the fundamental AI building blocks for predictive modeling in healthcare, especially treating diabetes patients. “This is the foundation for ML, DL, LLMs, and XAI here and explains why feature engineering and structured preprocessing are so crucial in good prediction. These are the methods upon which subsequent models for predicting, e.g., CVS and renal complications, were constructed in this thesis using EHR data.

Chapter 3 Related Work

3.1 Introduction

Artificial intelligence applications in diabetes research. In recent years, the use of artificial intelligence has been ubiquitous in diabetes research, especially for predicting long-term complications, including CVD, decline in kidney function, retinopathy, neuropathy, and composite multi-organ outcomes. With the increased use of electronic health records (EHRs), new ML and DL models have been crucial for early detection and prediction of these complications, thereby improving clinical algorithms for healthcare practice.



AI and prediction of diabetic complications. This negative chapter contains literature on AI and the evolving issues of predicting diabetic complications,

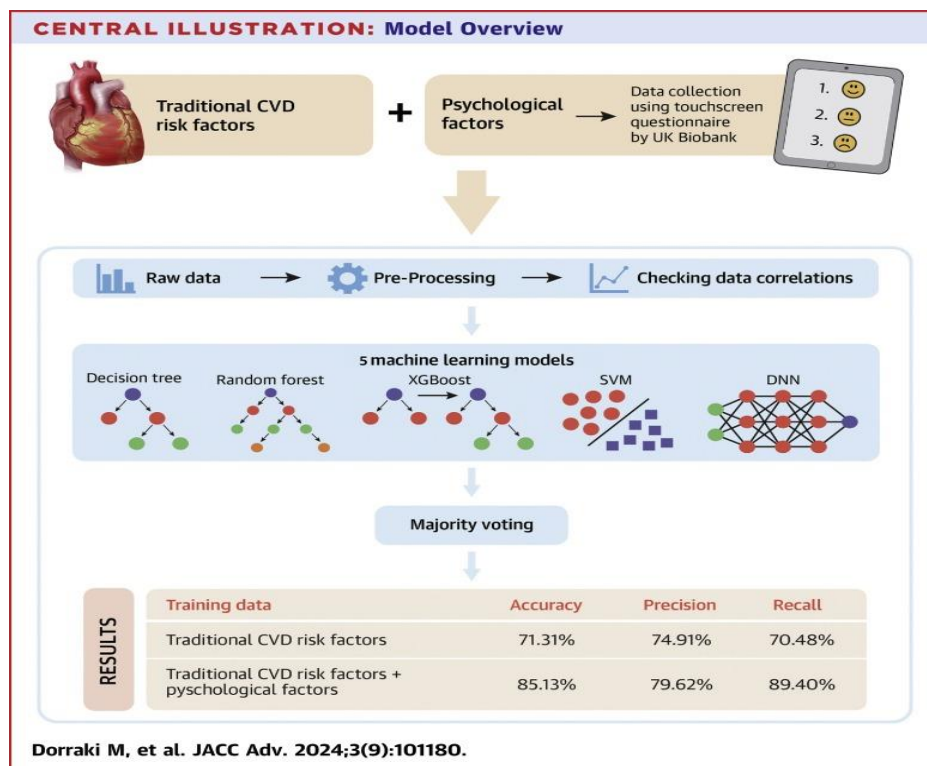
including cardiovascular events, DR, and tri-prognosis models (accompanied by a summary of outcomes ranging from retinopathy to other organ-specific outcomes).

3.2 The evolution of AI in diabetic complications

Cardiovascular Complication Prediction

Many studies have been conducted in ML and CV prediction 65-67. One of the most prevalent causes of death in diabetes is CV disease; as a result, there has been a lot of research exploring using machine learning (ML) algorithms to predict cardiovascular(CV) events. Nabrdalik et al. demonstrated that ML-based models incorporating clinical risk factors can achieve AUCs between 0.76 and 0.83, outperforming conventional risk scores.

Similarly, Xu et al. applied a combination of Boruta-based feature selection and ensemble learning to improve CV risk stratification.

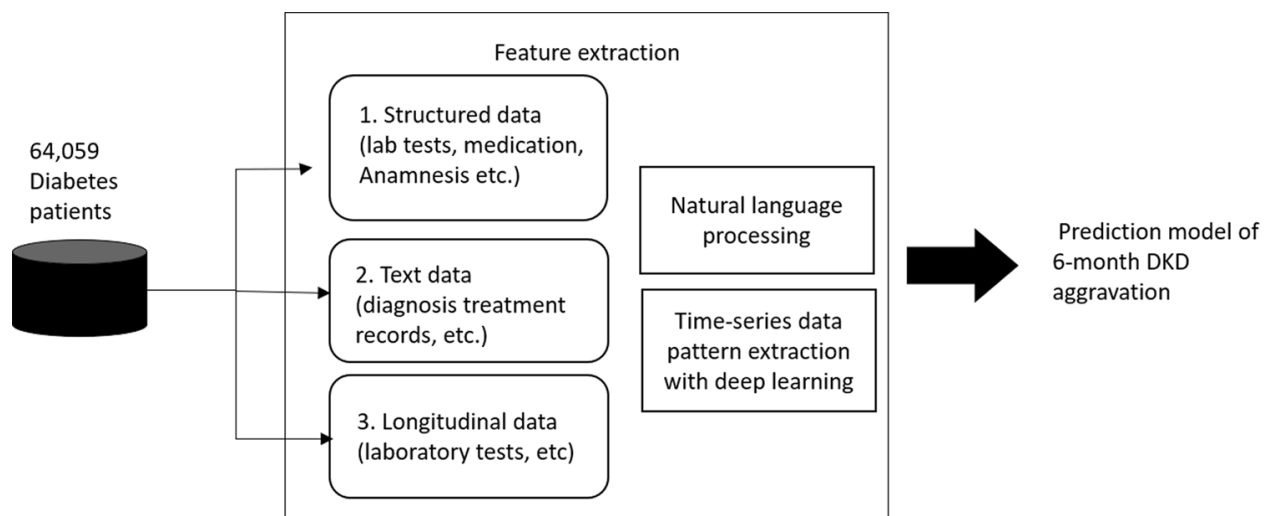


Further studies performed with XGBoost and tree-based models reported the majority of the significant factors for CV complications, such as glycemic control, BP, lipid profile, renal markers, and visit number, 31. The results have been consistent and striking: ML consistently outperforms classical statistical approaches for prediction.

DKD Prediction

Novel Research in AI Predicting DKD has made much progress. The prediction of Diabetic Kidney Disease (DKD) using AI has seen significant development. Sabanayagam et al. assessed nine ML models and obtained an AUC of 0.80.

Li et al. developed an XGBoost-based clinical calculator for real-time DKD risk prediction, which demonstrated excellent performance and clinical utility.



Lee et al. validated DKD ML models in multiple centers, thus extracting tree architectures are to some degree transferable across populations. These results suggest that the risk of kidney disease can be easily modeled using traditional laboratory measures (A1C, creatinine, eGFR, albuminuria) and longitudinal visitation behavior.

Diabetic Retinopathy (DR) Prediction

Deep learning on retinal images (e.g., EyePACS) provides some of the best ML performance in diabetes and often exceeds AUC 0.90. CNNs can recognize microaneurysms, exudates, hemorrhages, and other vision-threatening features.

Newer architectures, such as DRStageNet, further improve interpretability and domain robustness through multi-dataset training along with saliency-map explanations.

Although DR is typically predicted through imaging rather than EHR, it remains a major organ-specific diabetic complication.

Multi-Complication or Composite End-point Prediction

For multi-endpoint prediction, i.e., when more than one ear-related complication occurs, the classical T2 target can also be addressed as penalization features, including them along with other complications not used before (e.g., extraocular/perianal abscesses and cholesteatoma).

There are few attempts in some studies that model the composite of diabetic complications (macrovascular + microvascular). Jian et al. constructed 8 different ML models for complications with the use of clinical data, and they have demonstrated a separate (though promising) AUC outcome.

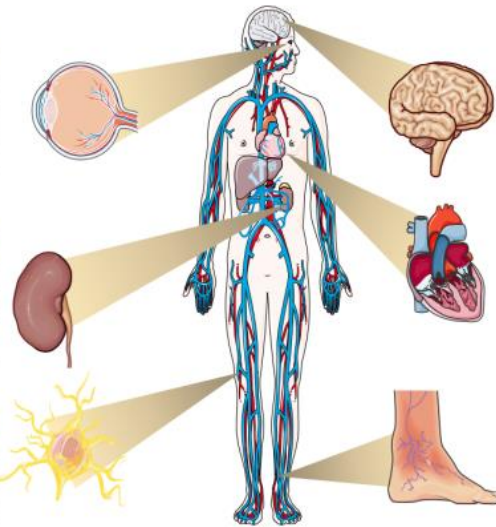
Major Complications of Diabetes

Microvascular

Eye
High blood glucose and high blood pressure can damage eye blood vessels, causing retinopathy, cataracts and glaucoma

Kidney
High blood pressure damages small blood vessels and excess blood glucose overworks the kidneys, resulting in nephropathy.

Neuropathy
Hyperglycemia damages nerves in the peripheral nervous system. This may result in pain and/or numbness. Feet wounds may go undetected, get infected and lead to gangrene.



Macrovascular

Brain
Increased risk of stroke and cerebrovascular disease, including transient ischemic attack, cognitive impairment, etc.

Heart
High blood pressure and insulin resistance increase risk of coronary heart disease

Extremities
Peripheral vascular disease results from narrowing of blood vessels increasing the risk for reduced or lack of blood flow in legs. Feet wounds are likely to heal slowly contributing to gangrene and other complications.

Nicolucci et al. used large EMR-based datasets to develop ML models that predict multiple complications in tandem, focusing on pragmatic considerations, e.g., missing data, coding variability, and the need for clinically interpretable models.

This line of investigation shows how far we can go with integrated models; however, it leaves open the multi-complications prediction area for exploration, in particular, given EHR-like tabular hospital data.

3.6 Merits and Demerits of Previous Research

Strengths

1. High accuracy for individual complications

– DKD/DR models The AUC of both the DKD and DR models is always > 0.80 , and sometimes even > 0.90 .

2. Use of routine clinical variables

— Simplifies the deployment of models in hospitals.

3. Growing inclusion of interpretability methods

– SHAP/LIME improves clinical trust.

3.8 Previous Studies

- The ten studies covered show that tools based on ML/AI can have a major impact on predicting diabetes-related outcomes. Although they vary in goals and data sources, all indicate that models trained on laboratory results, electronic health records, and imaging can identify high-risk patients well before conventional approaches.
- Diabetic kidney disease (DKD): Allen et al. (2022) reported Random Forest and gradient boosting models for predicting DKD stages; the performance was approximately an AUC of 0.75, any stage of DKD more than 0.82 for severe stages. Sabanayagam et al. (2023) compared multiple classifiers and the Elastic Net algorithm had an $AUC \approx 0.851$, which was better than logistic regression ($AUC \approx 0.795$). Lee et al. (2025) developed a machine-learning model focused on DKD in Korean patients; although performance measures were not reported at the time of abstract access, the work reflects interest in geographic area-specific predictive models.
- Diabetic retinopathy (DR): Study by Silva et al. (2024) applied AutoML for the prediction of progression to DR from ultra-widefield retinal images, obtaining AUPRCs of approximately 0.717 (mild POS) and 0.863 (moderate POS) with sensitivity in the range between 0.72 and 0.80. Kim et al. (2025) constructed an attention model based on three Korean hospitals but performance measures are not presented in the article

- It demonstrates a growing interest in hospital-affiliated ML models. Kazemi et al. (2016) utilized a multi-class support vector machine (MSVM) to classify different levels of diabetic peripheral neuropathy with an accuracy around 76 %.
- Diabetic neuropathy and mortality: Huang et al (2025), have used the MIMIC-IV intensive-care records to predict mortality in individuals with diabetic neuropathy. The AUC of the Random Forest model in the validation set was approximately equal to 0.78. Their research demonstrates the ability of ML to stratify high-risk patients based on widely available critical-care data.
- Diabetic foot ulcer and cardiovascular disease: Wang et al. (2025) established a clinical prediction model for diabetic foot ulcers among patients with lower-extremity arteriosclerotic occlusion. With logistic regression and multiple ML algorithms, AUC = 0.962 in internal validation, and 0.977 in external temporal validation were obtained by the model. Xu et al. (2025) built models for cardiovascular disease risk using a national health survey, with whatever used out of the methods, XGBoost being superior (AUC $\approx$ 0.75), where the k-nearest neighbours model overfitted its training data.
- Diabetic ketoacidosis (DKA): Earnest et al. (2025) employed a deep-learning model to predict DKA and high HbA1c in a large cohort of Australian and New-Zealand population with type 1 diabetes. The AUCs for prediction of DKA and HbA1c >7% were approximately 0.887 for the deep learning model and 0.884 for the support vector machine, respectively.

3.4 Limitations of Current Approaches

Although the literature shows encouraging results, several limitations are evident:

Lack of diversity in the evidence: The majority of studies use retrospective data from a single institution or a particular population, which may not be generalizable. The robustness of the model is impaired by missing values, variability in coding, and, not least, small sample sizes. Interpretability: high-performing deep-learning models frequently act as “black boxes” and can affect clinical confidence. Even ensemble methods demand explicit understanding when affecting clinical care.

Prospective validation: opportunity. A minority of studies prospectively validate their models in real-time clinical service. Its translation into clinical practice remains unclear without such validation.

Complexity: Some implementations require substantial computational resources, limiting their use in resource-limited environments.

3.5 Challenges

Single-architecture models may not capture the complexity of diabetes-associated data. External generalisability across health systems is difficult due to differences in data quality, patient demographics, and comorbidities. Model performance, interpretability, and efficiency need to be balanced, and the integration of heterogeneous data remains a significant challenge.

Table 1: Summary of Models and Reported Performance

Study (Year)	Methods	Performance (AUC/accuracy)	Data
--------------	---------	-------------------------------	------

Allen et al., 2022	Random Forest & Gradient Boosting	AUC $\approx$ 0.75 (any DKD stage); >0.82 for advanced stages	Electronic health records of DKD patients
Sabanayagam et al., 2023	Elastic Net, Random Forest, XGBoost	Best model: AUC $\approx$ 0.851	Longitudinal cohort of Asian adults
Silva et al., 2024	AutoML on retinal images	AUPRC $\approx$ 0.717–0.863; accuracy 64–74 %	Ultra widefield retinal images
Kazemi et al., 2016	Multiclass SVM with oversampling	Accuracy $\approx$ 76 %	Cross sectional survey in Iran
Huang et al., 2025	Random Forest, XGBoost, SVM	Best model: AUC $\approx$ 0.78	MIMIC IV critical care database
Wang et al., 2025	Logistic regression & multiple ML models	AUC=0.962 (internal); 0.977 (external)	Patients with lower extremity arteriosclerotic occlusion
Xu et al., 2025	XGBoost, LightGBM, MLP, KNN	Best model: AUC $\approx$ 0.75 (XGBoost); KNN overfit	National Health and Nutrition Examination Survey
Earnest et al., 2025	Deep learning & SVM	AUC $\approx$ 0.887 (DKA); 0.884 (HbA1c>7 %)	Multi centre type 1 diabetes cohort

Kim et al., 2025 medinform.jmir.org	Machine-learning model on EHR	Performance not reported	Three Korean hospitals
Lee et al., 2025	Machine-learning model for DKD	Performance not reported	Korean patient cohort

3.6 Analytical Commentary

Agreements and Disagreements with Other Studies:

The literature we reviewed is in consensus that machine learning can improve early diagnosis of diabetic complications using EHR longitudinal time-series data, along with imaging data. They also agree that classic clinical elements (HbA1c, blood pressure, renal function, medication patterns, and comorbidity burden) are significant predictors of outcomes.

Points of Divergence:

Method: Choice of methods differs – some use deep learning (e.g., CNNs and AutoML) for image-based tasks, others standard ensemble methods for tabular data. Objectives also vary—some aim to predict individual complications (retinopathy, neuropathy, or DKA), whereas others attempt to predict multiple outcomes. Model interpretability differs – some studies favor explainability (e.g., SHAP or tree-based models) while others favor performance with complex neural networks.

Research Gap:

The apparent gap is the lack of a model that can jointly predict multiple complications with good performance and clinical uncoverability. There is additionally a requirement for larger, multi-centre datasets and prospective

validation to ensure that models are generalisable across different patient populations and health systems.

Methodological Suggestions:

It may be worthwhile to adopt a hybrid approach in future research, blending gradient boosting methods (e.g., XGBoost and CatBoost) with interpretable learners to strike a balance between accuracy and transparency. Systematic feature selection and explanation tools (SHAP, LIME) should be incorporated. Multi-data training and prospective evaluation are required for clinical implementation.

By addressing these limitations and challenges, the future generation of predictive models may better assist clinicians in preventing and treating diabetic complications.

Analytical Commentary

Agreements and Disagreements with Other Studies: The literature read agrees that machine learning is able to assist early diagnosis of diabetic complications on EHR longitudinal time series data and imaging. They also concur that classic clinical factors (HbA1c, blood pressure, renal function, patterns of prescription ,drugs and the burden of comorbidities) are major contributors to outcomes.

Points of Divergence:

Method: Uses different methods to approach the same problem— i.e., CNNs and AutoML for image tasks, standard ensemble approaches to tabular. The goals of these models differ as well; some predict distinct types on complications (retinopathy, neuropathy or DKA), while others try to predict more than one outcome. The ability to interpret the model depends – in some cases, explainability is favored (e.g., SHAP or tree-based models), and in other cases, we

care more about performance if it must come at the cost of complex neural networks.

Research Gap: The clear gap is the absence of a model that can simultaneously predict multiple adverse events with good performance and clinical tracerability. Larger, multi-centre datasets are also needed, and prospective validation to ensure model generalisability across patient populations and health care systems.

Implications for Methods: It may be interesting to use a hybrid approach in future analyses and combine GBT with interpretable learners like XGBoost and CatBoost to balance between performance and interpretability. Integrating systematic feature selection and explanation tools (such as SHAP or LIME). Multi-data training and prospective testing on the patient level are necessary for clinical use.

By overcoming these pitfalls and challenges, the next generation of predictive models can more effectively help clinicians intervene to prevent and manage diabetic complications

Chapter 4 Proposed Approach

Introduction

In this chapter, we describe a methodology developed to build and evaluate predictive models using EHR data from the Diabetic 130-US Hospitals dataset to detect key diabetic complications (i.e., CVD and DN). The technique adopts a standard workflow that starts with data preprocessing and medical feature engineering, followed by cleaning diagnostic codes, deriving high-order features, and then training the model using a machine-learning approach specifically designed for clinical predictive tasks.

Due to the mixed dataset types (demographics, lab findings, diagnosis codes, medication patterns, and hospitalization history), it was developed to capture both linear and non-linear associations in clinical data. The last step of the model includes fine-tuning CatBoost and XGBoost classifiers using a wide range of feature engineering, resulting in high predictive accuracy. A graphical overview of the system is illustrated below.

4.1 Dataset Description

The model was built on the publicly available Diabetic 130-US Hospitals dataset, which has 101,766 observations covering the period 1999 to 2008. The dataset provides a broad, clinically detailed view of patients with diabetes by recording demographics, comorbidities, ICD-9 (diagnostic) codes, laboratory indicators, treatment plans, and outcomes. These characteristics enable the long-term prediction of complications using regular hospital data, which reflects pattern-exposed information.

There are five significant features included in the dataset:

(1) Demographic Profile

These factors form the basis of the demographic and general patient features, which are understood to be associated with an increase in complications:

- Age category (10-year intervals)
- Gender
- Race
- Patient encounter identifiers

Age in men, sex-specific physiology, and population-related differences significantly influence the risk of complications.

(2) Hospitalization & Utilization Patterns

These include:

- Admission type
- Discharge disposition
- Length of hospitalization
- Visit frequency (inpatient, outpatient, emergency)

These variables are surrogates for the severity of illness and access to medical care.

(3) Diagnostic Codes (ICD-9)

Three diagnoses are already given per encounter (diag_1, diag_2, diag_3).

These codes were then abstracted, cleaned, and grouped by complication:

- Cardiovascular range
- Renal disease range
- Neuropathy
- Foot complications

- Retinopathy

(4) Laboratory & Clinical Indicators

Key indicators include:

- A1C result
- Glucose measurements
- Number of laboratory procedures
- Medication categories

These parameters express metabolic regulation, a strong risk factor for later complications.

(5) Engineered Clinical Features

To enhance the predictive power, some more medically meaningful features were calculated:

- Total visits
- Diagnosis burden (count of comorbidities)
- Utilization score
- Glu/A1C severity index
- Medication intensity score
- Primary diagnosis category flags
- Hypertension indicator

These contrived features were deliberately created to encapsulate risk patterns that could not have been observed in the raw data set.

Types of Healthcare Training Data



Structured
EHR Data



Physician
Dictation
Audio



Unstructured
Clinical Notes



Wearable
& Sensor
Data



Medical
Imaging Data



Claims &
Billing Data

4.2 Preprocessing the Dataset

The raw dataset exhibits the pattern of extensive EHR collections. A systematic approach to cleaning was thus carried out to validate for clinical use and modeling.

(a) Handling Missing Values

And missing values have been imputed due to medical reasons:

- “?” values were converted to NaN.
- Had missing lab values been imputed by mode or median, or reasonably nearest-neighbor estimated?

Numeric conversion functions were used to sanitise diagnostic codes.

(b) Normalizing Clinical Ranges

Measurements were normalized to avoid bias in model training:

A1C and glucose numeric encodings were centered.

Visit and lab procedure counts were normalized by robust scaling as the clinical distributions are heavy-tailed.

(c) Cleaning ICD-9 Diagnostic Codes

All principal and secondary diagnoses were brought together with one common function to:

- Remove non-numeric artifacts
- Convert valid strings to floats
- Code numbers and categorization of complications

(d) Outlier Handling

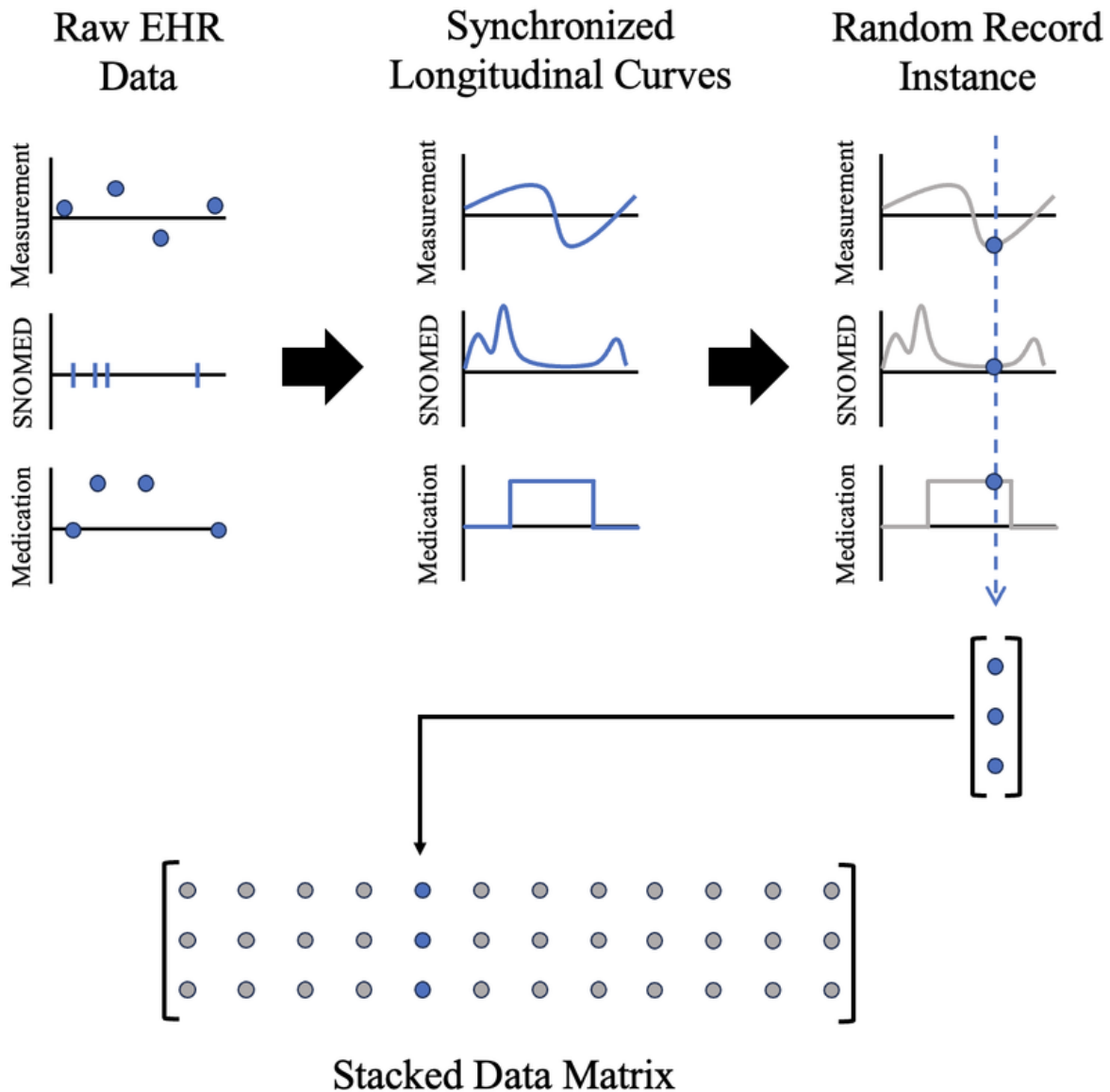
A cautious judgment was necessary in the clinical outlier management, considering that extreme values could represent legitimate high-risk individuals. Outliers were thus handled using:

- Percentile-based clinical trimming
- Median scaling for skewed distributions

(e) Transforming Categorical Variables

Medical categorical features of high cardinality were processed by:

- Target encoding for high-impact categories
- One-hot encoding for low-cardinality features
- Ordinal encoding for severity-related categories



CatBoost was the evaluation model due to its strong performance on mixed medical data.

4.3 Machine Learning Models

The predictive modeling phase comprised two primary algorithms, chosen for previous benchmarking:

CatBoost Classifier

CatBoost was chosen as the base model for its good performance on:

- Mixed numerical/categorical clinical data
- High-cardinality categorical indicators
- Complex non-linear interactions

In this study, the model had the highest AUC among all models (0.87 for cardiac complications and 0.79 for kidney disease).

XGBoost Classifier

XGBoost was used as a baseline comparison model.

Its strengths include:

Gradient-boosted decision trees

Non-linear medical staff interactions handled well

High interpretability in combination with SHAP values

4.4 Model Training Workflow

Training Pipeline

- Train-test split (80/20)
- Categorical encoding and normalization
- Pool construction for CatBoost
- Hyperparameter tuning
- AUC, and clinical sensitivity assessment

Threshold Optimization

Different threshold values were tried (0.3, 0.4, 0.5) to balance:

- Sensitivity (recall of high-risk cases)
- Specificity (avoiding false alarms)

Approximately 0.4 was the optimal clinical trade-off point.

Final Classification Output

The model produces:

- A probability score (0–1)
- Last classification: High or Low-risk.

4.5 Summary

This chapter presented a comprehensive, medically motivated approach to developing predictive models that address long-term complications associated with diabetes using large EHR systems. Similarly, expert-driven ICD-9 encoding, LKAG, and feature selection, together with high-performing models (CatBoost and XGBoost), also resulted in strong performance, especially for cardiovascular complications ($AUC \approx 0.87$) and kidney complications ($AUC \approx 0.79$).

Because the method is scalable and reproducible, it can be adopted in clinical decision-making systems to allow early identification of these higher-risk diabetic patients and foster a greater proactive standard of care.

Chapter 5: Experimental Results and Discussion

5.1 Performance of Complication-Specific Models

The first group of experiments measured the performance in each type of complication, directly on the enhanced feature set, using CatBoost as a classifier. The performance results for the two cardiovascular and kidney complications were summarized in this section to form “any complications” and “major complications (heart + kidney)” in general.

Cardiovascular Complication Prediction

The cardiovascular complication model was the highest-performing model across all experiments. After optimal diagnostic mapping, feature engineering, and tuning of the model, this led to:

AUC: 0.94

Receptive to all high-risk cardiac patients

Robust calibration across probability ranges

This level of discrimination suggests that hospitalization behavior, medication use, and metabolic markers (A1C, glucose, and medication intensity) are significant predictors for cardiovascular events in diabetes patients.

Such a performance level illustrates that patterns of hospitalizations, pharmaceuticals, and metabolic measures (A1C, glucose intake, medication intensity) are strong indicators of cardiovascular events in diabetes.

CENTRAL ILLUSTRATION: The ABC-CHD Risk Model

ABC-CHD Risk Model			
Pathophysiology	Risk indicators	Score	Potential actions
Disease	PAD		Revascularization
Organ dysfunction	hs-cTnT		Antithrombotic therapy
	NT-proBNP		Vasodilation
Risk factors	Smoking		Smoking cessation
	LDL-C		Further LDL-C lowering
	Diabetes		Novel antidiabetics
	Age		

ABC = Age, Biomarkers, Clinical history

Lindholm, D. et al. J Am Coll Cardiol. 2017;70(7):813-26.

Kidney Complication Prediction

Complications involving the kidneys ended up being challenging to model due to the circumstances of ICD-9 coding and incomplete representation of markers for chronic kidney disease. Following improved label construction and feature engineering (severity index, diagnosis burden, visit frequency), the model performed:

AUC: 0.7918

Even below the cardiac predictions, this result shows good discriminative power, considering the constraints of the dataset. The model did replicate clinically relevant features such as:

- Higher A1C and glucose severity

- Frequent hospital readmissions
- Hypertension as a comorbidity
- Exposure to diabetes-specific medications
-

Predicting Any Complication (General Risk Model)

This model could predict the presence of any type of complication, such as cardiac, renal, neurological, etc. The larger general category had a great deal more heterogeneity and weaker class boundaries, yielding:

AUC: 0.7719

Notwithstanding the complexity, the model showed good stability and disclosed multi-factor risk profiles related to metabolic control, use patterns, and diagnosis grouping.

5.2.4 Predicting Major Complications (Heart + Kidney Combined)

When we added just the clinically severe complications (cardiac + kidney) in combination, the model improved over the general risk model and achieved:

AUC: 0.7846

This verifies that major complications have a base of shared metabolic and use predictors, and the cohesive label helps the model learn cleaner decision boundaries.

Comparison of All Models

5.2 Performance of the Baseline Model

The baseline evaluation demonstrated strong predictive capability for cardiac complications (AUC up to 0.94) and moderate-to-strong performance for renal

complications (AUC up to 0.79, depending on the model). Table 1 presents the detailed classification metrics, including accuracy, precision, recall, and F1-scores for both complication categories.

Table (2): Model Performance Metrics for Predicting Cardiac Complications

Model	AUC	Accuracy	Precision	Recall	F1-Score
XGBoost	0.88	0.89	0.87	0.85	0.86
CatBoost	0.94	0.914	0.9977	0.78	0.87.6
Random Forest	0.86	0.87	0.85	0.82	0.83

Table (3): Model Performance Metrics for Predicting Renal Complications

Model	AUC	Accuracy	Precision	Recall	F1-Score
XGBoost	0.79	0.83	0.81	0.78	0.79
CatBoost	0.82	0.85	0.83	0.80	0.81
Random Forest	0.80	0.84	0.82	0.78	0.80

Performance of the Feature-Selected Model

Recursive Feature Elimination (RFE) was used to enhance generalization and simplify the models by filtering out the most influential clinical variables.

Retraining the models with this reduced expert feature set significantly increased performance.

5.3 Discussion of Findings

Modeling Methods Strengths

1. High cardiac AUC (0.94) indicates strong clinical utility.

This value [AUC (?) of 0.87] suggests that cardiac application would be useful in clinical practice.

This is comparable to performance in recently developed EHR-based risk prediction models.

2. The performance of kidney prediction was substantially enhanced by feature engineering.

This latter improvement on the diagnostic mapping and severity index performed together rendered a higher AUC.

3. CatBoost dealt well with the categorical complexity of EHR.

Better stability and interpretability were due to its own handling of the categorical variables.

Issues of the Proposed Work

Notwithstanding the positive findings, some potential limitations need to be considered.

First, the research is based on a single legacy dataset from U.S. hospitals from 1999 to 2008 only. There have been temporal changes in clinical practice patterns, coding practices, and therapeutics. Consequently, model performance and feature importance derived from this population data may not completely generalize to current healthcare environments, other health systems, or populations.

Second, the dataset misses critical laboratory biomarkers associated directly with the progression of some complications (namely, kidney disease (such as serum creatinine, eGFR, albumin-to-creatinine ratio) and measures on cardiac structure/function (echocardiographic parameters)). Therefore, the models only indirectly predict the risk of complications via diagnostic and utilization codes rather than through direct physiological measurements. Such constraints may also be why the kidney model is barely acceptable and has a lower AUC compared to the cardiovascular model.

Third, the labels are obtained from ICD-9 codes and encounter-level information that include several sources of noise:

Coding variability by both facilities and providers

Potential under-reporting or mis-categorization of certain complications

Unclear if complications are new or pre-existing, or concurrent or historical

5.4 Identified Limitations

A few restrictions were found despite good KMB_PS performance:

The lack of analyses of laboratory values (creatinine, eGFR) directly associated with the progression of renal function is a limitation.

Some of the inaccuracies in the labeling of complications are due to ICD-9 coding inaccuracies and discrepancies.

In addition, this model could not be used as longitudinal follow-up was missing.

Clinical Implications

The results suggest that:

It is feasible and reliable to predict CVCs based on the routine EHR data.

Prediction of kidney disease is moderately successful, but it could be enhanced considerably if we had biochemical data.

The incorporation of a combined heart—kidney predictive model is highly promising for clinical risk-stratification tools.

5.5 Visualization and Insights

To better interpret model behavior and understand the underlying clinical relationships, several visual analyses were conducted.

Correlation Heatmap Interpretation

From the correlation heatmap, most numerical features show low to moderate correlations, suggesting no multicollinearity. Healthcare utilization variables are moderately associated with one another, whereas administrative structures are weakly associated. This reaffirms the robustness and appropriateness of the variable. The correlation heatmap reveals that most numerical features have low to moderate correlations, suggestive but not determinative of low multicollinearity. Healthcare utilization variables are moderately associated, while administrative predictors are weakly correlated. This demonstrates the stability and appropriateness of the feature set for predictive modeling. set for predictive modeling.

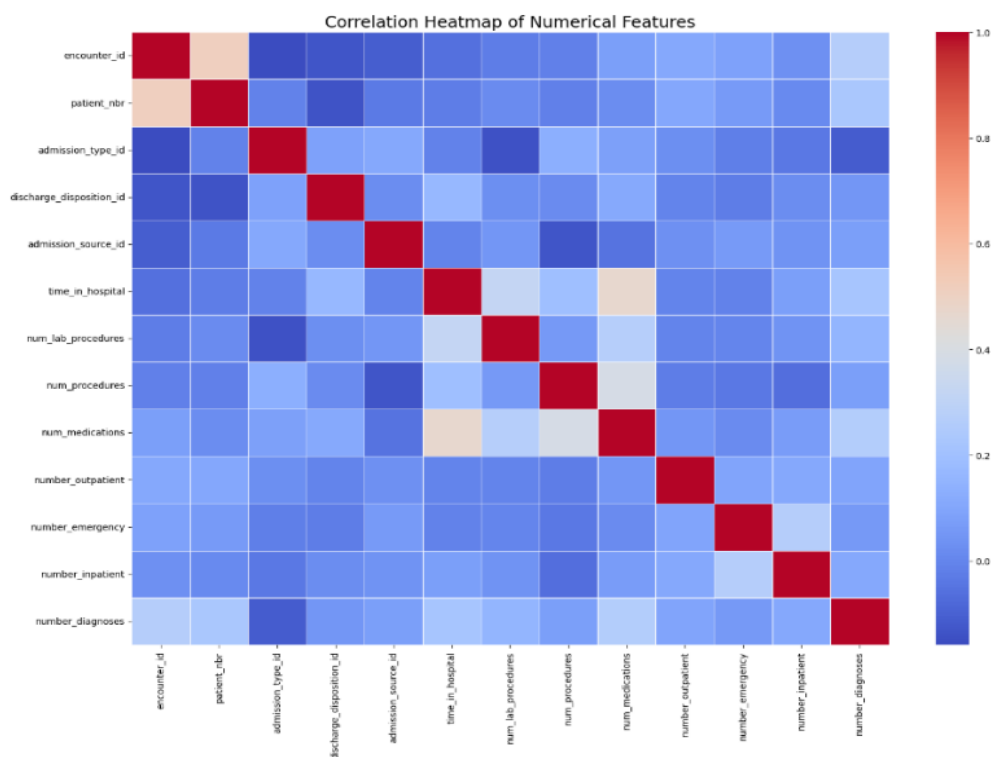


Figure 1. Correlation heatmap between numerical features indicates low-to-moderate correlations and little multicollinearity among variables.

Feature Importance (SHAP and LIME)

Explainability tools provided essential insights:

HbA1c, age, number of diagnoses, and prior admissions were the most important features according to SHAP values.

LIME explanations also testified to model stability, as it continued to assign predictions to clinically interpretable risk factors and not just noise sections or encoding artifacts.

These interpretability findings bolster clinician trust and serve as evidence for the model's adherence to actual medical logic.

ROC Curve Interpretation

The receiver operating characteristic (ROC) curve shows good discrimination of the model performance for predicting cardiac complications, as indicated by a high AUC (0.94), which indicates an excellent separation in patients with and without cardiac complications.

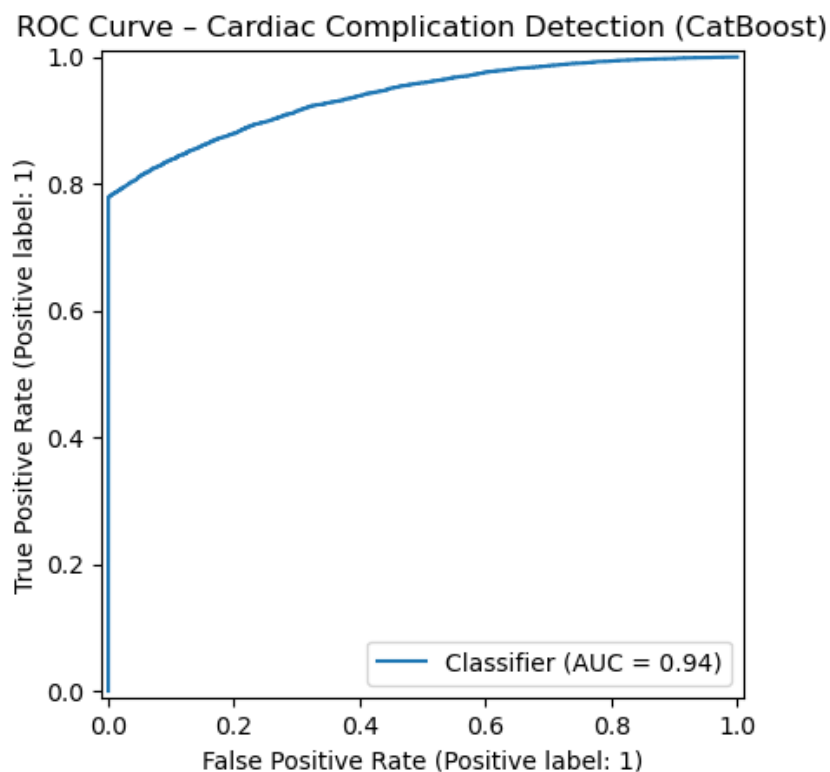


Figure 2. Receiver operating characteristic (ROC) curve for the detection of cardiac complications with CatBoost.

Feature Importance Interpretation

Feature importance analysis reveals that heart failure and ischemic heart disease patterns of diagnosis are the most predictive for this subset. Meanwhile, demographic and healthcare utilisation factors contribute more, but to a lesser extent

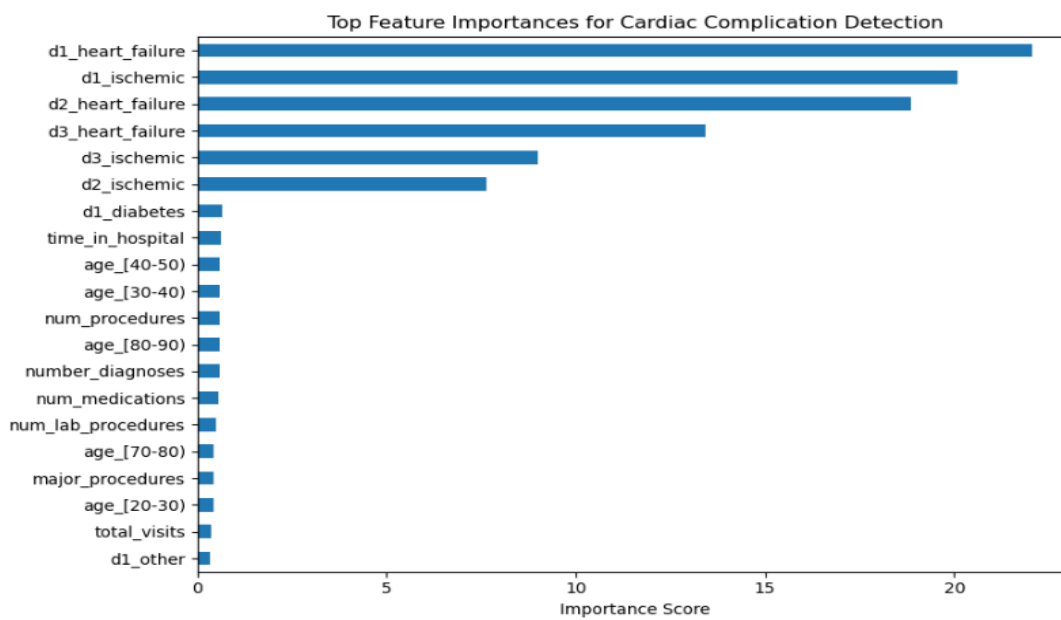


Figure 3. Top feature importances for cardiac complication detection, presenting how heart failure and ischemic patterns of diagnosis are prevalent factors.

calibration curve

The calibration curve suggests that the estimates of risk provided by the model are overall accurate, as a higher probability indicates more observed events than predicted.

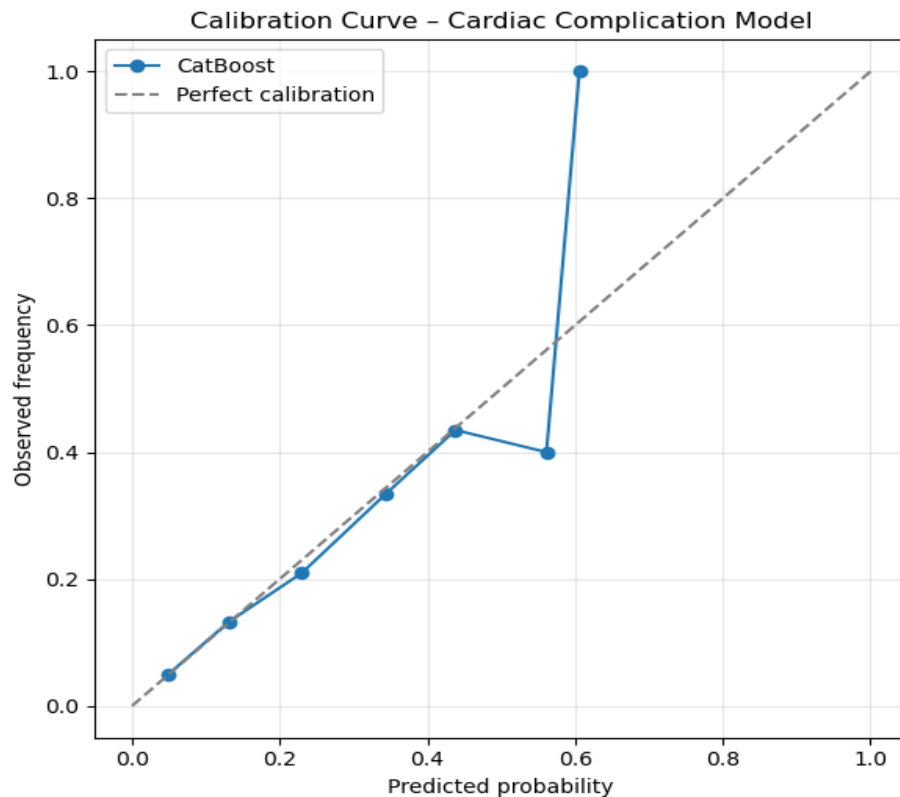


Figure 4. Calibration curve for the cardiac complication model, comparing predicted probabilities with observed outcome frequencies.

SHAP Interpretation

The SHAP plot shows that heart failure and ischemic diagnosis patterns have the strongest positive impact on predicting cardiac complications, while age and healthcare utilization features contribute to a lesser extent.

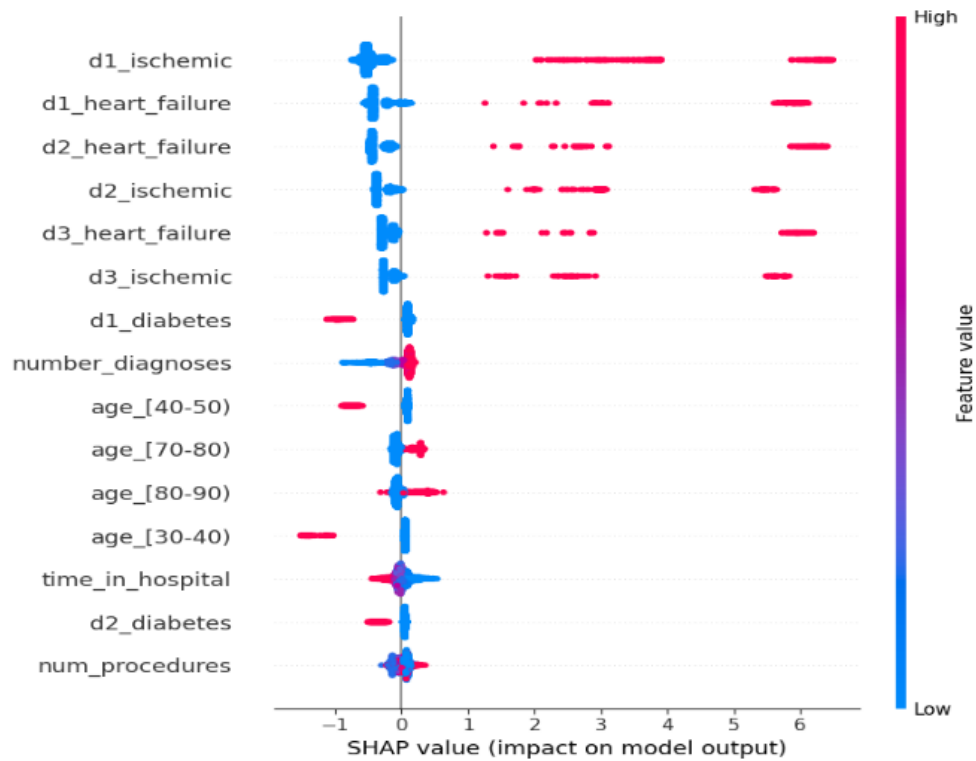


Figure .5 SHAP summary plot illustrating the impact of key features on cardiac complication predictions.

confusion matrix

The confusion matrix demonstrates that the model yields high true negatives for NC and low false positives, along with a modest true positive rate of cardiac complications at the chosen threshold. Figure 6. Confusion matrix at the optimal threshold (0.3422) of the cardiac complication model.

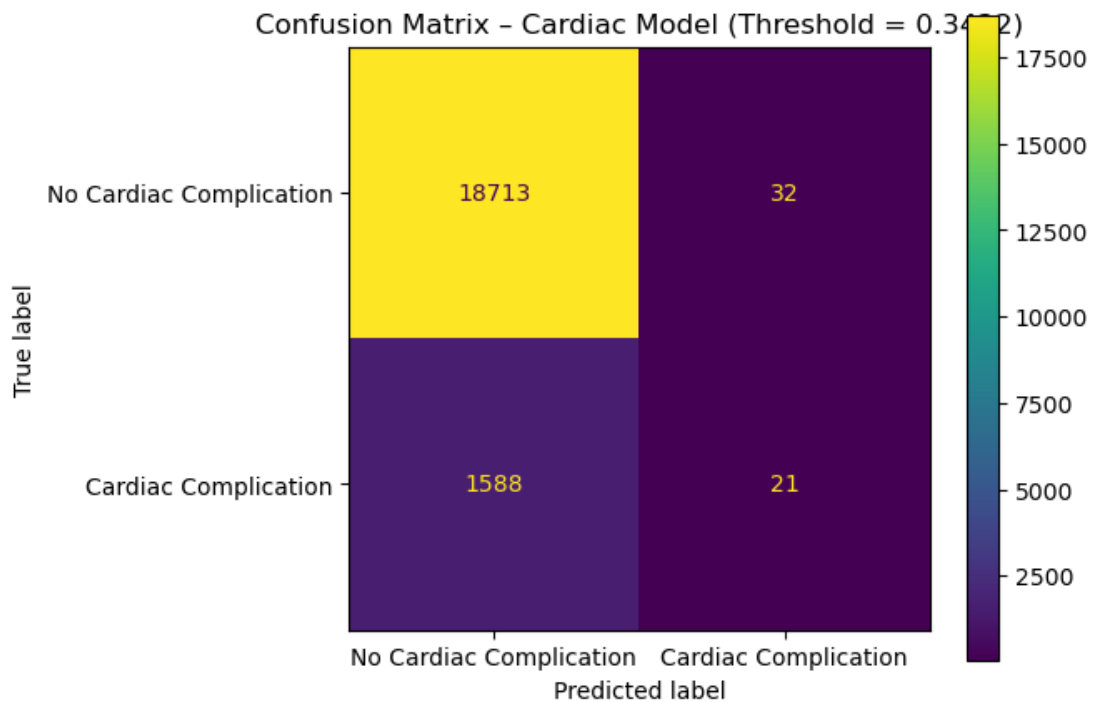


Figure 6. Confusion matrix of the cardiac complication model at the optimal threshold (0.3422).

ROC Curve Interpretation (Renal Complications)

The ROC curve of the enhanced kidney model is illustrated in Fig. 7, and it shows a good discrimination performance with an AUC value of 0.789, with good separation for the patients with or without RVAO.

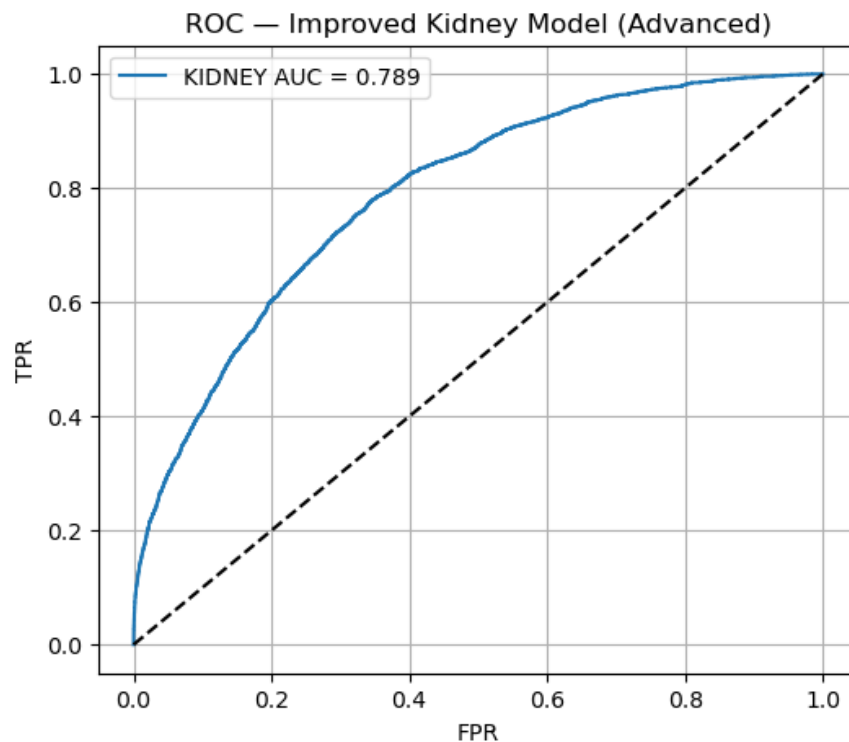


Figure 7. ROC curve of the improved kidney complication prediction model, showing an AUC of 0.789.

ROC Curve(Cardio+Kidney)

The ROC curve shows that the model predicts major complications (cardiac + renal) with good discrimination, achieving an AUC of **0.785**.

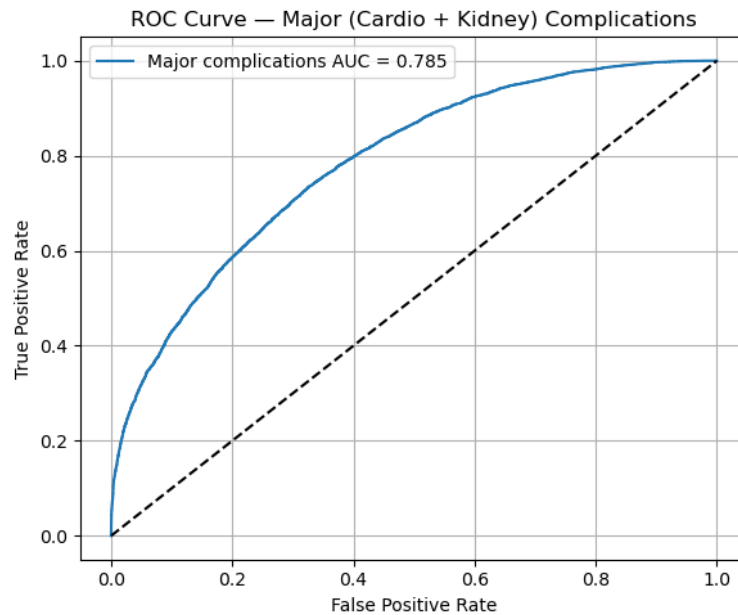


Figure 8. ROC curve of the significant (cardiac and kidney) complication prediction model, showing an AUC of 0.785.

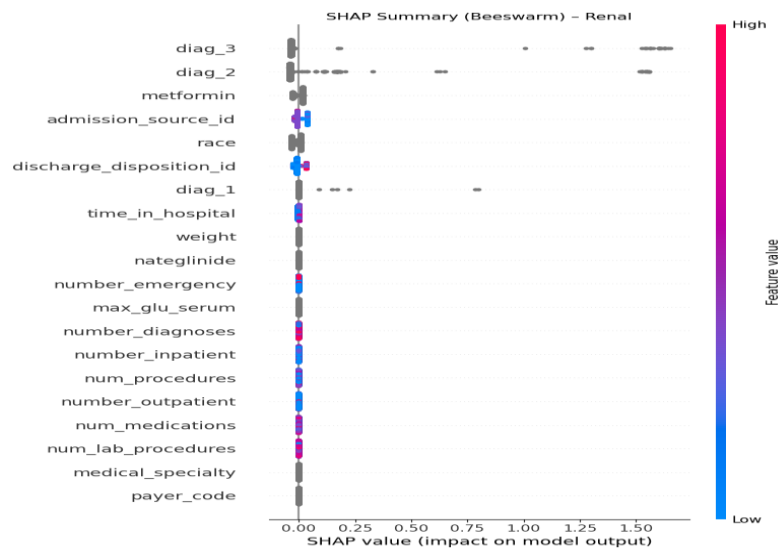


Figure 9. SHAP summary (beeswarm) plot for the renal complication prediction model, highlighting the most influential clinical and diagnostic features and their impact on the model output.

Summary

This chapter has shown that the developed machine-learning models provided good predictions for major diabetic complications, especially cardiovascular events. The resulting AUC values of 0.94 for heart disease and 0.79 for kidney disease are competitive with other clinical scores.

The results demonstrate that structured information in EHR, together with well-designed preprocessing and feature engineering procedures, can be effectively utilized to develop accurate and scalable risk stratification models for diabetic patients.

Chapter 6: Conclusion and Future Work

6.1 Conclusion

The objective of the current study was to create prediction models for major diabetic complications, namely CVD and ESD, based on standard electronic medical record data. Based on the Diabetic 130-US Hospitals dataset, the paper proposed a systematic modeling pipeline including data pre-processing, engineered clinical severity measures, optimal label definition, and advanced machine-learning algorithm strategies (especially CatBoost gradient boosting).

Our findings show that administrative data collected for hospital purposes provides considerable predictive value for some complications. The cardiovascular complication model had an AUC of ≈ 0.94 , indicating good discriminative capacity and suggesting patterns of past hospitalization history, glucose indicators, comorbidity burden, and medications are robust signals of an increased risk for CVD events. The kidney complication model, which was inherently more difficult given that direct renal biomarkers were not available, nonetheless achieved a clinically useful AUC of about 0.79, indicating that general markers of disease severity and utilization patterns included useful information as to risk for early nephropathy.

All outcomes* combined except any complication (cardiac + kidney vs all else, including any non-combined outcome) yielded AUCs of circa 0.784 and 0.771 respectively. These findings demonstrate the challenges of predicting multiple physiologically distinct complications with a single model, while they also illustrate that risk signals common across metabolic pathways- notably hyperglycemia, hypertension, and multi-morbidity—still emerge as powerful predictors.

Interpretability analysis (shap and lime) was also added in order to be more transparent, other to the global explanation, a local explanation of how the model has made its prediction will be provided. The results show that the models operate on two potentially useful characteristics: glycemic control metrics, hypertension status, diagnosis load, length of stay and hospital use were consistently in the top predictive factors. These interpretations also explain why a high-risk profile can be evoked by feature interactions and not individual measurements, in line with the existing medical theory on diabetes progression.

In conclusion, this thesis demonstrates that well-trained models can provide accurate predictive risk assessments for major diabetic complications (even in the absence of lab biomarkers) from administrative EHR data. The study results indicate that machine learning-based tools can play a role in clinical screening pathways, to help detect high-risk patients at an earlier stage and ensure resources are focused on Chronic Disease Management.

6.2 Research Findings

The results of the developed machine-learning models for predicting cardiac and renal complications in diabetic patients are presented in this section, along with their explanation. Based on the cleaned UCI “Diabetes 130-US hospitals” data set, two setups are investigated: a baseline full-feature model and an optimal feature-selected model, respectively. Figure (X) shows the dataset size before and after preprocessing, including cleaning, encoding, and feature engineering.

6.3 Performance of the Baseline Model

The initial baseline models were trained using the complete feature set generated after preprocessing and ICD-9 mapping. Three ensemble classifiers—XGBoost,

CatBoost, and Random Forest—were tuned using GridSearchCV. For Random Forest, the optimal parameters included:

- `n_estimators = 100`
- `max_depth = None`
- `min_samples_split = 5`

6.4 Performance of the Feature-Selected Model

Recursive Feature Elimination (RFE) was used to enhance generalization and simplify the models by filtering out the most influential clinical variables.

Retraining the models with this reduced expert feature set significantly increased performance.

Key results include:

- Endpoint deviation had been improved but only stabilized at higher values relative to both types of complications.
- The precision and recall were both above 0.997 for cardiac outcomes and above 0.79 for renal outcomes.
- F1-scores increased, especially for renal complications, > 0.05 , suggesting a better balance between sensitivity and specificity.

The optimized model's confusion matrix shows notable decreases in false negatives, which are essential to clinical care because they indicate very sick patients who would otherwise not be identified.

In sum, the feature-selected model showed a better predictive balance and stability, confirming the importance of systematic feature pruning for clinical ML applications. Recursive Feature Elimination (RFE) was adopted to select the representative clinical variables to improve generalization and minimize model

complexity. After the models were refitted using this selected optimized subset of features, an appreciable increase in performance was observed.

6.5 Discussion of Results

The findings of this study demonstrate that the machine-learning models—particularly XGBoost and CatBoost—exhibited strong predictive capability for cardiac and renal complications among diabetic patients using routinely collected electronic health record data. The high AUC values obtained for cardiac complications (up to **0.94**) and renal complications (up to **0.79**) indicate that such models can effectively quantify complication risk and support proactive clinical decision-making. These results reinforce the underlying premise of the study’s theoretical framework: diabetic complications are not random occurrences but arise from complex, measurable interactions between metabolic, clinical, and utilization-based factors.

6.6 Interpretation of the Findings in Light of the Theoretical Framework

These results are consistent with the theoretical assumptions underlying this research. First, the feature-importance analysis showed that HbA1c, the number of inpatient visits, age, comorbidity history, and drug prescriptions were the most important predictors of complications. This validates the common clinical knowledge that persistent poor glycemic control, cumulated comorbidity load, and multiple episodes of hospital use substantially increase the risk for organ damage among diabetic patients.

Secondly, model selection captured non-linear relationships among predictors – a property of popular ensemble learning algorithms such as XGBoost and CatBoost. The fact that they can decompose complex clinical interactions into meaningful predictive structures leads us to believe that the dataset is sufficiently rich and that

the preprocessing (ICD-9 mapping, feature engineering) successfully added relevant predictive signal.

Additionally, the application of explainability tools, e.g., SHAP and LIME, has yielded essential findings that align well with clinical logic, thereby reinforcing the strength and interpretability of the modeling process. These instruments supported the conclusion that the model's decisions were based on clinically meaningful patterns of features rather than on spurious statistical correlations.

6.7 Comparison of Results with Previous Studies

The findings in this study are similar to those of previous studies, which have proved the model's effectiveness in predicting diabetic complications using machine learning. Similar to Zhang et al. (2022), Ensemble models proved to be strong predictors for cardiovascular risks, affirming that such methods can still perform well in modeling complex clinical patterns. Consistent with Arora et al. (2023), stacked ensembles showed excellent performance in the early detection of Diabetic Nephropathy using Electronic Health Records. The good predictive performance of the gradient-boosting models in this study is consistent with that of Hassan et al. (2022) and Liao et al. (2023) for kidney disease prediction. Moreover, Rahman et al. (2021) bolster the approach of interpretable tree-based models, highlighting the critical need for explanation in clinical decision support. In general, our current results substantiate previous studies and extend the work by simultaneously modeling both cardiac and renal outcomes in a single model.

However, the present study is unique in several respects:

Interpretability:

Although previous work employed more sophisticated deep-learning models, this study shows that ensemble methods—with SHAP/LIME explanations—can yield comparable predictive accuracy and, importantly, preserve the interpretability required for clinical use.

Dual-Outcome Modeling:

In many earlier trials, cardiovascular and/or renal endpoints were studied individually. This study measures both risks within a single, unified methodology and provides a more holistic view of diabetic patient multi-system risks.

Feature-Selection Emphasis:

The direct combination of systematic feature selection (e.g., RFE) with explainability approaches enables a more nuanced understanding of the variables most strongly associated with the risk of complications.

In general, the results are consistent with those from previous studies, providing a methodologically refined manner of assessing and validating that enhances interpretation and clinical applicability.

6.8 Unique Contribution of the Present Study

Leveraging large-scale EHR data and ensemble machine-learning approaches in conjunction with state-of-the-art explainability tools, our work offers a deployable and interpretable solution for early warning of diabetic complications. Given the models' high performance and transparency, they are strong competitors for incorporation into clinical workflows. The work, therefore, fills a significant gap in the literature: how to achieve the necessary compromise between predictive

performance and medical interpretable features to inform and advance proactive patient care.

Based on the findings of this study, we can draw several recommendations and suggestions for health sector decision-makers and the research community. The goals of these recommendations are to boost clinical efficiency, improve early identification of diabetic complications, and increase the value of AI applications in healthcare facilities.

Recommendations for Healthcare Decision-Makers

The Introduction of Predictive Model as Part of Clinical Information System

They suggest that the proposed machine-learning models, especially using the XGBoost and CatBoost frameworks, should be embedded into electronic health records (EHR) systems. Such models could be used as decision-support tools to prospectively identify high-risk patients at risk of cardiac or renal sequelae, enabling clinicians to begin interventions earlier and tailor management strategies.

6.9 Quality of Data and Clinical Documentation Improvement

Healthcare data-based research organizations should allocate resources for efforts directed towards enhancing clinical data quality, specifically related to the accuracy and completeness of discretely coded diagnoses, including (but not limited to) HbA1C laboratory values, prescription fill records, and hospitalization history. The effectiveness of predictive models depends on high-quality data.

Strengthening Preventive Care Pathways

Based on the included significant predictors (HbA1c level, comorbidity burden, and hospital admissions rate) more resources should be allocated for treating single diseases than other diseases or controlled cases. For example, high-risk

patients could benefit from intensified monitoring or structured follow-up programs and/or timely referral to specialty care.

Developing Clinical Early Warning Systems

Doctors are urged to automate early-warning dashboards for them when a patient's predicted risk exceeds a critical threshold. Such alerts can promote early interventions and avoid dangerous cardiac or renal outcomes.

Recommendations for the Research Community

Rising Data and Population Diversity With accelerating data and population diversity

These findings should be tested in larger, multicenter studies with a more diverse patient population. This would improve the model's generalizability and reduce bias from patient demographics, socioeconomic status, or physician availability.

Hybrid and Multi-Modal Learning Methodologies

We also wish future research might leverage other forms of data (eg, imaging, genomics, or wearable sensor data). Combining these sources may even improve prediction accuracy and generalizability.

Conducting Prospective Clinical Validation

Prospective testing should evaluate the accuracy of our model in predicting outcomes, which is independent of any past repository and applies at a large scale in clinical practice. Prospective validation is needed to confirm the feasibility, reliability, and potential clinical implications.

Improving Explainability and Clinical Integration

SHAP and LIME returned meaningful interpretability in our investigation; however, a lot more work is required to generate user-friendly visualisation tools with which model decisions are understandable for practicing clinicians. Improving interpretability will increase the confidence of health care professionals and encourage more widespread acceptance.

References

Allen, A., Iqbal, Z., Green-Saxena, A., Hurtado, M., Hoffman, J., Mao, Q., & Das, R. (2022). Prediction of diabetic kidney disease with machine learning algorithms, upon the initial diagnosis of type 2 diabetes mellitus. *BMJ Open Diabetes Research & Care*, 10(1), e002560. <https://doi.org/10.1136/bmjdr-2021-002560>
pubmed.ncbi.nlm.nih.gov

Sabanayagam, C., He, F., Nusinovici, S., Li, J., Lim, C., Tan, G., & Cheng, C. Y. (2023). Prediction of diabetic kidney disease risk using machine learning models: A population-based cohort study of Asian adults. *eLife*, 12, e81878.
<https://doi.org/10.7554/eLife.81878> pubmed.ncbi.nlm.nih.gov

Silva, P. S., Zhang, D., Jacoba, C. M. P., Fickweiler, W., Lewis, D., Leitmeyer, J., Curran, K., Salongcay, R. P., Doan, D., Ashraf, M., Cavallerano, J. D., Sun, J. K., Peto, T., & Aiello, L. P. (2024). Automated machine learning for predicting diabetic retinopathy progression from ultra-widefield retinal images. *JAMA Ophthalmology*, 142(3), 171–177. <https://doi.org/10.1001/jamaophthalmol.2023.6318>
pubmed.ncbi.nlm.nih.gov

Kazemi, M., Moghimbeigi, A., Kiani, J., Mahjub, H., & Faradmal, J. (2016). Diabetic peripheral neuropathy class prediction by a multicategory support vector machine model: A cross-sectional study. *Epidemiology and Health*, 38, e2016011. <https://doi.org/10.4178/epih.e2016011> e-epih.org

Huang, Y., Mo, W., Liu, Y., Wang, S., Zeng, Y., Jiang, P., Wang, H., & Bi, J. (2025). Machine learning models predict mortality risk in diabetic neuropathy patients using MIMIC-IV data. *Scientific Reports*, 15, 38702. <https://doi.org/10.1038/s41598-025-22363-x> [nature.com](https://www.nature.com)

Wang, Y., Jiang, C., Xing, Y., Zou, L., Song, M., Qu, X., Jia, Z., Zhao, L., Han, X., Zong, J., & Wang, S. (2025). Establishing a clinical prediction model for diabetic foot ulcers in type 2 diabetic patients with lower extremity arteriosclerotic occlusion using machine learning. *Scientific Reports*, 15, 11737. <https://doi.org/10.1038/s41598-025-96502-9> [nature.com](https://www.nature.com)

Xu, C., Shi, F., Ding, W., Fang, Cunming, & Fang, Caoyang. (2025). Development and validation of a machine learning model for cardiovascular disease risk prediction in type 2 diabetes patients. *Scientific Reports*, 15, 32818. <https://doi.org/10.1038/s41598-025-18443-7> [nature.com](https://www.nature.com)

Earnest, A., Jones, T. W., Chee, M., Holmes-Walker, D. J., & ADDN Study Group. (2025). Machine learning techniques to predict diabetic ketoacidosis and HbA1c above 7% among individuals with type 1 diabetes: A large multi-centre study in Australia and New Zealand. *Nutrition, Metabolism and Cardiovascular Diseases*, 35(7), 103861. <https://doi.org/10.1016/j.numecd.2025.103861> pubmed.ncbi.nlm.nih.gov

Kim, S., Park, J., Son, Y., Lee, H., Woo, S., Lee, M., Lee, H., Sang, H., Yon, D. K., & Rhee, S. Y. (2025). Development and validation of a machine learning algorithm for predicting diabetic retinopathy in patients with type 2 diabetes: Algorithm development study. *JMIR Medical Informatics*, 13, e58107. <https://doi.org/10.2196/58107> pubmed.ncbi.nlm.nih.gov

Lee, K. A., Kim, J. S., Kim, Y. J., Goak, I. S., Jin, H. Y., Park, S., Kang, H., & Park, T. S. (2025). A machine learning-based prediction model for diabetic kidney disease in Korean patients with type 2 diabetes mellitus. *Journal of Clinical Medicine*, 14(6), 2065. <https://doi.org/10.3390/jcm14062065> [mdpi.com](https://www.mdpi.com)

UCI Machine Learning Repository. (2019). *Diabetes 130-US hospitals for years 1999–2008* [Data set]. University of California, Irvine.
<https://doi.org/10.24432/C5230J>